Université de Montréal

Finite population inference for population with a large number of zero-valued observations

par

Isabelle Nolet-Pigeon

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

11 août 2020

Université de Montréal

Faculté des études supérieures et postdoctorales

Ce mémoire intitulé

Finite population inference for population with a large number of zero-valued observations

présenté par

Isabelle Nolet-Pigeon

a été évalué par un jury composé des personnes suivantes :

Pierre Duchesne
(président-rapporteur)

David Haziza
(directeur de recherche)

Mylène Bédard
(membre du jury)

Mémoire accepté le :
1er décembre 2020

Résumé

Dans certaines enquêtes auprès des entreprises, il n'est pas rare de s'intéresser à estimer le total ou la moyenne d'une variable qui, par sa nature, prend souvent une valeur nulle. En présence d'une grande proportion de valeurs nulles, les estimateurs usuels peuvent s'avérer inefficaces. Dans ce mémoire, nous étudions les propriétés des estimateurs habituels pour des populations exhibant une grande proportion de zéros. Dans un contexte d'une approche fondée sur le modèle, nous présentons des prédicteurs robustes à la présence de valeurs influentes pour ce type de populations. Finalement, nous effectuons des études par simulation afin d'évaluer la performance de divers estimateurs/prédicteurs en termes de biais et d'efficacité.

Mots-clefs: Robustesse ; Unités influentes ; Inférence basée sur le modèle ; Inférence basée sur le plan de sondage ; Biais conditionnel.

Abstract

In business surveys, we are often interested in estimating population means or totals of variables which, by nature, will often take a value of zero. In the presence of a large proportion of zero-valued observations, the customary estimators may be unstable. In this thesis, we study the properties of commonly used estimators for populations exhibiting a large proportion of zero-valued observations. In a model-based framework, we present some robust predictors in the presence of influential units. Finally, we perform simulation studies to evaluate the performance of several estimators in terms of bias and efficiency.

Keywords: Robustness ; Influential units ; Model-based inference ; Design-based inference ; Conditional bias.

Contents

Résumé 5

Abstract 7

List of tables 11

List of figures 13

Remerciements 15

Introduction 17

Chapter 1. The design-based approach and the model-based approach 19

1.1. Finite population and sample 19

1.2. Sampling design 19

1.3. The design-based approach 20

1.3.1. The Horvitz-Thompson estimator 21

1.3.2. The ratio estimator 22

1.3.3. The GREG estimator 23

1.3.4. Conditional Bias 25

1.4. Model-based approach 26

1.4.1. Conditional Bias 27

Chapter 2. Inference for populations with a large number of zero-valued observations 29

2.1. Design-based approach 30

2.1.1. The Horvitz-Thompson estimator 30
2.1.2. The ratio estimator 33
2.1.3. The GREG estimator 34
2.2. Model-based approach 35
2.2.1. Best Linear Unbiased Predictor (BLUP) 36
2.2.2. Empirical Best Predictor (EBP) 36
2.2.3. Estimation of the conditional bias 40
Chapter 3. Robust prediction 41
3.1. Robust regression 42
3.2. Predictor of Chambers 46
3.3. Predictor based on the conditional bias 47
Chapter 4. Empirical investigations 49
4.1. Design-based approach 49
4.2. Model-based approach 51
4.3. Robust Prediction 57
Conclusion 67
Bibliography 69
Appendix A. ..... A-i
A.1. Proof of conditional bias estimator ..... A-i
A.2. Graphs ..... A-iv
A.3. Parameters used in simulations ..... A-vii

List of tables

3.1 Objective function and corresponding ϕ-function for Huber and bisquare function 44

4.1 Results for the model-based predictors for Mechanisms 1, 2, and 3. 54

4.2 Results for the model-based predictors for Mechanisms 4, 5, and 6. 55

4.3 Results for Population 1 with $p_0 = 0.23$. 62

4.4 Results for Population 2 with $p_0 = 0.23$. 63

4.5 Results for Population 3 with $p_0 = 0.23$. 64

4.6 Results for Population 4 with $p_0 = 0.23$. 65

4.7 Results for Population 5 with $p_0 = 0.23$. 66

List of figures

2.1 Design effect of Bernoulli sampling as a function of the $CV_1(y)$ for $p_0 = 0.5$. 32

2.2 Design effect of Bernoulli sampling as a function of the proportion p_0 for $CV_1(y) = 1$. 33

3.1 Example of inliers and outliers for a linear model. 42

3.2 Huber function with $k = 1.345$. 44

3.3 Bisquare function with $k = 4.685$. 45

3.4 Example of the customary least squares fit (purple) and the fit from a robust method (red) in presence of outliers. 46

4.1 Relationship between y and x under Mechanism 1 50

4.2 Relationship between y and x under Mechanism 2. 50

4.3 RMSE as a function of ϕ_0 for Mechanism 1 (in blue) and Mechanism 2 (in red). . 51

4.4 RMSE as a function of p_0 for Mechanism 1 (in blue) and Mechanism 2 (in red). . 51

4.5 Relationship between y and x under Mechanism 3. 52

4.6 Relationship between y and x under Mechanism 4. 52

4.7 Relationship between y and x under Mechanism 5. 53

4.8 Relationship between y and x under Mechanism 6. 53

4.9 Example of population with Mechanism 1 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right. 57

4.10 Example of population with Mechanism 2 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right. 57

4.11 Example of Population 1 with a normal distribution. 58

4.12 Example of Population 2 with a gamma distribution. 58
4.13 Example of Population 3 with a lognormal distribution. 59
4.14 Example of Population 4 with a Pareto distribution. 59
4.15 Example of Population 5 with a mixture distribution. 59

Remerciements

Tout d'abord, je tiens à remercier mon superviseur, David Haziza pour toute l'aide et tous les conseils qu'il m'a apportés durant ce (long) parcours ainsi que pour l'opportunité qu'il m'a offerte avec un stage à Statistique Canada.

Je remercie également mes parents et mes amies MoJi et Victoire de m'avoir appuyée sans trop me poser de questions.

Introduction

Surveys are regularly conducted to gather information about a certain finite population. In most surveys, information is collected on many variables of interest (also called survey variables or characteristics of interest) and the aim is to estimate many population parameters; such surveys are thus often referred to as multipurpose surveys.

In some surveys, especially in business surveys, it is not unusual to encounter survey variables exhibiting a large proportion of zero-valued observations. For instance, we may be interested in the consumption of propane used by Canadian businesses. For this type of variable, we expect to observe a large proportion of zero-valued observations in the sample as most businesses do not use propane but use another type of energy, such as electricity. Depending on the sampling design, a large proportion of zero-valued observations may lead to unstable estimators of population totals or population means. If the proportion of zero-valued observations is very large, we could obtain a sample with only zero-valued observations which would lead to an estimated total equal to zero when, in fact, the real value for the total is larger than zero.

The objective of this work is to examine the properties of estimators/predictors in the presence of zero-valued observations. In Chapter 1, we introduce the main inferential approaches in survey sampling: the design-based approach and the model-based approach. We also describe some commonly used estimators and predictors such as the Horvitz-Thompson estimator, the ratio estimator, the generalized regression estimator and the best unbiased linear predictor. In Chapter 2, we examine the behavior of these estimators/predictors of population totals in the presence of different proportions of zero-valued observations. We

also discuss the empirical best predictor based on a mixture model that accounts for the zero-valued observations. In Chapter 3, in the context of the model-based approach, we consider the problem of influential units for populations exhibiting a large proportion of zero-valued observations. We describe several robust predictors to the presence of influential units in the sample. In Chapter 4, we conduct several simulation studies to evaluate the performance of customary estimators/predictors with varying proportions of zero-valued observations.

Chapter 1

The design-based approach and the model-based approach

1.1. Finite population and sample

Let us consider a finite population U consisting of N units. We write:

$$U = \{1, ..., i, ..., N\}.$$

For each unit, we collect a survey variable y. The aim is to estimate a finite population parameter, which describes some aspect of the finite population. In this thesis, we focus on the population total of the variable y, t_y.

Most often, conducting a census is not an option due to a lack of resources and time. Instead, it is common practice to select a sample s of size n, to estimate the parameters of interest.

A sample can always be viewed as the result of a two-phase process: first, the finite population U is generated from an infinite population, often referred to as a superpopulation, according to a given model m. For instance, the values $y_1, \ldots, y_N$, of a variable of interest y may be generated from a normal infinite population with mean μ and variance σ^2. Then, a random sample s is selected from the finite population according to a sampling design $p(s)$.

1.2. Sampling design

A sampling design is a function $p(\cdot)$ that assigns to every possible sample s its probability of being selected. Since $p(s)$ is a probability distribution, it must satisfy:

(1) $p(s) \geq 0 \ \ \forall s \in \Omega$,

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

where Ω denotes the set of all the possible samples. A sample is characterized by the vector of sample selection indicators $\boldsymbol{I} = (I_1, \ldots, I_i, \ldots, I_N)^\top$, where

$$I_i = \begin{cases} 1 & \text{if unit } i \in s \\ 0 & \text{otherwise.} \end{cases}$$

Let $\pi_i = P(i \in s) = P(I_i = 1)$ be the first-order inclusion probability of unit i in the sample and $\pi_{ij} = P(i \in s, j \in s) = P(I_i = 1, I_j = 1)$ the second-order inclusion probability of units i and j, $i \neq j$. A basic sampling design is simple random sampling without replacement (SRSWOR). SRSWOR assigns to each of the $\binom{N}{n}$ possible samples of size n the same probability of being selected. That is,

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

As a result, we have $\pi_i = n/N$ for all $i \in U$.

Another simple sampling design is Poisson sampling, which is a random-sized design unlike SRSWOR. Let π_i be the first-order inclusion probability attached to unit i and set prior to sampling, $i = 1, \ldots, N$. Each of the N population units is subject to an independent Bernoulli trial with probability π_i. If the trial results in a success, the unit is included in the sample, otherwise, it is rejected. When $\pi_i = \pi$ for all $i \in U$, Poisson sampling is referred to as Bernoulli sampling. In that case, the probability of selecting a given sample is

$$p(s) = \pi^{n_s}(1-\pi)^{N-n_s},$$

where n_s denotes the random size of s and $\pi = \mathbb{E}(n_s)/N$.

After selecting a sample, there are two main approaches to inference: the design-based approach and the model-based approach. In the former approach, the y-values $\boldsymbol{y} = (y_1, \ldots, y_i, \ldots, y_N)^\top$ are treated as fixed and the inferences are conducted with respect to the sampling design $p(s)$. In the latter approach, the sample s is treated as fixed while the y_i's are random. Inferences are made with respect to a specified model m.

1.3. The design-based approach

For the design-based approach (Lohr, 2009), the subscript p is used to denote expectations and variances with respect to the sampling design.

1.3.1. The Horvitz-Thompson estimator

One of the most common estimator of a population total t_y is the Horvitz–Thompson estimator $\widehat{t}_y^{HT}$, also called expansion estimator. It is defined as

$$\widehat{t}_y^{HT} = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i \in s} d_i y_i, \tag{1.3.1}$$

where $d_i = 1/\pi_i$ denotes the design (or sampling) weight attached to unit i. It is a linear estimator because it is expressed as $\sum_{i \in s} w_i y_i$, where $w_i = d_i$. When $\pi_i > 0$ for all i, $\widehat{t}_y^{HT}$ is design-unbiased for t_y. That is, $\mathbb{E}_p(\widehat{t}_y^{HT}) = t_y$.

For a fixed-sized sampling design, if $y_i = c\pi_i$ for some constant c, we have

$$\widehat{t}_y^{HT} = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i \in s} \frac{1}{\pi_i} c\pi_i = cn = \sum_{i \in U} c\pi_i = t_y.$$

The second to last equality follows from the fact that $\sum_{i \in U} \pi_i = \sum_{i \in U} \mathbb{E}\left(I_i\right) = \mathbb{E}\left(\sum_{i \in U} I_i\right) = \mathbb{E}\left(n\right) = n$.

As a result, both the bias and the variance are equal to zero. Thus, we expect $\widehat{t}_y^{HT}$ to be efficient if there exists a linear relationship between the inclusion probabilities π_i and the variable of interest y, the relationship goes through the origin and the relationship is strong. The design-variance of $\widehat{t}_y^{HT}$ is given by

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

where $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$. For SRSWOR, the variance (1.3.2) reduces to

$$\mathbb{V}_p(\widehat{t}_y^{HT}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}, \tag{1.3.3}$$

where

$$S_y^2 = \frac{1}{N-1} \sum_{i \in U} \left(y_i - \overline{Y}\right)^2.$$

For Bernoulli sampling, (1.3.2) reduces to

$$\mathbb{V}_p\left(\widehat{t}_y^{HT}\right) = N^2 \left(1 - \frac{\mathbb{E}\left(n_s\right)}{N}\right) \frac{\sum_{i \in U} y_i^2}{N\mathbb{E}\left(n_s\right)}. \tag{1.3.4}$$

1.3.2. The ratio estimator

Consider the case of a quantitative auxiliary variable x such that x is observed for all $i \in s$ and $t_x = \sum_{i \in U} x_i$ is known. The ratio estimator of t_y is given by

$$\begin{aligned} \widehat{t}_y^{ra} &= \frac{\widehat{t}_y^{HT}}{\widehat{t}_x^{HT}} t_x, \\ &= \sum_{i \in s} w_i y_i, \end{aligned} \tag{1.3.5}$$

where

$$w_i = d_i \frac{t_x}{\widehat{t}_x^{HT}}.$$

Thus, the ratio estimator belongs to the class of linear estimators. If we apply the weights w_i to the x-variable, we have $\widehat{t}_x^{ra} = \sum_{i \in s} w_i x_i = t_x$. This property is often referred to as the calibration property. When $y_i = cx_i$ for some constant c, we have

$$\widehat{t}_y^{ra} = \sum_{i \in s} w_i y_i = \sum_{i \in s} d_i \frac{t_x}{\widehat{t}_x^{HT}} c x_i = c \frac{t_x}{\widehat{t}_x^{HT}} \widehat{t}_x^{HT} = c t_x = \sum_{i \in U} c x_i = t_y.$$

Thus, we expect the ratio estimator to be efficient if there is a linear relationship between y and x going through the origin and if the relationship is strong. Since the ratio estimator is a non-linear function of x and y, its variance is intractable and we rely on a first-order Taylor expansion to get an approximation of $\mathbb{V}_p\left(\widehat{t}_y^{ra}\right)$. It leads to the following approximate variance:

$$A\mathbb{V}_p\left(\widehat{t}_y^{ra}\right) = \sum_{i \in U} \sum_{j \in U} \Delta_{ij} \frac{E_i}{\pi_i} \frac{E_j}{\pi_j}, \tag{1.3.6}$$

where $E_i = y_i - Rx_i$ with $R = t_y / t_x$.

In the special case of SRSWOR, Expression (1.3.6) reduces to

$$A\mathbb{V}_p\left(\widehat{t}_y^{ra}\right) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_E^2}{n}, \tag{1.3.7}$$

where

$$S_E^2 = \frac{1}{N-1} \sum_{i \in U} \left(y_i - Rx_i\right)^2.$$

For Bernoulli sampling, (1.3.6) reduces to

$$A\mathbb{V}_p\left(\widehat{t}_y^{ra}\right) = N^2 \left(1 - \frac{\mathbb{E}[n_s]}{N}\right) \frac{1}{\mathbb{E}\left(n_s\right)} \frac{N-1}{N} S_E^2. \tag{1.3.8}$$

If we assume that $(N-1) \approx N$, we note that (1.3.7) and (1.3.8) are identical.

1.3.3. The GREG estimator

Let $\boldsymbol{x}_i = (x_{i1}, ..., x_{iq})^\top$ be the vector of the q auxiliary variables available for $i \in s$ and let $\boldsymbol{t}_{\boldsymbol{x}} = (t_{x1}, ..., t_{xq})^\top$ be the vector of totals in the population, which we assume to be known. We assume that the relationship between y and $\boldsymbol{x}$ can be described by

$$m : y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{1.3.9}$$

such that

$$\mathbb{E}_m(\epsilon_i \mid \boldsymbol{x}_i) = 0, \quad \mathbb{E}_m(\epsilon_i \epsilon_j \mid \boldsymbol{x}_i, \boldsymbol{x}_j, i \neq j) = 0, \quad \mathbb{V}_m(\epsilon_i \mid \boldsymbol{x}_i) = \sigma^2 c_i,$$

where $c_i > 0$ is a known coefficient attached to unit i.

The Generalized REGression (GREG) estimator is given by

$$\widehat{t}_y^{GREG} = \sum_{i \in U} \boldsymbol{x}_i^\top \widehat{\mathbf{B}} + \sum_{i \in s} d_i e_i, \tag{1.3.10}$$

where

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

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

$$\widehat{t}_y^{GREG} = \sum_{i \in s} w_i y_i,$$

where

$$w_i = d_i \left\{ 1 + c_i^{-1} \left(\boldsymbol{t}_x - \widehat{\boldsymbol{t}}_x^{HT} \right)^\top \widehat{\boldsymbol{T}}^{-1} \boldsymbol{x}_i \right\}$$

and $\widehat{\boldsymbol{T}} = \sum_{i \in s} d_i \boldsymbol{x}_i c_i^{-1} \boldsymbol{x}_i^\top$. The ratio estimator (1.3.5) is a special case of the GREG estimator with $\boldsymbol{x_i} = x_i$ and $c_i = x_i$.

The GREG estimator is calibrated on the vector of known population totals $\boldsymbol{t}_x$ since

$$\widehat{\boldsymbol{t}}_x^{GREG} = \sum_{i \in s} w_i \boldsymbol{x}_i = \boldsymbol{t}_x.$$

A consequence of this result is the following: if $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ for some $\boldsymbol{\beta}$, then

$$\widehat{t}_y^{GREG} = \sum_{i \in s} w_i y_i = \sum_{i \in s} w_i \boldsymbol{x}_i^\top \boldsymbol{\beta} = \sum_{i \in U} \boldsymbol{x}_i^\top \boldsymbol{\beta} = t_y.$$

That is, the GREG provides a perfect estimate of t_y if there is a perfect linear relationship between y and $\boldsymbol{x}$.

Using a first-order Taylor expansion, the approximate variance of the GREG can be written as

$$A\mathbb{V}_p\left(\widehat{t}_y^{GREG}\right) = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\frac{E_i}{\pi_i}\frac{E_j}{\pi_j}, \tag{1.3.11}$$

where $E_i = y_i - \boldsymbol{x}_i^\top \mathbf{B}$ with

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

For SRSWOR, (1.3.11) reduces to

$$A\mathbb{V}_p\left(\widehat{t}_y^{GREG}\right) = N^2\left(1-\frac{n}{N}\right)\frac{S_E^2}{n}, \tag{1.3.12}$$

where

$$S_E^2 = (N-1)^{-1}\sum_{i\in U}\left(E_i - \overline{E}\right)^2. \tag{1.3.13}$$

For Bernoulli sampling, (1.3.11) reduces to

$$A\mathbb{V}_p\left(\widehat{t}_y^{GREG}\right) = N^2\left(1-\frac{\mathbb{E}(n_s)}{N}\right)\frac{\sum_{i\in U}E_i^2}{N\mathbb{E}(n_s)}. \tag{1.3.14}$$

Proposition 1.3.1. *If there exists a vector of constant* $\boldsymbol{\lambda}$ *such that* $c_i = \boldsymbol{\lambda}^\top \boldsymbol{x}_i$, *then* $\sum_{i\in U} E_i = 0$ *and* S_E^2 *in (1.3.13) reduces to* $S_E^2 = (N-1)^{-1}\sum_{i\in U}E_i^2$.

Proof.

$$\begin{aligned}
\sum_{i\in U} E_i &= \sum_{i\in U}\left(y_i - \boldsymbol{x}_i^\top\mathbf{B}\right)\\
&= \sum_{i\in U} y_i - \sum_{i\in U}\frac{c_i}{c_i}\boldsymbol{x}_i^\top\mathbf{B}\\
&= \sum_{i\in U} y_i - \boldsymbol{\lambda}^\top\sum_{i\in U}\boldsymbol{x}_i c_i^{-1}\boldsymbol{x}_i^\top\mathbf{B}\\
&= \sum_{i\in U} y_i - \boldsymbol{\lambda}^\top\sum_{i\in U}\boldsymbol{x}_i c_i^{-1} y_i\\
&= \sum_{i\in U} y_i - \sum_{i\in U}\left(\boldsymbol{\lambda}^\top\boldsymbol{x}_i\right)c_i^{-1} y_i\\
&= \sum_{i\in U} y_i - \sum_{i\in U} y_i\\
&= 0.
\end{aligned}$$

□

Therefore, if $c_i = \boldsymbol{\lambda}^\top \boldsymbol{x}_i$ and if we assume that $(N-1)/N \approx 1$, then (1.3.12) and (1.3.14) are identical.

1.3.4. Conditional Bias

The conditional bias is a measure of influence of a unit proposed by Moreno Rebollo et al. (1999) in the context of the design-based approach. For an estimator $\widehat{\theta}$, the conditional bias of a sampled unit i ($I_i = 1$) is defined as

$$B_{1i}^{\widehat{\theta}} = \mathbb{E}_p\left(\widehat{\theta} \mid I_i = 1\right) - \mathbb{E}_p\left(\widehat{\theta}\right). \tag{1.3.15}$$

For a non-sampled unit, the conditional bias is defined as

$$B_{0i}^{\widehat{\theta}} = \mathbb{E}_p\left(\widehat{\theta} \mid I_i = 0\right) - \mathbb{E}_p\left(\widehat{\theta}\right). \tag{1.3.16}$$

Since $\mathbb{E}_p(\widehat{\theta}) = \mathbb{E}\left(\mathbb{E}_p\left(\widehat{\theta} \mid I_i\right)\right)$, (1.3.16) can be written as

$$\begin{aligned}
B_{0i}^{\widehat{\theta}} &= \mathbb{E}_p\left(\widehat{\theta} \mid I_i = 0\right) - \mathbb{E}\left(\mathbb{E}_p\left(\widehat{\theta} \mid I_i\right)\right) \\
&= \mathbb{E}_p\left(\widehat{\theta} \mid I_i = 0\right) - \left[\pi_i \mathbb{E}_p\left(\widehat{\theta} \mid I_i = 1\right) + (1-\pi_i)\mathbb{E}_p\left(\widehat{\theta} \mid I_i = 0\right)\right] \\
&= \pi_i\left[\mathbb{E}_p\left(\widehat{\theta} \mid I_i = 0\right) - \mathbb{E}_p\left(\widehat{\theta} \mid I_i = 1\right)\right] \\
&= \pi_i\left(B_{0i}^{\widehat{\theta}} - B_{1i}^{\widehat{\theta}}\right).
\end{aligned}$$

It follows that

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

The conditional bias of either a sampled or a non-sampled unit can be viewed as a measure of its influence and units with a large conditional bias tend to be influential. However, at the estimation stage, only the influence of sampled units can be reduced and nothing can be done for the non-sampled units. Therefore, in the sequel, we focus on the conditional bias of sampled units.

For the Horvitz-Thompson estimator, the conditional bias of a sampled unit is

$$\begin{aligned}
B_{1i}^{HT} &= \mathbb{E}_p\left(\widehat{t}_y^{HT} - t_y \mid I_i = 1\right) \\
&= (d_i - 1)\, y_i + \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) y_j.
\end{aligned} \tag{1.3.17}$$

Note that the conditional bias in (1.3.17) requires the first-order inclusion probability π_i and the second-order inclusion probability π_{ij}. This is due to the fact that $\mathbb{E}_p(I_j|I_i = 1) = \pi_{ij}/\pi_i$.

In the special case of SRSWOR, Expression (1.3.17) reduces to

$$B_{1i}^{HT} = \frac{N}{N-1}\left(\frac{N}{n} - 1\right)\left(y_i - \overline{Y}\right). \tag{1.3.18}$$

Thus, for SRSWOR, unit i has a large influence if its y-value is far from the mean $\overline{Y}$. For Bernoulli sampling, (1.3.17) reduces to

$$B_{1i}^{HT} = \left(\frac{N}{\mathbb{E}\left(n_s\right)} - 1\right) y_i. \tag{1.3.19}$$

Thus, for Bernoulli sampling, unit i has a large influence if its y-value is far from zero.

For the GREG estimator (1.3.10), the conditional bias is approximated by Taylor expansion, which leads to

$$B_{1i}^{ra} \approx (d_i - 1)E_i + \sum_{\substack{j\in U \\ j\neq i}} \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right) E_j,$$

where $E_i = y_i - \boldsymbol{x}_i^\top \mathbf{B}$. If $\sum_{i\in U} E_i = 0$ (which occurs if $c_i = \boldsymbol{\lambda}^\top \boldsymbol{x}_i$), the conditional bias for unit i is given by

$$B_{1i}^{GREG} \approx \begin{cases} \frac{N}{N-1}\left(\frac{N}{n} - 1\right) E_i & \text{for SRSWOR} \\ \left(\frac{N}{\mathbb{E}(n_s)} - 1\right) E_i & \text{for Bernoulli sampling.} \end{cases} \tag{1.3.20}$$

Hence, if we assume that $(N-1)/N \approx 1$, the conditional bias of a sampled unit is the same irrespectively of the sampling design, SRSWOR or Bernoulli sampling. This wasn't the case for the Horvitz-Thompson estimator. From (1.3.20), it follows that a unit has a large influence if it is associated with a large residual E_i.

1.4. Model-based approach

In the context of the model-based approach (Chambers and Clark, 2012), the sampling design does not play an explicit role in the inference unlike in the design-based approach. The population total t_y is now a random variable that we wish to predict. It can be written as

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

The first term is known as it is a function of the sampled observations only. The goal is to predict the second term: the total of the non-sampled units, $\sum_{i\in U-s} Y_i$. To that end,

we postulate a model describing the relationship between the variable of interest y and a vector of auxiliary variables. A commonly used model is the linear regression model given by (1.3.9). The model-based approach is interested in the prediction error $\widehat{t}_y - t_y$ and all the properties are evaluated with respect to the assumed model m.

The Best Linear Unbiased Predictor (BLUP) of t_y is of the form $\widehat{t}_y = \sum_{i \in s} w_i Y_i$, where the weights w_i satisfy:

$$\mathbb{E}_m \left(\widehat{t}_y - t_y \mid s \right) = 0, \tag{1.4.1}$$

$$\mathbb{V}_m \left(\widehat{t}_y - t_y \mid s \right) \leq \mathbb{V}_m \left(\widehat{t}_y^* - t_y \mid s \right), \tag{1.4.2}$$

where $\widehat{t}_y^*$ is any other linear unbiased predictor of t_y. In other words, $\widehat{t}_y$ is model-unbiased and has the smallest variance among all linear unbiased predictors of t_y.

For Model (1.3.9), the weights w_i satisfying (1.4.1) and (1.4.2) are given by

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

The BLUP is then given by

$$\begin{aligned} \widehat{t}_y^{BLUP} &= \sum_{i \in s} w_i Y_i \\ &= \sum_{i \in s} Y_i + \sum_{i \in U - s} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{WLS}, \end{aligned} \tag{1.4.3}$$

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

$$\widehat{\boldsymbol{\beta}}_{WLS} = \left(\sum_{i \in s} \boldsymbol{x}_i c_i^{-1} \boldsymbol{x}_i^\top \right)^{-1} \sum_{i \in s} \boldsymbol{x}_i c_i^{-1} Y_i. \tag{1.4.4}$$

The prediction variance of $\widehat{t}_y^{BLUP}$ is

$$\mathbb{V}_m \left(\widehat{t}_y^{BLUP} - t_y \mid s \right) = \sigma^2 \left\{ \sum_{i \in s} (w_i - 1)^2 c_i + \sum_{i \in U - s} c_i \right\}. \tag{1.4.5}$$

The reader is referred to Chambers and Clark (2012) for a more detailed discussion.

1.4.1. Conditional Bias

The definition of the conditional bias for the model-based approach is different than the one for the design-based approach. For the model-based approach, the concept of conditional

bias was defined by Beaumont et al. (2013). It is given by

$$B_i = \mathbb{E}_m \left(\widehat{t}_y - t_y \mid s, Y_i = y_i \right).$$

For the BLUP, for $i \in s$, we have

$$B_i^{BLUP} = (w_i - 1) \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right). \tag{1.4.6}$$

For $i \in U - s$, we have

$$B_i^{BLUP} = - \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right). \tag{1.4.7}$$

Thus, a sampled unit has a large influence when its weight w_i and/or its residual $\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right)$ is large; see Beaumont et al. (2013). Again, both sampled and non-sampled units may have a large influence on the predictor $\widehat{t}_y^{BLUP}$. However, at the estimation stage, nothing can be done for non-sampled units.

Chapter 2

Inference for populations with a large number of zero-valued observations

In practice, some variables of interest are prone to a large proportion of zero-valued observations. In this case, we can think of the finite population as being generated from a mixture of distributions.

For instance, in the Industrial Consumption of Energy Survey (ICES) conducted at Statistic Canada, we are interested in learning about the consumption of some type of energy (propane, electricity, natural gas, etc.) in the manufacturing sector in Canada. The sample is selected from a sampling frame referred to as the business register, which is the repository of baseline information on enterprises and establishments operating in Canada. With the information from the business register, the enterprises are assigned to strata defined by the type industry, the location and the size of the enterprise (often defined as a function of revenue). The industry of an enterprise is defined by the North American Industry Classification System (NAICS) which employs up to six digits in the most detailed industry level, to classify businesses by type of economic activities. In each stratum, a sample is selected using Bernoulli sampling and the estimates are provided at the industry level. Because some types of energy are rarely used (e.g., propane), it follows that a large proportion of businesses report a value equal to zero when asked about their consumption. For example, in 2015 for miscellaneous manufacturing (NAICS 339), in a sample of 75 enterprises, the proportion of zero-valued observations for electricity, a commonly used energy, was about 19%, whereas it

was equal to 43% for natural gas and 75% for propane. In the bakeries and tortillas manufacturing (NAICS 3118), with a sample size of 74 enterprises, the proportion of zero was about 3% for electricity, 10% for natural gas and 95% for propane.

Zero-valued observations are also very common in audit sampling (Liu et al., 2005). Auditing is the process by which a company's financial records are examined and since companies do their own financial record we might be interested, for example, in estimating the amount subject to sales tax, the amount deductible from income tax and compare them to the company's results in order to detect manipulation or fraud. The sampling unit is typically an invoice and the observed value is the qualified amount, meaning the amount that satisfies tax requirement. Hence, this observed value ranges from zero, when the invoice is not qualified, to the full amount. In that scenario, the proportion of zero-valued observations can be quite large.

As mentioned above, finite populations involving a large number of zero-valued observations may be viewed as being generated from a mixture of populations. Let $U_1 \subset U$, of size N_1 be the population consisting of units with $y > 0$ and $U_0 \subset U$, of size N_0, the population consisting of the units with zero-valued observations. We have that $U = U_1 \cup U_0$ and $N = N_1 + N_0$. From the population U, a sample s of size n is selected and we have $s = s_1 \cup s_0$, where s_1, of size n_1, is the subset of s units exhibiting a strictly positive y-value and s_0, of size n_0, is the subset of s units with a zero-valued observation. We have $n = n_0 + n_1$.

2.1. Design-based approach

2.1.1. The Horvitz-Thompson estimator

In the situation introduced above, the Horvitz-Thompson estimator of t_y in (1.3.1) reduces to

$$\widehat{t}_y^{HT} = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i \in s_0} \frac{1}{\pi_i} y_i + \sum_{i \in s_1} \frac{1}{\pi_i} y_i = \sum_{i \in s_1} \frac{1}{\pi_i} y_i.$$

The design variance of $\widehat{t}_y^{HT}$ given by (1.3.2) reduces to

$$\mathbb{V}_p(\widehat{t}_y^{HT}) = \sum_{i \in U_1} \sum_{j \in U_1} \frac{\Delta_{ij}}{\pi_i \pi_j} y_i y_j. \tag{2.1.1}$$

We now examine the case of SRSWOR. Let $p_0 = N_0/N$ be the proportion of zero-valued observations. By approximating $(N_1 - 1)$ and $(N - 1)$ by N_1 and N respectively, the variance

of $\widehat{t}_y^{HT}$ given by (2.1.1) reduces to

$$\mathbb{V}_p(\widehat{t}_y^{HT}) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n}(1-p_0) \left(S_1^2 + p_0 \overline{Y}_1^2\right), \tag{2.1.2}$$

where

$$S_1^2 = (N_1 - 1)^{-1} \sum_{i \in U_1} \left(y_i - \overline{Y}_1\right)^2$$

and $\overline{Y}_1 = N_1^{-1} \sum_{i \in U_1} y_i$. It is often useful to make use of the design coefficient of variation, which is a standardized measure of variance of an estimator. For a parameter θ and an estimator $\widehat{\theta}$, it is denoted by $CV(\widehat{\theta})$ and defined as

$$CV(\widehat{\theta}) = \frac{\sqrt{\mathbb{V}_p(\widehat{\theta})}}{\mathbb{E}_p(\widehat{\theta})}.$$

Noting that

$$\overline{Y} = \frac{\sum_{i \in U} y_i}{N} = \frac{\sum_{i \in U_1} y_i}{N} = (1-p_0)\overline{Y}_1,$$

it follows from (2.1.2) that the squared coefficient of variation of $\widehat{t}_y^{HT}$ is given by

$$CV^2\left(\widehat{t}_y^{HT}\right) \equiv \frac{\mathbb{V}_p(\widehat{t}_y^{HT})}{t_y^2} \approx \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{1}{(1-p_0)} \left(p_0 + CV_1^2(y)\right), \tag{2.1.3}$$

where $CV_1(y) = \frac{S_1}{\overline{Y}_1}$ denotes the coefficient of variation of the y-variable in the population U_1.

For Bernoulli sampling, the design variance of $\widehat{t}_y^{HT}$ in (2.1.1) simplifies to

$$\mathbb{V}\left(\widehat{t}_y^{HT}\right) \approx N^2 \left(1 - \frac{\mathbb{E}\left(n_s\right)}{N}\right) \frac{1}{\mathbb{E}\left(n_s\right)} (1-p_0)(S_1^2 + \overline{Y}_1^2)$$

and the squared coefficient of variation of $\widehat{t}_y^{HT}$ is given by

$$CV^2\left(\widehat{t}_y^{HT}\right) \approx \left(1 - \frac{\mathbb{E}\left(n_s\right)}{N}\right) \frac{1}{\mathbb{E}\left(n_s\right)} \frac{1}{(1-p_0)} \left(1 + CV_1^2(y)\right). \tag{2.1.4}$$

Expressions (2.1.3) and (2.1.4) suggest that the coefficient of variation increases as $CV_1(y)$ increases for a fixed value of p_0 or increases as p_0 also increases, because of the $\frac{1}{1-p_0}$ term, for a fixed value of $CV_1(y)$.

We now compare Expressions (2.1.3) and (2.1.4). The main difference is p_0 in the last term on the right-hand side of (2.1.3), which is replaced by a one in (2.1.4). The coefficient of variation of $\widehat{t}_y^{HT}$ for SRSWOR is always smaller than the one for Bernoulli sampling if we do not consider the extreme case where all units have a zero-valued observation such that $p_0 = 1$.

To compare both sampling designs, we use the design effect of Bernoulli sampling defined as

$$\text{deff(BE)} = \frac{\mathbb{V}_{\text{BE}}\left(\widehat{t}_y\right)}{\mathbb{V}_{SRSWOR}\left(\widehat{t}_y\right)}.$$

Fig. 2.1 shows the design effect of Bernoulli sampling as a function of $CV_1(y)$ with $p_0 = 0.5$. As the coefficient of variation $CV_1(y)$ increases, the difference between the two sampling designs becomes smaller. When $CV_1(y) \geq 3$, the design effect is very close to 1 and both sampling methods are essentially equivalent in terms of efficiency.

Fig. 2.2 shows the design effect as a function of p_0 with $CV_1(y) = 1$. As p_0 increases,

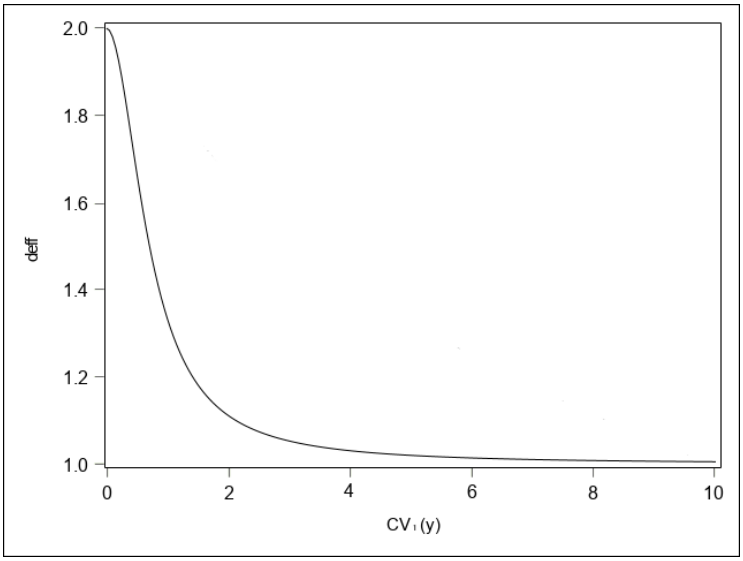


Fig. 2.1. Design effect of Bernoulli sampling as a function of the $CV_1(y)$ for $p_0 = 0.5$.

the difference between both designs becomes smaller and the design effect approaches 1. In the extreme situation where $p_0 = 1$, meaning that all population units have zero-valued observations, both variances are equal and we have $\mathbb{V}_{\text{BE}}\left(\widehat{t}_y\right) = \mathbb{V}_{SRSWOR}\left(\widehat{t}_y\right) = 0$.

Indeed, the conditional bias of a unit $i \in s_0$ is

$$B_{1i}^{HT} = \begin{cases} \frac{-N}{N-1}\left(\frac{N}{n} - 1\right)\overline{Y} & \text{for SRSWOR} \\ 0 & \text{for Bernoulli sampling.} \end{cases}$$

That is, a unit with a zero-valued observation has no influence under Bernoulli sampling, whereas it could have a substantial influence under SRSWOR if the population mean $\overline{Y}$ is far from zero. Therefore, Bernoulli sampling becomes increasingly efficient relative to SRSWOR as the proportion of zero-valued observations, p_0, increases.

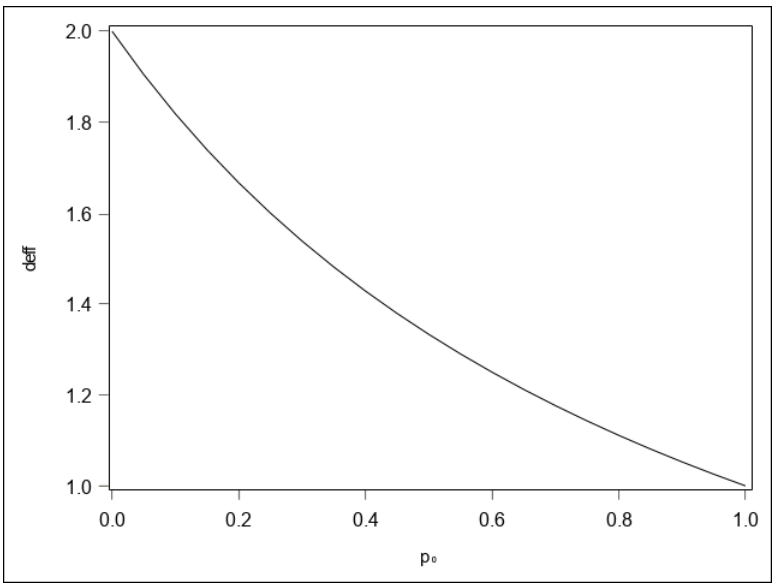


Fig. 2.2. Design effect of Bernoulli sampling as a function of the proportion p_0 for $CV_1(y) = 1$.

2.1.2. The ratio estimator

In the presence of zero-valued observations, the ratio estimator of t_y (1.3.5) reduces to

$$\begin{aligned}\widehat{t}_y^{ra} &= \sum_{i\in s} w_i y_i \\ &= \sum_{i\in s_0} w_i y_i + \sum_{i\in s_1} w_i y_i \\ &= \sum_{i\in s_1} w_i y_i \\ &= \sum_{i\in s_1} d_i \left(\frac{t_x}{\widehat{t}_x^{HT}}\right) y_i.\end{aligned}$$

Let $R_1 = \sum_{i\in U_1} y_i / \sum_{i\in U_1} x_i = t_{y,1}/t_{x,1}$ and $\phi_0 = \sum_{i\in U_0} x_i/t_x = t_{x,0}/t_x$, the fraction of t_x corresponding to population U_0. When N is large, we can safely replace $N-1$ by N and Equations (1.3.7) and (1.3.8) are identical. For both SRSWOR and Bernoulli sampling, the variance of $\widehat{t}_y^{ra}$ reduces to

$$\begin{aligned}A\mathbb{V}_p(\widehat{t}_y^{ra}) \approx N^2 \frac{1-n/N}{n}\Bigg\{&(1-p_0)S_{E_1}^2 + 2(1-p_0)R_1\phi_0 S_{E_1,x} \\ &+ \frac{\phi_0^2 R_1^2 \sum_{i\in U_1} x_i^2}{N-1} + \frac{(1-\phi_0)^2 R_1^2 \sum_{i\in U_0} x_i^2}{N-1}\Bigg\},\end{aligned} \tag{2.1.5}$$

where $S_{E_1}^2 = (N_1 - 1)^{-1} \sum_{i \in U_1} (y_i - R_1 x_i)^2$.

In Chapter 4, we assess the efficiency of $\widehat{t}_y^{ra}$ as a function of p_0 and ϕ_0 through a simulation study.

The influence of a sampled unit having a zero-valued observation can, once again, be measured with the conditional bias from (1.3.20). It reduces to

$$B_{1i}^{ra} \approx \begin{cases} -\frac{N}{N-1}\left(\frac{N}{n} - 1\right) Rx_i & \text{for SRSWOR} \\ -\left(\frac{N}{n} - 1\right) Rx_i & \text{for Bernoulli sampling.} \end{cases}$$

The ratio estimator involves a straight line passing through the origin. Therefore, a sample zero-valued observation with a large x-value will have a large influence as its residual $y_i - Rx_i = -Rx_i$ will be large. On the other hand, a sample zero-valued observation with an x-value close to zero will have virtually no influence.

2.1.3. The GREG estimator

In the presence of zero-valued observations, the GREG estimator of t_y given by (1.3.10) reduces to

$$\begin{aligned} \widehat{t}_y^{GREG} &= \sum_{i \in s} w_i y_i \\ &= \sum_{i \in s_0} w_i y_i + \sum_{i \in s_1} w_i y_i \\ &= \sum_{i \in s_1} w_i y_i \\ &= \sum_{i \in U} \boldsymbol{x}_i^\top \widehat{\mathbf{B}} + \sum_{i \in s_1} d_i e_i - \sum_{i \in s_0} d_i \boldsymbol{x}_i^\top \widehat{\mathbf{B}}. \end{aligned}$$

We focus on the case whereby $c_i = \boldsymbol{\lambda}^\top \boldsymbol{x}_i$ so that $\sum_{i \in U} E_i = 0$. Then, (1.3.12) and (1.3.14) are identical and they reduce to

$$A\mathbb{V}\left(\widehat{t}_y^{GREG}\right) \approx N^2 \frac{1 - n/N}{n} \Bigg\{ (1-p_0) S_{E_1}^2 + \sum_{i \in U_1} \left[\boldsymbol{x}_i^\top (\mathbf{B}_1 - \mathbf{B})\right]^2 + 2 \sum_{i \in U_1} (y_i - \boldsymbol{x}_i^\top \mathbf{B}) \left\{\boldsymbol{x}_i^\top (\mathbf{B}_1 - \mathbf{B})\right\} + \sum_{i \in U_0} \left(\boldsymbol{x}_i^\top \mathbf{B}\right)^2 \Bigg\},$$

where

$$\mathbf{B}_1 = \left(\sum_{i \in U_1} \boldsymbol{x}_i c_i^{-1} \boldsymbol{x}_i^\top \right)^{-1} \sum_{i \in U_1} \boldsymbol{x}_i c_i^{-1} y_i$$

and

$$S_{E_1}^2 = (N_1 - 1)^{-1} \sum_{i \in U_1} (y_i - \boldsymbol{x}_i^\top \mathbf{B}_1)^2.$$

The efficiency of $\widehat{t}_y^{GREG}$ depends, among others, on the difference between $\mathbf{B}_1$ and $\mathbf{B}$, where $\mathbf{B}_1$ is the vector of estimated regression coefficients that would have been obtain had we fitted a linear regression model at the population level, based on the nonzero valued observations only, whereas $\mathbf{B}$ is the vector of estimated regression coefficients based on all population units (zero-valued and nonzero-valued observations). The approximate variance of $\widehat{t}_y^{GREG}$ increases as the difference between $\mathbf{B}_1$ and $\mathbf{B}$ increases. The efficiency also depends on the proportion of nonzero valued observation $(1 - p_0)$. We assess the efficiency of $\widehat{t}_y^{GREG}$ through a simulation study in Chapter 4.

For the GREG estimator, the influence of a sample zero-valued observation from (1.3.20) reduces to

$$B_{1i}^{GREG} \approx \begin{cases} -\frac{N}{N-1}\left(\frac{N}{n} - 1\right) \boldsymbol{x}_i^\top \mathbf{B} & \text{for SRSWOR} \\ -\left(\frac{N}{n} - 1\right) \boldsymbol{x}_i^\top \mathbf{B} & \text{for Bernoulli sampling.} \end{cases}$$

As for the ratio estimator, a sample zero-valued observation will have a large influence if its population fit $\boldsymbol{x}_i^\top \mathbf{B}$ is large.

2.2. Model-based approach

For the model-based approach, the population U can be viewed as being generated from the model

$$m : Y_i = \delta_i(\boldsymbol{x}_i^\top \boldsymbol{\beta}_1 + \epsilon_i) + (1 - \delta_i) \times 0 \tag{2.2.1}$$

where δ_i is an indicator variable associated with unit i such that

$$\delta_i = \begin{cases} 1 & \text{if } y_i > 0 \\ 0 & \text{if } y_i = 0. \end{cases}$$

We assume that

$$\mathbb{E}_m(\epsilon_i \mid \delta_i = 1) = 0, \quad \mathbb{E}_m(\epsilon_i \epsilon_j \mid \delta_i = 1, \delta_j = 1, i \neq j) = 0, \quad \mathbb{V}_m(\epsilon_i \mid \delta_i = 1) = \sigma^2 c_i.$$

In sections 2.2.1 and 2.2.2, we consider two predictors of t_y when the population contains a large number of zero-valued observations. The first predictor is the customary BLUP of t_y given by (1.4.3) derived under model (1.3.9) while the second predictor is based on the mixture model (2.2.1).

2.2.1. Best Linear Unbiased Predictor (BLUP)

A first predictor of t_y is the BLUP given by (1.4.3), based on the customary linear model (1.3.9):

$$\begin{aligned}\widehat{t}_y^{BLUP} &= \sum_{i\in s} w_i Y_i \\ &= \sum_{i\in s_0} w_i Y_i + \sum_{i\in s_1} w_i Y_i \\ &= \sum_{i\in s_1} w_i Y_i \\ &= \sum_{i\in s_1} Y_i + \sum_{i\in U-s} \boldsymbol{x_i}^\top \widehat{\boldsymbol{\beta}}_{WLS}, \end{aligned} \tag{2.2.2}$$

where $\widehat{\boldsymbol{\beta}}_{WLS}$ is given by (1.4.4). A unit in U_1 has the same conditional bias as in (1.4.6) and (1.4.7), depending on whether unit i has been selected in the sample or not. For a unit in U_0, the conditional bias with respect to model (1.3.9) becomes

$$B_i^{BLUP} = \begin{cases} -(w_i - 1)\boldsymbol{x_i}^\top \boldsymbol{\beta} & \text{if } i \in s \\ \boldsymbol{x_i}^\top \boldsymbol{\beta} & \text{if } i \in U_0 - s. \end{cases}$$

Thus, a sampled unit i from U_0 has a large influence if its weight w_i is large and/or the census "fit" $\left(\boldsymbol{x_i}^\top \boldsymbol{\beta}\right)$ is large.

2.2.2. Empirical Best Predictor (EBP)

Under the mixture model (2.2.1), we can derive the Empirical Best Predictor (EBP) of t_y. We start by noting that

$$\begin{aligned}\mathbb{E}_m\left(Y_i\right) &= \mathbb{E}\left(\mathbb{E}\left(Y_i \mid \delta_i\right)\right) \\ &= p_i \mathbb{E}(\boldsymbol{x}_i^\top \boldsymbol{\beta}_1 + \epsilon_i \mid \delta_i = 1) \\ &= p_i \boldsymbol{x}_i^\top \boldsymbol{\beta}_1, \end{aligned}$$

where $p_i = P(\delta_i = 1)$. Also, the variance of Y_i under model (2.2.1) is

$$\begin{aligned}\mathbb{V}_m\left(Y_i\right) &= \mathbb{E}\left(\mathbb{V}\left(Y_i \mid \delta_i\right)\right) + \mathbb{V}\left(\mathbb{E}\left(Y_i \mid \delta_i\right)\right) \\ &= p_i \mathbb{V}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_1 + \epsilon_i \mid \delta_i = 1\right) + p_i\left(1 - p_i\right) \mathbb{E}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_1 + \epsilon_i \mid \delta_i = 1\right)^2 \\ &= p_i \sigma^2 c_i + p_i\left(1 - p_i\right)\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_1\right)^2. \end{aligned} \tag{2.2.3}$$

The EBP of t_y denoted by $\widehat{t}_y^{EBP}$ is given by

$$\widehat{t}_y^{EBP} = \sum_{i \in s_1} Y_i + \sum_{i \in U-s} \widehat{p}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_1, \tag{2.2.4}$$

where $\widehat{p}_i$ and $\widehat{\boldsymbol{\beta}}_1$ are consistent estimators for p_i and $\boldsymbol{\beta}_1$. In this paper, we use the weighted least squares estimator of $\boldsymbol{\beta}_1$:

$$\widehat{\boldsymbol{\beta}}_{1,WLS} = \left(\sum_{i \in s_1} \boldsymbol{x}_i c_i^{-1} \boldsymbol{x}_i^\top \right)^{-1} \sum_{i \in s_1} \boldsymbol{x}_i c_i^{-1} Y_i.$$

The predictor (2.2.4) involves the estimated probabilities $\widehat{p}_i$. To obtain these estimated probabilities, we may postulate a parametric model of the form

$$p_i = P(\delta_i = 1) = \psi(\boldsymbol{x}_i, \boldsymbol{\gamma}),$$

where $\psi(\cdot)$ is a predetermined functional and $\boldsymbol{\gamma}$ is a vector of unknown coefficients. A frequently encountered parametric model is the logistic regression model given by

$$p_i = \frac{\exp\left(\boldsymbol{x}_i^\top \boldsymbol{\gamma}\right)}{1 + \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\gamma}\right)}.$$

The estimated probabilities $\widehat{p}_i$ are given by $\widehat{p}_i = \psi(\boldsymbol{x}_i, \widehat{\boldsymbol{\gamma}})$, where $\widehat{\boldsymbol{\gamma}}$ is a suitable estimator (e.g., the maximum likelihood estimator) of $\boldsymbol{\gamma}$. However, point estimators based on parametric imputation procedures may suffer from bias if the form of the model is misspecified or if the specified $\mathbf{X}$ fails to include interactions or predictors accounting for curvature.

Alternatively, one may use a nonparametric procedure in order to obtain estimates of the p_i's. In contrast to parametric methods, the shape of the relationship is left unspecified. Also, nonparametric procedures have the ability to capture nonlinear trends in the data and tend to be robust to the non-inclusion of interactions or predictors accounting for curvature. However, many traditional nonparametric procedures tend to breakdown when the dimension of the $\mathbf{x}$-vector is large, a problem often referred as the curse of dimensionality.

A simple nonparametric procedure is Kernel Density Estimation (KDE). For simplicity, we assume a scalar x, in which case a KDE of p_i is

$$\widehat{\psi}(x, h) = \frac{1}{nh} \sum_{i \in s} K\left(\frac{x - x_i}{h} \right),$$

where K is the so-called kernel function and $h > 0$ is a smoothing parameter called the bandwidth. Popular kernel functions include the Epanechnikov (Epanechnikov, 1969), Tricube,

and Gaussian kernels. The bandwidth h is a parameter whose value may considerably affect the resulting estimate. The bandwidth is usually selected so as to achieve a compromise between bias and variance. As h increases, $\widehat{\psi}(x,h)$ approaches the constant value $1-n_0/n$, the sample proportion of nonzero-valued observations.

In the context of highly skewed distributions with a large proportion of zero-valued observations, Karlberg (2000) studied the EBP with the p_i's estimated through a logistic regression model and the nonzero-valued observations y_i's modeled by a lognormal distribution.

For simplicity, we assume that the p_i's are known for all i in the sequel. The predictor (2.2.4) can be written as

$$\widehat{t}_y^{EBP} = \sum_{i\in s_1} w_i Y_i,$$

where

$$w_i = 1 + \frac{\boldsymbol{x}_i^\top}{c_i}\left(\sum_{i\in s_1}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{c_i}\right)^{-1}\left(\sum_{i\in U} p_i\boldsymbol{x}_i - \sum_{i\in s} p_i\boldsymbol{x}_i\right).$$

Its prediction variance is given by

$$\begin{aligned}
\mathbb{V}_m\left(\widehat{t}_y^{EBP} - t_y\right) &= \mathbb{V}_m\left(\sum_{i\in s} w_i Y_i - \sum_{i\in U} Y_i\right) \\
&= \mathbb{V}_m\left(\sum_{i\in s} w_i Y_i - \sum_{i\in s} Y_i - \sum_{i\in U-s} Y_i\right) \\
&= \sum_{i\in s}(w_i-1)^2\,\mathbb{V}_m(Y_i) + \sum_{i\in U-s}\mathbb{V}_m(Y_i),
\end{aligned}$$

where $\mathbb{V}_m(Y_i)$ is defined by (2.2.3).

We now compute the conditional bias of unit i. There are four cases to consider: unit i is included or not in the sample and $\delta_i = 1$ or $\delta_i = 0$. If a unit i has a nonzero-valued

observation and is selected in the sample, its conditional bias is given by

$$\begin{aligned}
B_i^{EBP} &= \mathbb{E}_m\left(\widehat{t}_y^{EBP} - t_y \mid s, Y_i = y_i\right) \\
&= \mathbb{E}_m\left(\sum_{j\in s} w_j Y_j - \sum_{j\in U} Y_j \mid s, Y_i = y_i\right) \\
&= \mathbb{E}_m\left(w_i Y_i - Y_i + \sum_{\substack{j\in s \\ j\neq i}} w_j Y_j - \sum_{\substack{j\in U \\ j\neq i}} Y_j \mid s, Y_i = y_i\right) \\
&= (w_i - 1)(y_i - p_i \boldsymbol{x_i}^\top \boldsymbol{\beta}_1) + \sum_{j\in s} w_j p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1 - \sum_{j\in U} p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1.
\end{aligned}$$

If unit i is not included in the sample, its conditional bias is

$$\begin{aligned}
B_i^{EBP} &= \mathbb{E}_m\left(\widehat{t}_y^{EBP} - t_y \mid s, Y_i = y_i\right) \\
&= \mathbb{E}_m\left(\sum_{j\in s} w_j Y_j - \sum_{j\in U} Y_j \mid s, Y_i = y_i\right) \\
&= \mathbb{E}_m\left(-Y_i + \sum_{j\in s} w_j Y_j - \sum_{\substack{j\in U \\ j\neq i}} Y_j \mid s, Y_i = y_i\right) \\
&= -(y_i - p_i \boldsymbol{x_i}^\top \boldsymbol{\beta}_1) + \sum_{j\in s} w_j p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1 - \sum_{j\in U} p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1.
\end{aligned}$$

Therefore, the conditional bias of unit i with respect to EBP is given by

$$B_i^{EBP} = \begin{cases} (w_i - 1)(y_i - p_i \boldsymbol{x_i}^\top \boldsymbol{\beta}_1) + \sum_{j\in s} w_j p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1 - \sum_{j\in U} p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1 & \text{if } i \in s \\ -(y_i - p_i \boldsymbol{x_i}^\top \boldsymbol{\beta}_1) + \sum_{j\in s} w_j p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1 - \sum_{j\in U} p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1 & \text{if } i \in U - s. \end{cases} \tag{2.2.5}$$

In particular when $\delta_i = 0$, (2.2.5) reduces to

$$B_i^{EBP} = \begin{cases} -(w_i - 1) p_i \boldsymbol{x_i}^\top \boldsymbol{\beta}_1 + \sum_{j\in s} w_j p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1 - \sum_{j\in U} p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1 & \text{if } i \in s \\ p_i \boldsymbol{x_i}^\top \boldsymbol{\beta}_1 + \sum_{j\in s} w_j p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1 - \sum_{j\in U} p_j \boldsymbol{x_j}^\top \boldsymbol{\beta}_1 & \text{if } i \in U - s. \end{cases}$$

Thus, a sampled unit i from U_0 has a large influence if its weight w_i is large and/or the census "fit" $\left(p_i \boldsymbol{x_i}^\top \boldsymbol{\beta}_1\right)$ is large. The conditional bias (2.2.5) is unknown and must be estimated, this is considered in the next section.

2.2.3. Estimation of the conditional bias

The conditional bias B_i^{BLUP} in (1.4.6) and (1.4.7) are unknown since $\boldsymbol{\beta}$ is unknown. Therefore, we must estimate it. We note that it is not possible to estimate the conditional bias of a unsampled unit as its y-value is not observed. For a sample unit, the conditional bias B_i^{BLUP} can be estimated by replacing the unknown $\boldsymbol{\beta}$ by an estimator $\tilde{\boldsymbol{\beta}}$:

$$\widehat{B}_i^{BLUP} = \begin{cases} (w_i - 1)(y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) & \text{if } i \in s_1 \\ -(w_i - 1)\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} & \text{if } i \in s_0. \end{cases} \tag{2.2.6}$$

If $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{LS}$, then $\widehat{B}_i^{BLUP}$ is unbiased for B_i^{BLUP}. If $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{LS}^{(-i)}$, then $\widehat{B}_i^{BLUP}$ is conditionally unbiased for B_i^{BLUP} in the sense that

$$\mathbb{E}_m(\widehat{B}_i^{BLUP} | s, Y_i = y_i) = B_i^{BLUP}, \tag{2.2.7}$$

where $\widehat{\boldsymbol{\beta}}_{LS}^{(-i)}$ is the least squares estimator calculated without unit i. The proof is given in the Appendix.

For the conditional bias of the EBP given in (2.2.5) and (2.2.2), both $\boldsymbol{\beta}_1$ and p_i's are unknown. As before, we can only estimate the conditional bias of a unit selected in the sample. The conditional bias B_i^{EBP} can be estimated by replacing $\boldsymbol{\beta}_1$ and the p_i's with some estimator $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{p}_i$, respectively.

Chapter 3

Robust prediction

In this chapter, we focus on the model-based approach, whereby the observations are assumed to have been generated from a given model. In some cases, one must face the presence of outliers in the sample. Outliers correspond to observations that have been generated from a model different from the one that has generated the majority of the observations. The non-outliers are often referred to as inliers. An outlier may be due to a measurement error or it may be a legitimate observation. In this chapter, we focus on the latter case. Measurement errors are typically detected and corrected at the editing stage. Figure 3.1 shows the relationship between a variable y and an auxiliary variable x. From Figure 3.1, most of the observations follow a linear model (in blue), while a few of them are outliers (in red).

Chambers (1986) identified two types of outliers. The first type is called nonrepresentative. The latter can either be due to an error or it is believed to be unique. The second type is a representative outlier, which is a valid sample observation that "represents" other similar units in the non-sampled part of the population. If a nonrepresentative outlier is deemed legitimate, it makes sense to assign it a weight equal to 1. For a representative outlier, the situation is more intricate as we do not know how many observations this outlier represents in the non-sampled part of the population. As a result, it is harder to decide which weight to attribute to this outlier. Representative outliers may be influential in the sense that including or excluding this unit from the sample may have a drastic impact on the estimate. Influential units make the usual predictors (e.g., the BLUP or the EBP) unstable (i.e., exhibiting a large prediction variance).

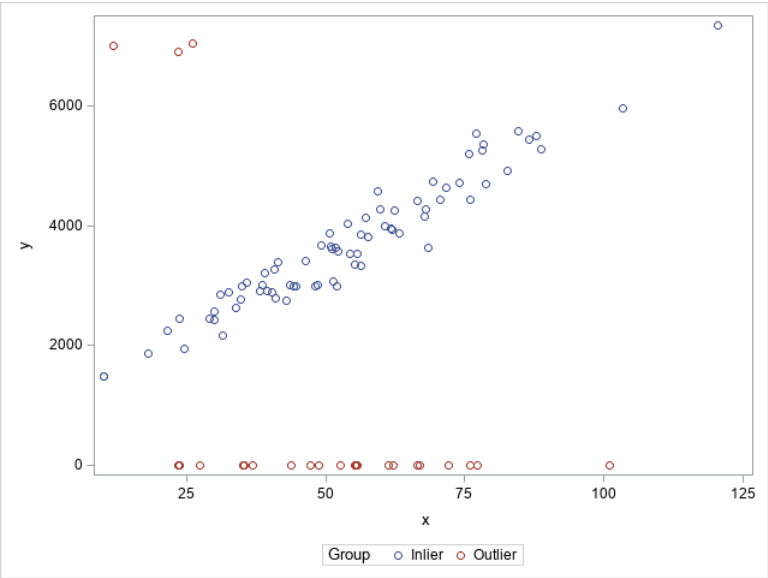


Fig. 3.1. Example of inliers and outliers for a linear model

In this chapter, we are interested in predicting the population total of a survey variable y in the presence of influential units in the sample. More specifically, we describe some robust predictors of t_y in the presence of influential units. By robust, we mean a predictor whose efficiency is close to that of the optimal estimator (e.g., the BLUP) when the model holds but whose efficiency is not affected by a small deviation from the model. A small deviation from the model corresponds to a small proportion of observations that do not follow the model that generated the rest of the observations or to having the first two moments of the model correctly specified but that the distribution of the errors is highly skewed. A robust predictor is expected to exhibit a mean square error smaller than that of a non-robust predictor when influential units are present in the sample. This is achieved at the expense of introducing a bias.

3.1. Robust regression

A naive approach to robust prediction of t_y is to replace the weighted least squares estimator $\widehat{\boldsymbol{\beta}}_{WLS}$ in (1.4.3) by a robust estimator $\widehat{\boldsymbol{\beta}}_R$. A naive predictor $\widehat{t}_y^R$ is thus defined as

$$\widehat{t}_y^R = \sum_{i \in s} Y_i + \sum_{i \in U-s} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_R. \tag{3.1.1}$$

There exist many approaches for robust estimation of $\boldsymbol{\beta}$. Three of the most common ones are M-estimation, least-trimmed squares estimation and MM-estimation.

The least squares estimator $\widehat{\boldsymbol{\beta}}_{LS}$ is obtained by minimizing the sum of the squared errors $e_i = Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_{LS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i \in s} \left(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2. \tag{3.1.2}$$

The rationale behind M-estimation is to minimize an alternative function whose role is to reduce the influence of units exhibiting large residuals (Huber, 1981). M-estimation can be viewed as a generalization of maximum-likelihood estimation. The estimator of $\boldsymbol{\beta}$ is determined by minimizing a function ρ, called the objective function:

$$\widehat{\boldsymbol{\beta}}_M = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i \in s} \rho(e_i) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i \in s} \rho(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

where ρ is nonnegative, monotone, symmetric and behaves like the identity function around the origin. An example of such function is $\rho(e_i) = e_i^2$, which gives back the least squares estimator (3.1.2). Let $\psi = \rho'$ be the derivative of ρ. By taking the derivative of the objective function with respect to $\boldsymbol{\beta}$ and setting every partial derivatives to zero, we get a system of estimating equations to obtain the coefficients:

$$\sum_{i \in s} \psi(e_i) \boldsymbol{x}_i = 0.$$

Two of the most commonly used ρ-functions in a classical setup are the Huber function and bisquare function, also called Tukey's biweight. The objective functions and the derivative associated with these functions are given in Table (3.1). Both depend on k, a tuning constant. A small value of k makes the predictor more resistant to outliers but may lead to a loss of efficiency when the errors are normally distributed. The tuning constant is usually chosen so as to achieve 95% efficiency under the normal model. The value is usually set to $k = 1.345\sigma$ for the Huber function and to $k = 4.685\sigma$ for the bisquare function. These functions are illustrated in Figure 3.2 and Figure 3.3.

Another approach for obtaining a robust estimator of $\boldsymbol{\beta}$ is least-trimmed squares (LTS) regression. We consider the ordered residuals of the sample from smallest to largest:

$$e_{(1)}, \ldots, e_{(n)}.$$

The LTS estimator $\widehat{\boldsymbol{\beta}}_{LTS}$ minimizes the sum of the smallest m (say) squared residuals:

$$\widehat{\boldsymbol{\beta}}_{LTS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{m} e_{(i)}^2.$$

Method	Objective Function	ϕ-Function
Huber	$\rho_H(e) = \begin{cases} \frac{1}{2}e^2 & \text{for } \lvert e\rvert \le k \\ k\lvert e\rvert - \frac{1}{2}k^2 & \text{for } \lvert e\rvert > k. \end{cases}$	$\psi_H(e) = \begin{cases} k & \text{for } e > k \\ e & \text{for } \lvert e\rvert \le k \\ -k & \text{for } e < k. \end{cases}$
Bisquare	$\rho_B(e) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{e}{k}\right)^2\right]^3\right\} & \text{for } \lvert e\rvert \le k \\ \frac{k^2}{6} & \text{for } \lvert e\rvert > k. \end{cases}$	$\psi_B(e) = \begin{cases} e\left[1 - \left(\frac{e}{k}\right)^2\right]^2 & \text{for } \lvert e\rvert \le k \\ 0 & \text{for } \lvert e\rvert > k \end{cases}$

Tab. 3.1. Objective function and corresponding ϕ-function for Huber and bisquare function

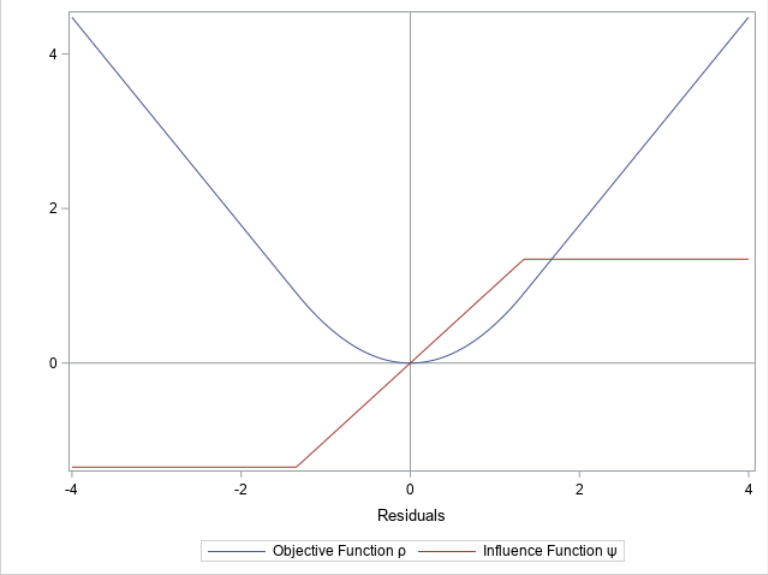


Fig. 3.2. Huber function with $k = 1.345$.

In this case, the units having the m smallest residuals are considered to be the inliers and when $m = n$, we get back the least squares estimator.

Finally, the last method is MM-estimator (Yohai, 1987) which is obtained in three stages:

(1) Compute an initial robust estimate of $\boldsymbol{\beta}$ denoted by $\widehat{\boldsymbol{\beta}}_0$. The initial estimator should have a high breakdown point but may possibly suffer from a low efficiency. The breakdown point of an estimator is the minimum proportion of incorrect observations leading to the breakdown of the estimator.

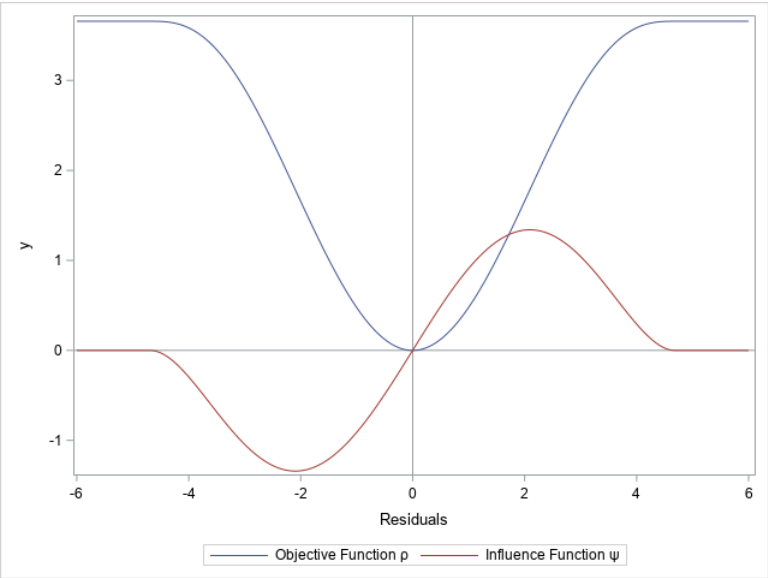


Fig. 3.3. Bisquare function with $k = 4.685$.

(2) Compute a robust estimate of σ, which is the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_0}{\widehat{\sigma}}\right) = \mathbb{E}_{\Phi}\left\{\rho(e)\right\},$$

where Φ is the standard normal distribution.

(3) Let ρ_1 be another objective function and let $\psi_1 = \rho_1'$. The MM-estimator $\widehat{\boldsymbol{\beta}}_{MM}$ is defined as the solution of

$$\sum_{i=1}^{n}\psi_1\left(\frac{Y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\widehat{\sigma}}\right)\boldsymbol{x}_i = 0,$$

such that

$$\sum_{i=1}^{n}\rho_1\left(\frac{Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_{MM}}{\widehat{\sigma}}\right) \leq \sum_{i=1}^{n}\rho_1\left(\frac{Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_0}{\widehat{\sigma}}\right).$$

The predictor (3.1.1) based on a robust estimator of $\boldsymbol{\beta}$ is called naive because it is essentially designed to deal with nonrepresentative outliers. If the sample contains nonrepresentative outliers only, we expect the naive predictor to perform very well in terms of mean squared error since it reduces the influence of these unique observations on the estimation of $\boldsymbol{\beta}$. However, if the outliers are representative, then the naive predictor may be substantially biased. In Fig 3.4, the purple line corresponds to the customary least squares fit. It is clear that the line is highly influenced by the outliers (represented by the red crosses). Predictions based on the least squares lines are poor for most of the observations. The red line

corresponds to the fit obtained by a robust method (M-estimation). It is clear that for the outlying observations, the predictions based on the robust line would be too small. If these outliers are representative, we expect the naive predictor (3.1.1) to be biased negatively. A solution to this problem was first suggested by Chambers (1986).

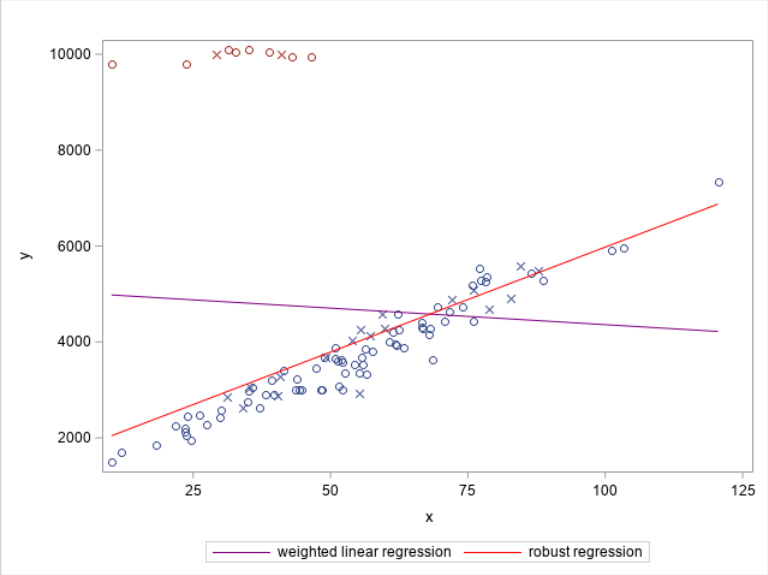


Fig. 3.4. Example of the customary least squares fit (purple) and the fit from a robust method (red) in presence of outliers.

3.2. Predictor of Chambers

Chambers (1986) suggests a bias-adjusted robust predictor:

$$\widehat{t}_y^C(k,c) = \widehat{t}_y^R(k) + \sum_{i\in s}(w_i - 1)\widehat{\sigma}_i\psi_2\left(\frac{Y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_R}{\widehat{\sigma}_i}; c\right), \tag{3.2.1}$$

where ψ_2 is a ψ-function based on the tuning constant c and on $\widehat{\sigma}_i$, which is a robust estimator of σ. The estimator $\widehat{\boldsymbol{\beta}}_R$ in (3.2.1) denotes any robust estimator of $\boldsymbol{\beta}$ based on a ψ-function ψ_1 and tuning constant k. The first term on the right hand-side of (3.2.1) is the naive robust prediction, whereas the second term can be viewed as a bias correction term. The tuning constant k is selected to obtain a highly efficient estimator, whereas the tuning constant c should be large enough. Chambers (1986) advocated a value of c lying between 4 and 6. When $c = 0$, $\widehat{t}_y^C(k,c)$ reduces to

$$\widehat{t}_y^C(k,0) = \widehat{t}_y^R(k),$$

the naive robust predictor, which is expected to be stable but biased. When $c = \infty$, $\widehat{t}_y^C(k,c)$ reduces to

$$\widehat{t}_y^C(k,\infty) = \widehat{t}_y^{BLUP},$$

which is unbiased but unstable. Therefore, c is chosen to achieve a good compromise between bias and variance. This is often referred to as the bias-variance trade-off.

3.3. Predictor based on the conditional bias

Beaumont, Haziza and Ruiz-Gazen (2013) suggested an alternative robust predictor based on the concept of conditional bias. It is defined as

$$\widehat{t}_y^{CB}(c) = \widehat{t}_y^{BLUP} - \sum_{i\in s} \widehat{B}_i^{BLUP} + \sum_{i\in s} \psi(\widehat{B}_i^{BLUP}; c), \tag{3.3.1}$$

where $\widehat{B}_i^{BLUP}$ is an estimator of the conditional bias defined in (2.2.6) and ψ is usually the Huber function with a tuning constant c. An alternative expression of (3.3.1) is given by

$$\widehat{t}_y^{CB}(c) = \sum_{i\in s} Y_i + \sum_{i\in U-s} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} + \sum_{i\in s} \psi\left\{(w_i - 1)(Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}); c\right\}.$$

When the tuning constant $c = 0$, the estimator $\widehat{t}_y^{CB}(c)$ reduces to

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

which depends on the choice of $\widetilde{\boldsymbol{\beta}}$. If $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_R$, we obtain the naive predictor $\widehat{t}_y^R$, which is stable but biased. When $c = \infty$, the estimator $\widehat{t}_y^{CB}(c)$ reduces to

$$\widehat{t}_y^{CB}(\infty) = \widehat{t}_y^{BLUP},$$

which is unbiased but unstable. To choose the tuning constant c, Beaumont et al. (2013) suggest to select the value of c that minimizes the largest absolute estimated conditional bias of $\widehat{t}_y^{CB}(c)$. That is, we seek the value of c that minimizes

$$\max\left\{|\widehat{B}_i^{CB}(c)|\right\},$$

where $\widehat{B}_i^{CB}(c)$ is an estimator of the conditional bias attached to unit i of $\widehat{t}_y^{CB}$. By rewriting (3.3.1) as

$$\widehat{t}_y^{CB}(c) = \widehat{t}_y^{BLUP} - \Delta(c),$$

where $\Delta(c) = \sum_{i\in s} \widehat{B}_i - \sum_{i\in s} \psi(\widehat{B}_i; c)$, we have

$$\begin{aligned} B_i^{BC}(c) &= \mathbb{E}_m\left(\widehat{t}_y^{BC} - t_y | s, Y_i = y_i\right) \\ &= \mathbb{E}_m\left(\widehat{t}_y^{BLUP} - \Delta(c) - t_y | s, Y_i = y_i\right) \\ &= \mathbb{E}_m\left(\widehat{t}_y^{BLUP} - t_y - \Delta(c) | s, Y_i = y_i\right) \\ &= B_i - \mathbb{E}_m\left(\Delta(c) | s, Y_i = y_i\right). \end{aligned}$$

A conditionally unbiased estimator of $B_i^{BC}(c)$ is then given by

$$\widehat{B}_i^{BC}(c) = \widehat{B}_i - \Delta(c).$$

Therefore, we want to minimize $\max\left\{|\widehat{B}_i - \Delta(c)|\right\}$. Beaumont et al. (2013) showed the solution to be

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

where $\widehat{B}_{min}$ and $\widehat{B}_{max}$ are the smallest and largest estimated conditional bias in the sample of the nonrobust estimator $\widehat{t}_y^{BLUP}$. This leads to the robust predictor based on conditional bias:

$$\widehat{t}_y^{CB}(c_{opt}) = \widehat{t}_y^{BLUP} - \frac{1}{2}\left(\widehat{B}_{min}^{BLUP} + \widehat{B}_{max}^{BLUP}\right). \tag{3.3.2}$$

Unlike in robust statistics, the cut-off value c_{opt} is adaptative in the sense that it depends on the sample size. That is, it increases as the sample size increases, which is a desirable property.

In the presence of zero-valued observations, a robust version of the EBP given by (2.2.4) can be obtained in a similar fashion. This leads to

$$\widehat{t}_y^{CB}(c_{opt}) = \widehat{t}_y^{EBP} - \frac{1}{2}\left(\widehat{B}_{min}^{EBP} + \widehat{B}_{max}^{EBP}\right). \tag{3.3.3}$$

In this context, a naive EBP is obtained by replacing $\boldsymbol{\beta}_1$ with a robust estimator $\widehat{\boldsymbol{\beta}}_{1R}$. This leads to

$$\widehat{t}_y^{R} = \sum_{i\in s_1} Y_i + \sum_{i\in U-s} \widehat{p}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{1R}.$$

The performance of these predictors in terms of bias and efficiency will be assessed in Chapter 4.

Chapter 4

Empirical investigations

In this section, we present the results from three simulation studies: the first assesses the performance of design-based estimators in terms of bias and efficiency while the second investigates the performance of predictors in a model-based framework. The third compares several robust predictors presented in Chapter 3 in the presence of influential units, again in terms of bias and efficiency.

4.1. Design-based approach

We generated several populations of size $N = 1000$ consisting of a single auxiliary variable x and a survey variable y. The x-variable was first generated from a Gamma distribution with shape parameter equal to 2 and scale parameter equal to 5. Each unit was assigned to either the population of zero-valued observations, U_0, or to the population of nonzero-valued observations, U_1. To that end, we performed 1000 Bernoulli trials with probability p_i. That is, we generated the indicator variables δ_i from a Bernoulli distribution with probability p_i. If a trial was a success the unit was assigned to the population U_1. The p_i's were generated according to a logistic model:

$$p_i = \frac{\exp(\gamma_0 + \gamma_1 x_i)}{1 + \exp(\gamma_0 + \gamma_1 x_i)},$$

where the values of γ_0 and γ_1 were set so as (i) to obtain an overall proportion, p_0, of zero-valued observations from 0.1 to 0.9 and (ii) for each value of p_0, the probability p_i was either increasing with x_i (i.e., larger values of x_i exhibited a larger proportion of zero-valued observations), which will be called Mechanism 1 below, or decreasing with x_i (i.e., smaller

values of x_i exhibited a larger proportion of zero-valued observations), which will be called Mechanism 2 below.

Given the x-values, for the units belonging to U_1, the y-values were generated according to the so-called ratio model:

$$y_i = 10x_i + \sqrt{x_i}\epsilon_i, \tag{4.1.1}$$

where the errors ϵ_i were generated from a normal distribution with mean equal to zero and variance equal to 25.

Figures 4.1 and 4.2 show simultaneously the relationship between y and x and the relationship between p_i and x_i for an overall proportion p_0 of zero-valued observations equal to 40% for Mechanisms 1 and 2, respectively.

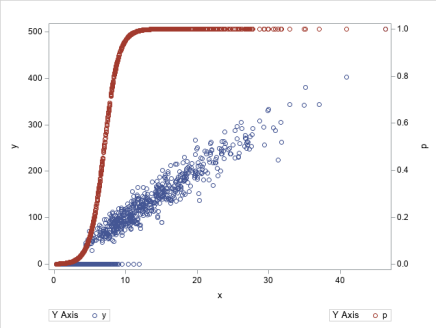


Fig. 4.1. Relationship between y and x under Mechanism 1

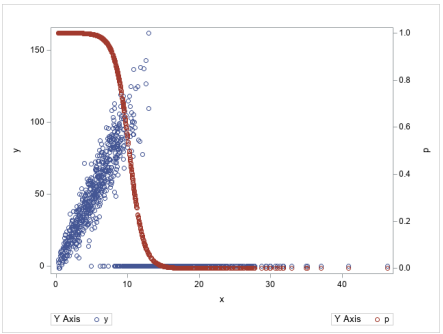


Fig. 4.2. Relationship between y and x under Mechanism 2

From each population, we selected $R = 1000$ samples, of size $n = 100$, according to simple random sampling without replacement. In each sample, we computed the ratio estimator given by (1.3.5).

The Monte Carlo average of the estimator $\widehat{t}_{y,r}^{ra}$ is defined as

$$\mathbb{E}_{MC}(\widehat{t}_y^{ra}) = \frac{1}{R}\sum_{r=1}^{R}\widehat{t}_{y,r}^{ra}, \tag{4.1.2}$$

where $\widehat{t}_{y,r}^{ra}$ denotes the ratio estimator in the rth iteration, $r = 1, \ldots, R$. We computed the relative mean squared error (RMSE) of $\widehat{t}_{y,r}^{ra}$ defined as

$$\text{RMSE}(\widehat{t}_y^{ra}) = \frac{\text{MSE}_{MC}(\widehat{t}_y^{ra})}{t_y}, \tag{4.1.3}$$

where

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

Figure 4.3 shows the RMSE of $\widehat{t}_y^{ra}$ as a function of ϕ_0 (see Equation (2.1.5)) for Mechanisms 1 and 2, whereas Figure 4.4 shows the RMSE as function of p_0.

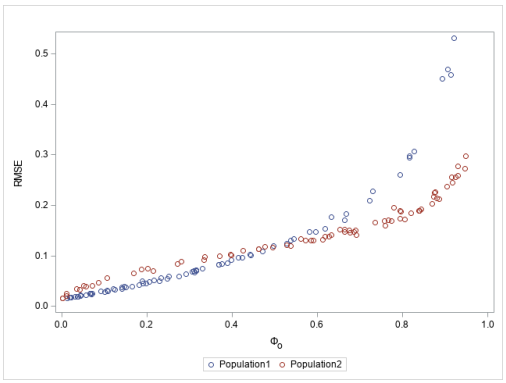
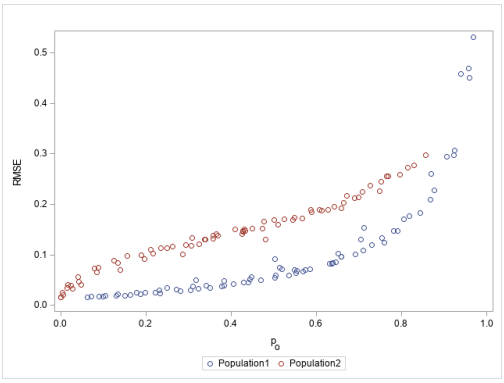


Fig. 4.3. RMSE as a function of ϕ_0 for Mechanism 1 (in blue) and Mechanism 2 (in red).



.png

Fig. 4.4. RMSE as a function of p_0 for Mechanism 1 (in blue) and Mechanism 2 (in red).

From Figure 4.3, we note that, as ϕ_0 increases, the RMSE of $\widehat{t}_y^{ra}$ increases for both mechanisms. The same is true in Figure 4.4, where the RMSE increases as the proportion of zero-valued observations, p_0, increases. This can be explained by the fact that, as p_0 increases, we are getting farther from the ratio model, which assumes a straight line that passes through the origin, making the ratio estimator less efficient.

4.2. Model-based approach

We conducted a model-based simulation to assess the performance of the BLUP and the EBP in terms of bias and efficiency. For each scenario, we repeated $R = 1{,}000$ iterations of the following process:

(i) A finite population of size $N = 1{,}000$ was generated. First, the sub-populations U_0 and U_1 were generated according to six mechanisms. In addition to Mechanisms 1 and 2 used in Section 4.1, we used the following four additional mechanisms:

(3) $p_i = \gamma_0$;

(4) $p_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2$;

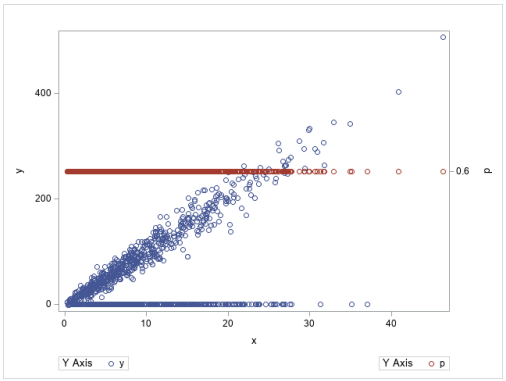


Fig. 4.5. Relationship between y and x under Mechanism 3.

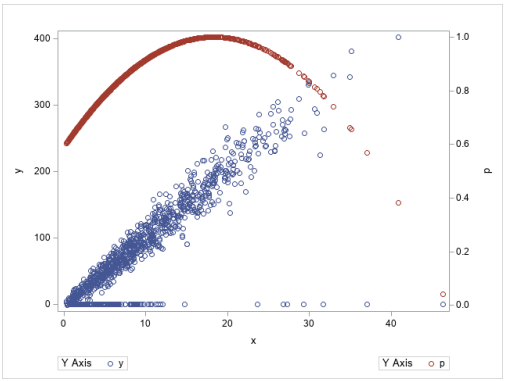


Fig. 4.6. Relationship between y and x under Mechanism 4.

(5) $p_i = \frac{\exp\left[\gamma_0+\gamma_1 x_i+\gamma_2 x_i^2+\gamma_3\cos(\gamma_4 x_i^2)\sin(\gamma_5 x_i)\right]}{1+\exp\left[\gamma_0+\gamma_1 x_i+\gamma_2 x_i^2+\gamma_3\cos(\gamma_4 x_i^2)\sin(\gamma_5 x_i)\right]}$;

(6) $p_i = |\cos(\gamma_0 + \gamma_1 x_i)|$.

For all mechanisms the parameters $\gamma_0, \ldots, \gamma_5$ were set so as to obtain an overall proportion, p_0, of zero-valued observations equal to $0.1, 0.3, 0.5, 0.7$, and 0.9.

Again, for the units in U_1, the y-values were generated according to the ratio model given by (4.1.1). Figures 4.5-4.8 show simultaneously the relationship between y and x and the mechanism for generating the p_i's for Mechanisms 3-6.

(ii) From the finite population generated in Step (i), a sample of size $n = 100$ was selected according to simple random sampling without replacement.

(iii) In each sample, we computed the following predictors of t_y:

(a) the BLUP given by (2.2.2).

(b) the EBP given by (2.2.4), where the probabilities $p_i = P(\delta_i = 1)$ were estimated using three procedures: a logistic regression model, the overall proportion of nonzero-valued observations, $1 - n_0/n$, and a kernel density estimator based on the Gaussian kernel, where the bandwidth h was defined as $h = H \times \text{range}(x)$, and the values of H were set to 0.1, 0.2, and 0.5.

Let $\widehat{t}_y$ be a generic notation. We computed the Monte Carlo percent relative bias (RB) defined as

$$\text{RB}_{MC}\left(\widehat{t}_y\right) = 100 \times \mathbb{E}_{MC}\left(\frac{\widehat{t}_y - t_y}{t_y}\right). \tag{4.2.1}$$

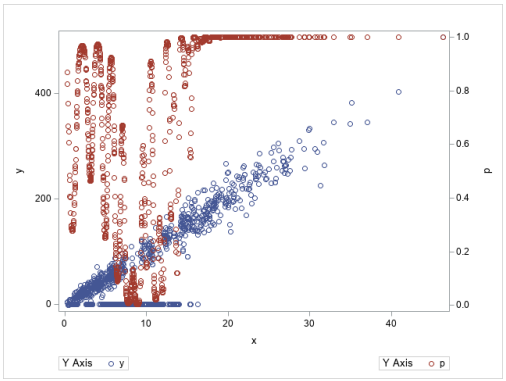


Fig. 4.7. Relationship between y and x under Mechanism 5.

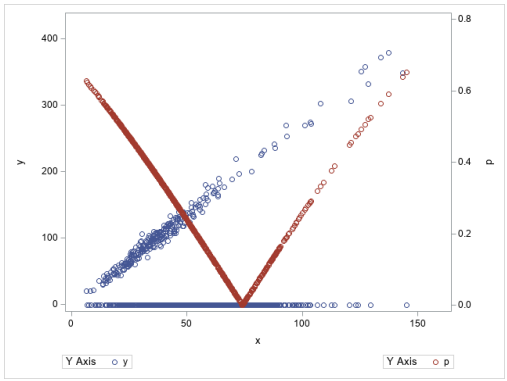


Fig. 4.8. Relationship between y and x under Mechanism 6.

We also computed a measure of relative efficiency (RE) defined as

$$\text{RE}\left(\widehat{t}_y^{EBP}\right) = 100 \times \frac{\text{MSE}_{MC}\left(\widehat{t}_y^{EBP}\right)}{\text{MSE}_{MC}\left(\widehat{t}_y^{BLUP}\right)}. \tag{4.2.2}$$

				Mech.1		Mech.2		Mech.3	
$\widehat{t}_y$	$\widehat{p}_i$	H	p_0	RB	RE	RB	RE	RB	RE
BLUP			0.1	0.0	100	0.8	100	-0.2	100
			0.3	0.0	100	0.6	100	-0.3	100
			0.5	-0.3	100	0.5	100	-0.1	100
			0.7	-0.6	100	0.0	100	-0.2	100
			0.9	-1.7	100	0.1	100	-0.1	100
EBP	Logistic		0.1	0.0	96	0.1	29	-0.3	108
			0.3	0.0	82	-0.1	34	-0.3	104
			0.5	0.0	59	-0.2	51	-0.1	103
			0.7	0.0	35	-0.2	75	-0.2	104
			0.9	0.0	25	0.6	87	0.4	107
EBP	Constant		0.1	-7.1	2065	17.4	572	-0.2	74
			0.3	-19.3	6276	51.1	2833	-0.1	71
			0.5	-30.8	5378	91.0	6066	0.0	70
			0.7	41.0	2574	154.9	8912	-0.1	74
			0.9	-54.8	606	291.2	8338	-0.2	69
EBP	NP	0.1	0.1	-1.8	219	-1.1	33	0.0	110
			0.3	-3.7	248	-0.4	34	0.0	113
			0.5	-2.1	85	7.4	92	0.1	116
			0.7	0.6	34	26.0	402	0.3	115
			0.9	4.7	29	88.8	837	-0.5	104
EBP	NP	0.2	0.1	-2.5	383	-1.9	47	-0.1	108
			0.3	-5.9	647	1.9	41	0.0	105
			0.5	-5.6	247	17.9	285	0.0	110
			0.7	-1.2	36	51.8	1080	-0.1	110
			0.9	7.8	45	129.6	1697	-0.1	104
EBP	NP	0.5	0.1	-3.2	478	-1.8	51	-0.1	116
			0.3	-7.1	908	5.0	65	0.0	103
			0.5	-7.7	417	24.1	476	-0.1	108
			0.7	-3.4	52	61.5	1498	0.4	107
			0.9	7.7	51	146.5	2154	0.2	102

Tab. 4.1. Results for the model-based predictors for Mechanisms 1, 2, and 3.

				Mech.4		Mech.5		Mech.6	
$\widehat{t}_y$	$\widehat{p}_i$	H	p_0	RB	RE	RB	RE	RB	RE
BLUP			0.1	0.1	100	0.0	100	-0.1	100
			0.3	0.2	100	-0.3	100	0.6	100
			0.5	0.6	100	-0.2	100	0.5	100
			0.7	1.6	100	-0.1	100	0.5	100
			0.9	1.2	100	-1.7	100	-0.3	100
EBP	Logistic		0.1	0.2	97	0.0	99	-0.2	103
			0.3	0.2	97	-0.3	87	0.5	99
			0.5	0.6	101	0.0	74	0.4	101
			0.7	1.7	103	-0.2	52	-0.2	98
			0.9	5.3	120	-0.2	30	-1.1	98
EBP	Constant		0.1	-2.8	140	-1.2	126	2.5	79
			0.3	-3.7	178	-6.6	228	5.5	99
			0.5	-5.1	221	-15.9	472	9.0	121
			0.7	-8.4	328	-32.2	922	16.6	168
			0.9	-12.2	121	-55.8	588	26.9	137
EBP	NP	0.1	0.1	-0.1	100	0.9	83	0.0	118
			0.3	-0.1	101	2.8	79	0.2	109
			0.5	0.0	102	3.3	61	0.7	116
			0.7	0.2	94	3.4	50	-0.4	106
			0.9	-0.1	116	4.7	25	3.1	104
EBP	NP	0.2	0.1	-0.7	104	1.2	88	-0.2	116
			0.3	-0.8	107	3.3	90	0.0	105
			0.5	-1.0	107	3.8	71	0.7	109
			0.7	-1.4	103	4.0	57	-0.3	102
			0.9	-2.0	113	7.8	42	6.7	108
EBP	NP	0.5	0.1	-0.8	97	0.9	86	-0.2	111
			0.3	-0.9	100	2.5	81	-0.1	102
			0.5	-1.1	99	2.3	63	0.6	105
			0.7	-1.4	93	2.1	51	-0.3	98
			0.9	-1.3	113	7.6	48	7.3	100

Tab. 4.2. Results for the model-based predictors for Mechanisms 4, 5, and 6.

Tables 4.1 and 4.2 show the Monte Carlo RB and RE for the BLUP and EBP for each scenario.

As expected, the BLUP $\widehat{t}_y^{BLUP}$ showed negligible bias in all scenarios. For the EBP, the bias varied depending on the method used for estimating the p_i's:

- When using a logistic model to estimate p_i, the predictor $\widehat{t}_y^{EBP(log)}$ showed negligible bias for all mechanisms except for Mechanism 4, where it showed a slight bias (5.3%) when $p_0 = 0.9$.
- When using the overall proportion of zero to estimate p_i, the predictor $\widehat{t}_y^{EBP(const)}$ was biased for every mechanisms except for Mechanism 3, as expected. Except for Mechanism 3, the bias increased as p_0 increased.
- When using a kernel density estimator, the predictor $\widehat{t}_y^{EBP(NP)}$ showed negligible bias for Mechanisms 3 and 4 for all values of p_0. For Mechanism 6, it is only biased when $p_0 = 0.9$. All the other mechanisms had a small bias (around 2% to 7%), especially for the larger values of p_0. The value of $H = 0.1$ leads to the least amount of bias.

As expected, the predictor $\widehat{t}_y^{EBP(log)}$ did great when the real model used to generate the probability p_i was a logistic model (Mechanisms 1 and 2) with a relative efficiency ranging from 25% to 96%. For Mechanism 1, the relative efficiency decreased as p_0 increased while for Mechanism 2, it is the inverse. The reason can be seen in Figure 4.9, where the blue dots represent the y-values for the non-sampled units, the black dots represent the values predicted by the EBP and the red line is the BLUP prediction. For Mechanism 1, when $p_0 = 0.1$, the zero-valued observations have only a small influence on the BLUP prediction since there are few of them and their x-values are small. Hence, the predictions from the BLUP and EBP are similar. However, when $p_0 = 0.9$, the zero-valued observations have a greater impact on the BLUP predictions and the predictions of EBP are much better. For Mechanism 2, the inverse happens and it can be seen in Figure 4.10.
Similar graphs can be found in the Appendix for all mechanisms.

For all other mechanims, the predictor $\widehat{t}_y^{EBP(log)}$ has a relative efficiency near 100% except for Mechanism 5 where some gain was obtained (RE of 30% to 99%), especially when the proportion of zero is high. The predictor $\widehat{t}_y^{EBP(const)}$ did better than the BLUP for Mechanism 3 and for low values of p_0 for Mechanism 6. In the other scenarios, it did poorly which was

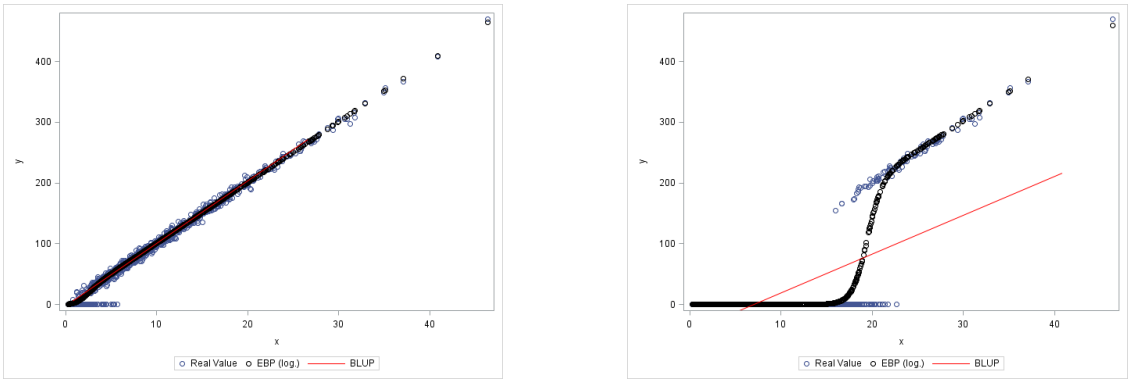


Fig. 4.9. Example of population with Mechanism 1 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

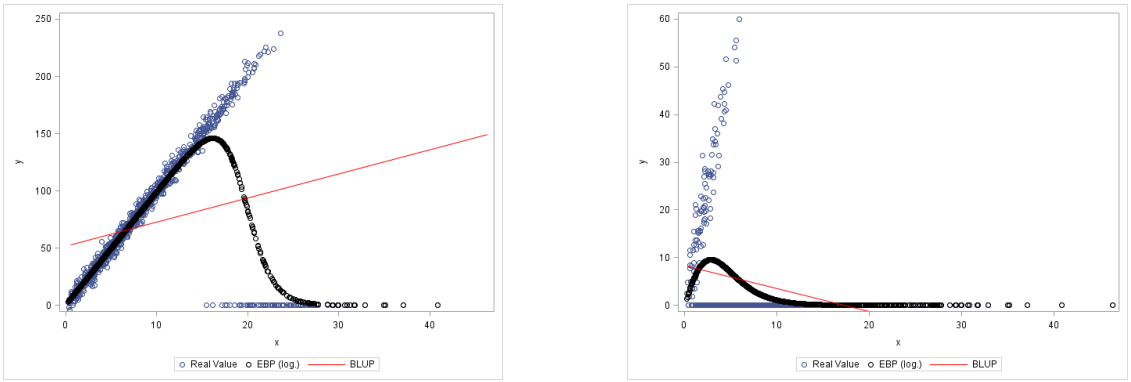


Fig. 4.10. Example of population with Mechanism 2 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

expected. The predictors $\widehat{t}_y^{EBP(NP)}$ and $\widehat{t}_y^{EBP(log)}$ had similar performances: they did well for high values of p_0 for Mechanism 1, low values of p_0 for Mechanism 2, and did better than the BLUP for Mechanism 5. For other mechanisms, it performed similarly to the BLUP (RE is close to 100%). In general, the value of H did not affect the relative efficiency.

4.3. Robust Prediction

We conducted a simulation study to assess the performance of the robust predictors presented in Chapter 3, in terms of bias and efficiency. Again, we repeated 1,000 iterations of the process described in Section 4.2.

We first generated an x-variable from a Gamma distribution with shape parameter equal to 2 and scale parameter equal to 5. Then, the survey variable y was generated from the conditional distribution

$$Y_i|x_i \sim \mathcal{D}(\mu_i,\nu_i), \tag{4.3.1}$$

where $\mu_i = \beta_0 + \beta_1 x_i$ and $\nu_i = \sigma^2 x_i$. We used four different distributions $\mathcal{D}$: normal, gamma, lognormal and Pareto. The parameters of each distribution were chosen such that μ_i and ν_i were the same for every distribution. We also generated a fifth population from a mixture of two normal distributions:

$$Y_i = \Delta_i \mathcal{N}(\mu_1,\sigma_1^2) + (1-\Delta_i)\mathcal{N}(\mu_2,\sigma_2^2), \tag{4.3.2}$$

where $P(\Delta_i = 1) = 0.95$ and μ_1, μ_2, σ_1 and σ_2 are the parameters of the two normal distributions (See the Appendix for the values of the parameters). Zero-valued observations were then generated according to Bernoulli trials with probability

$$p_i = \frac{1}{1+\exp\left(-10+0.15x_i\right)}.$$

This led to an overall proportion of zero-valued observations, $p_0 = 0.23$.

From each of the 1,000 populations, we selected a sample of size $n = 100$ according to simple random sampling without replacement. Examples of the five types of populations are shown in Figures 4.11-4.15.

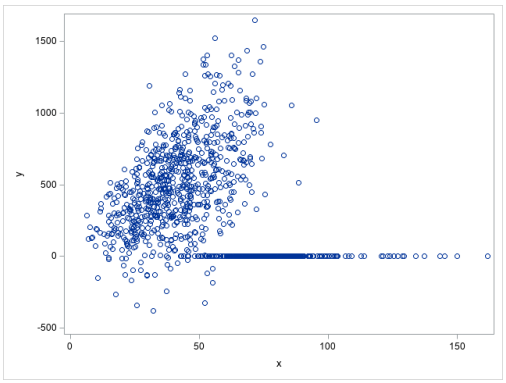


Fig. 4.11. Example of Population 1 with a normal distribution.

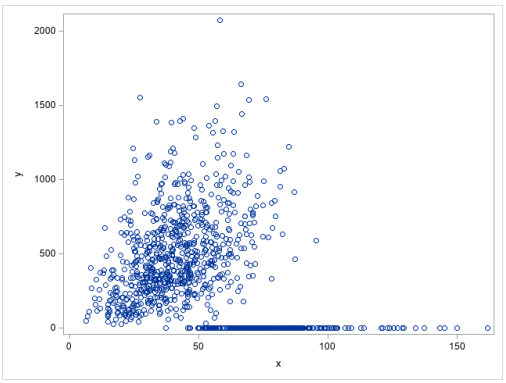


Fig. 4.12. Example of Population 2 with a gamma distribution.

In each sample, the following predictors were computed:

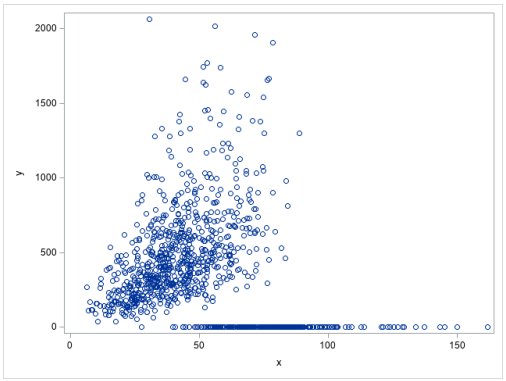


Fig. 4.13. Example of Population 3 with a lognormal distribution.

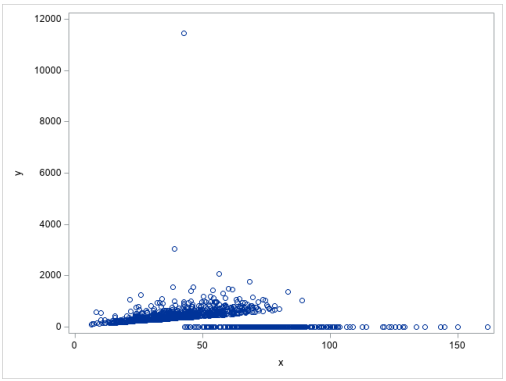


Fig. 4.14. Example of Population 4 with a Pareto distribution.

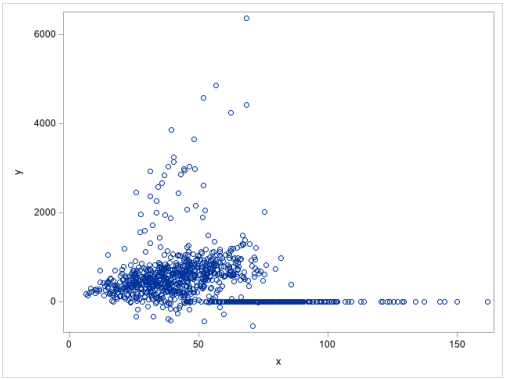


Fig. 4.15. Example of Population 5 with a mixture distribution.

(1) The BLUP:

$$\widehat{t}_y^{BLUP} = \sum_{i \in s} Y_i + \sum_{i \in U-s} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{WLS}.$$

(2) The EBP:

$$\widehat{t}_y^{EBP} = \sum_{i \in s_1} Y_i + \sum_{i \in U-s} \widehat{p}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{WLS1},$$

where p_i was estimated using a logistic model.

(3) The naive robust predictors BLUP and EBP:

$$\widehat{t}_y^{RBLUP}(k) = \sum_{i \in s} Y_i + \sum_{i \in U-s} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_R$$

and

$$\widehat{t}_y^{REBP}(k) = \sum_{i \in s_1} Y_i + \sum_{i \in U-s} \widehat{p}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{R1},$$

where $\widehat{\boldsymbol{\beta}}_R$ and $\widehat{\boldsymbol{\beta}}_{R1}$ were either a Huber M-estimator with $k = 0.1, 0.8, 1.345, 2$, a Bisquare M-estimator with $k = 3.5, 4.685, 6$, a LTS estimator or a MM-estimator.

(4) The predictor of Chambers (1986):

$$\widehat{t}_y^C(k,c) = \widehat{t}_y^{RBLUP}(k) + \sum_{i \in s}(w_i - 1)\widehat{\sigma}\psi_2\left(\frac{Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_R}{\widehat{\sigma}}; c\right),$$

where ψ_2 is the Huber function with $c = 2, 4, 6$ and 8 and $\widehat{\sigma}$ was estimated by the median absolute deviation. Note that $\widehat{t}_y^C(k,0) \equiv \widehat{t}_y^{RBLUP}(k)$.

(5) The predictor based on conditional bias for both BLUP and EBP:

$$\widehat{t}_y^{CB(BLUP)}(c_{opt}) = \widehat{t}_y^{BLUP} - \frac{1}{2}\left(\widehat{B}_{min}^{BLUP} + \widehat{B}_{max}^{BLUP}\right)$$
$$\widehat{t}_y^{CB(EBP)}(c_{opt}) = \widehat{t}_y^{EBP} - \frac{1}{2}\left(\widehat{B}_{min}^{EBP} + \widehat{B}_{max}^{EBP}\right),$$

where $\widehat{B}^{BLUP}$ was estimated by (2.2.6) with $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{WLS}$. The conditional bias in the predictor $\widehat{t}_y^{CB(EBP)}(c_{opt})$ was estimated by

$$\widehat{B}_i^{EBP} = -(w_i - 1)\widehat{p}_i \boldsymbol{x_i}^\top \widehat{\boldsymbol{\beta}}_{WLS1}.$$

We used the Monte Carlo relative bias (%) and relative efficiency (%) to compare the predictors to one another. Tables 4.3-4.7 show the results for all the predictors and populations.

The naive predictor $\widehat{t}_y^{RBLUP}$ was generally highly biased in all the scenarios. The bias increased as k decreased. It was also less efficient than the BLUP in all the scenarios. These results suggest that using a naive predictor in the context of survey sampling can lead to poor performances. The same was true for the naive EPB, $\widehat{t}_y^{REBP}$.

Turning to the predictor of Chambers (1986), the best performances were obtained for $c = 4$ and $c = 6$. This is consistent with what was suggested by Chambers. For Populations 1-3, the predictor of Chambers showed a value of efficiency close to 100. Some gains were observed for Population 4 and Population 5 with values of relative efficiency ranging from 77 to 80.

The predictor based on the conditional bias, $\widehat{t}_y^{CB(BLUP)}(c_{opt})$ was never less efficient than $\widehat{t}_y^{BLUP}$ but the gains were modest with values of relative efficiency ranging from 91

to 100. On the other hand, the predictor $\widehat{t}_y^{CB(EBP)}(c_{opt})$ leads to good gains in efficiency, with values of relative efficiency ranging from 68% to 92%. The greatest gains were observed for Population 4 (i.e., the Pareto distribution). It is worth pointing out that both $\widehat{t}_y^{CB(BLUP)}(c_{opt})$ and $\widehat{t}_y^{CB(EBP)}(c_{opt})$ exhibited small biases with an absolute relative bias less than 1.5%.

		BLUP					EBP
$\widehat{t}_y$		0.4 (100)					-0.1 (80)
$\widehat{t}_y^{CB}$		0.0 (100)					0.5 (80)
$\widehat{t}_y^{C}(k,c)$		Huber c					$c=0$
		0	2	4	6	8	
Huber k	0.1	-15.7 (528)	-3.9 (125)	0.2 (99)	0.4 (99)	0.4 (100)	0.1 (96)
	0.8	-11.2 (295)	-3.4 (116)	0.2 (99)	0.4 (100)	0.4 (100)	0.2 (83)
	1.345	-4.4 (137)	-2.7 (110)	0.3 (99)	0.4 (100)	0.4 (100)	0.2 (79)
	2	-1.0 (102)	-2.4 (107)	0.3 (99)	0.4 (100)	0.4 (100)	0.1 (79)
Bisquare k	3.5	-10.9 (320)	-3.4 (118)	0.2 (99)	0.4 (100)	0.4 (100)	0.3 (89)
	4.685	-5.2 (158)	-2.8 (111)	0.3 (100)	0.4 (100)	0.4 (100)	0.2 (81)
	6	-2.8 (116)	-2.6 (108)	-0.3 (100)	0.4 (100)	0.4 (100)	0.2 (79)
LTS		-25.9 (300)	-4.3 (193)	0.2 (105)	0.4 (101)	0.4 (100)	0.3 (90)
MM		-10.4 (271)	-3.3 (115)	0.2 (99)	0.4 (100)	0.4 (100)	0.2 (86)

Tab. 4.3. Results for Population 1 with $p_0 = 0.23$.

		BLUP					EBP
$\widehat{t}_y$		0.3 (100)					0.3 (80)
$\widehat{t}_y^{CB}$		-0.7 (99)					0.7 (80)
$\widehat{t}_y^{C}(k,c)$		Huber c 0	2	4	6	8	$c=0$
Huber k	0.1	-19.6 (698)	-7.0 (165)	-0.6 (98)	0.4 (99)	0.5 (100)	-8.7 (190)
	0.8	-15.0 (427)	-6.4 (151)	-0.5 (98)	0.4 (99)	0.5 (100)	-7.0 (143)
	1.345	-7.0 (173)	-5.5 (135)	-0.4 (98)	0.4 (99)	0.5 (100)	-4.7 (104)
	2	-2.2 (107)	-5.0 (128)	-0.3 (98)	0.4 (99)	0.5 (100)	-2.4 (84)
Bisquare k	3.5	-16.3 (526)	-6.6 (158)	-0.6 (98)	0.4 (99)	0.5 (100)	-9.2 (193)
	4.685	-8.8 (230)	-5.7 (141)	-0.4 (98)	0.4 (99)	0.5 (100)	-5.9 (125)
	6	-5.0 (141)	-5.2 (133)	-0.3 (98)	0.4 (99)	0.5 (100)	-3.8 (103)
LTS		-33.3 (1838)	-8.5 (225)	-0.9 (99)	0.3 (99)	0.5 (100)	-17.5 (500)
MM		-14.9 (427)	-6.4 (151)	-0.5 (98)	0.4 (99)	0.5 (100)	-8.2 (163)

Tab. 4.4. Results for Population 2 with $p_0 = 0.23$.

		BLUP					EBP
$\widehat{t}_y$		0.4					-0.1
		(100)					(80)
$\widehat{t}_y^{CB}$		-0.8					0.2
		(99)					(79)
$\widehat{t}_y^{C}(k,c)$		Huber					$c = 0$
		c					
		0	2	4	6	8	
Huber k	0.1	-19.2	-8.2	-1.6	-0.1	0.2	-10.6
		(605)	(170)	(96)	(96)	(99)	(241)
	0.8	-15.4	-7.7	-1.5	-0.1	0.2	-8.9
		(402)	(159)	(95)	(96)	(99)	(184)
	1.345	-7.9	-6.8	-1.2	0.0	0.2	-6.2
		(179)	(144)	(95)	(97)	(99)	(127)
	2	-3.1	-6.3	-1.1	0.0	0.3	-3.5
		(110)	(134)	(95)	(97)	(99)	(93)
Bisquare k	3.5	-17.2	-8.0	-1.5	-0.1	0.2	-11.7
		(509)	(167)	(95)	(97)	(98)	(270)
	4.685	-10.2	-7.1	-1.3	-0.1	0.2	-8.1
		(245)	(150)	(96)	(97)	(98)	(172)
	6	-6.2	-6.6	-1.2	0.0	0.3	-5.7
		(153)	(141)	(95)	(97)	(98)	(123)
LTS		-32.6	-9.7	-1.9	-0.2	-0.2	-19.0
		(1618)	(230)	(98)	(96)	(98)	(598)
MM		-15.9	-7.8	-1.5	-0.1	0.2	-10.2
		(423)	(161)	(95)	(96)	(98)	(218)

Tab. 4.5. Results for Population 3 with $p_0 = 0.23$.

		BLUP					EBP
$\widehat{t}_y$		0.4					0.0
		(100)					(77)
$\widehat{t}_y^{CB}$		-0.8					-0.1
		(91)					(68)
$\widehat{t}_y^{C}(k,c)$		Huber c					$c=0$
		0	2	4	6	8	
Huber k	0.1	-9.7	-6.1	-3.1	-1.5	-0.7	-13.4
		(351)	(185)	(85)	(79)	(81)	(305)
	0.8	-10.2	-7.7	-3.2	-1.5	-0.7	-11.6
		(261)	(169)	(83)	(78)	(81)	(242)
	1.345	-7.2	-7.7	-3.2	-1.4	-0.7	-8.6
		(141)	(146)	(80)	(77)	(81)	(158)
	2	-4.1	-7.2	-3.1	-1.4	-0.7	-6.0
		(90)	(133)	(80)	(77)	(81)	(107)
Bisquare k	3.5	-12.1	-8.3	-3.3	-1.6	-0.8	-14.6
		(317)	(177)	(86)	(78)	(81)	(357)
	4.685	-8.9	-8.0	-3.2	-1.5	-0.7	-11.8
		(183)	(152)	(82)	(77)	(80)	(251)
	6	-6.7	-7.6	-3.2	-1.5	-0.7	-9.5
		(127)	(142)	(80)	(77)	(80)	(184)
LTS		1.9	4.1	1.2	-0.7	-0.6	-16.9
		(1363)	(448)	(136)	(84)	(83)	(453)
MM		-10.2	-7.6	-3.0	-1.5	-0.7	-12.4
		(302)	(182)	(88)	(78)	(80)	(207)

Tab. 4.6. Results for Population 4 with $p_0 = 0.23$.

		BLUP					EBP
$\widehat{t}_y$		0.6 (100)					0.6 (92)
$\widehat{t}_y^{CB}$		-1.5 (93)					-1.0 (87)
$\widehat{t}_y^{C}(k,c)$		Huber c					$c=0$
		0	2	4	6	8	
Huber k	0.1	-25.3 (540)	-13.6 (182)	-7.0 (92)	-3.8 (80)	-1.8 (83)	-12.4 (146)
	0.8	-20.8 (368)	-13.0 (171)	-6.8 (91)	-3.8 (80)	-1.7 (83)	-12.0 (138)
	1.345	-14.2 (196)	-12.3 (158)	-6.6 (90)	-3.6 (80)	-1.6 (83)	-11.4 (128)
	2	-9.7 (121)	-11.8 (150)	-6.5 (89)	-3.5 (80)	-1.5 (83)	-10.3 (114)
Bisquare k	3.5	-23.7 (471)	-13.4 (178)	-6.9 (92)	-3.8 (80)	-1.8 (83)	-14.7 (185)
	4.685	-18.4 (296)	-12.8 (166)	-6.8 (90)	-3.7 (80)	-1.7 (83)	-14.4 (176)
	6	-15.4 (221)	-12.4 (160)	-6.7 (90)	-3.7 (80)	-1.7 (83)	-13.8 (166)
LTS		-35.7 (1073)	-14.6 (214)	-7.3 (95)	-4.1 (80)	-1.9 (82)	-15.0 (217)
MM		-22.2 (404)	-13.2 (173)	-6.9 (91)	-3.8 (80)	-1.7 (83)	-14.6 (182)

Tab. 4.7. Results for Population 5 with $p_0 = 0.23$.

Conclusion

In this thesis, we examined the use of design-based estimators and model-based predictors of a population total t_y when the population is prone to a large number of zero-valued observations. We also developed robust predictors based on the concept of conditional bias when influential observations are present in the sample.

We began by introducing usual estimators and predictors and examined their properties when used for populations with a large number of zero-valued observations. For the design-based approach, the results of a simulation study presented in Chapter 4 suggested that the variance of the customary estimators increases as the proportion of zeroes p_0 increases. For the ratio and GREG estimators, their variance not only depends on the proportion p_0 but also on ϕ_0, the fraction of the total t_x corresponding to the zero-valued observations. This suggests that the distribution of the zero/nonzero status has an impact on the variance of these estimators. In the context of the model-based approach, we studied the BLUP and the EBP based on a mixture model. We saw that, when the assumption made on the model about the probability of zero/nonzero status, p_i, is correct, the EBP leads to a gain in efficiency compared to the BLUP. However, when the model is misspecified, the resulting predictor may be highly biased.

In the second simulation, we compared the robust predictors in the presence of influential observations. The results suggested that using naive predictors may lead to poor performances. On the other hand, the performance of the predictor of Chambers (1986) depends on the choice of the tuning constant c. The values $c = 4$ or $c = 6$ turned out to be the best values, in general, which is consistent with what was advocated by Chambers (1986). The robust version of the EBP based on the conditional bias did well in terms of

efficiency and was never less efficient than the BLUP. Also, the predictor $\widehat{t}_y^{EBP}$ in the presence of influential units still does better than the robust BLUP $\widehat{t}_y^{CB(BLUP)}$ but we note that when predicting the EBP in the simulation study, we used the correctly specified model of p_i which is, most of the time, unknown. The result from the first simulation showed that a misspecification of p_i can have a great effect on the performance. Hence, when unsure about the distribution, it might be better to use the predictor based on the conditional bias for BLUP $\widehat{t}_y^{CB(BLUP)}$ as it will treat the zero-valued observations as outliers and still give better results than the simple BLUP, without any assumption made on p_i.

Bibliography

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

[2] Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1066–1069.

[3] Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14, 153–158.

[4] Huber, P.J. (1981). *Robust Statistics.* John Wiley & Sons, Inc., New York.

[5] Kalberg, F. (2000). Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16, 229–241.

[6] Liu, Y., Batcher, M., and Scheuren, F. (2005). Efficient Sampling Design in Audit Data. *Journal of Data Science*, 3, 213–222.

[7] Lohr, S. L. (2009). *Sampling: Design and Analysis.* Duxbury Press, Pacific Grove, CA.

[8] 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.

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

[10] Yohai, V.J. (1987). High breakdown point and high efficiency robust estimates for regression, *The Annals of Statistics*, 15, 642–656.

Appendix A

A.1. Proof of conditional bias estimator

i) When $\tilde{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}_{LS}$:

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

ii) When $\tilde{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}_{LS}^{(-i)}$:

First,

$$\begin{aligned}
(y_i-\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{LS}^{(-i)})(1-h_{ii}) &= y_i-y_ih_{ii}-\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{LS}^{(-i)}+\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{LS}^{(-i)}h_{ii} \\
&= y_i-y_ih_{ii}-\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{LS}^{(-i)}\sum_{\substack{j\in s\\ j\neq i}}\boldsymbol{x}_j\boldsymbol{x}_j^\top\left(\sum_{j\in s}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)^{-1} \\
&= y_i-y_i\boldsymbol{x}_j\boldsymbol{x}_j^\top\left(\sum_{j\in s}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)^{-1} \\
&\quad -\boldsymbol{x}_i^\top\sum_{\substack{j\in s\\ j\neq i}}\boldsymbol{x}_jY_j\left(\sum_{\substack{j\in s\\ j\neq i}}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)^{-1}\sum_{\substack{j\in s\\ j\neq i}}\boldsymbol{x}_j\boldsymbol{x}_j^\top\left(\sum_{j\in s}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)^{-1} \\
&= y_i-\boldsymbol{x}_i^\top y_i\boldsymbol{x}_i\left(\sum_{j\in s}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)^{-1}-\boldsymbol{x}_i^\top\sum_{\substack{j\in s\\ j\neq i}}\boldsymbol{x}_jY_j\left(\sum_{j\in s}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)^{-1}
\end{aligned}$$

$$= y_i - \boldsymbol{x}_i^\top \sum_{j\in s} \boldsymbol{x}_j Y_j \left(\sum_{j\in s} \boldsymbol{x}_j \boldsymbol{x}_j^\top\right)^{-1}$$
$$= y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{LS}.$$

Hence,

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

Then,

$$\begin{aligned}
\mathbb{E}_m\left(\widehat{B}_i^{BLUP} | s, Y_i = y_i\right) &= \mathbb{E}_m\left\{(w_i - 1)\frac{y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{LS}}{1-h_{ii}} | s, Y_i = y_i\right\} \\
&= \frac{w_i - 1}{1-h_{ii}}\left(y_i - \mathbb{E}_m\left\{\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{LS} | s, Y_i = y_i\right\}\right) \\
&= \frac{w_i - 1}{1-h_{ii}}\left(y_i - \mathbb{E}\left\{\boldsymbol{x}_i^\top \sum_{j\in s} \boldsymbol{x}_j Y_j \left(\sum_{j\in s} \boldsymbol{x}_j \boldsymbol{x}_j^\top\right)^{-1} | s, Y_i = y_i\right\}\right) \\
&= \frac{w_i - 1}{1-h_{ii}}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{x}_i y_i \left(\sum_{j\in s} \boldsymbol{x}_j \boldsymbol{x}_j^\top\right)^{-1}\right. \\
&\quad \left. - \mathbb{E}_m\left\{\boldsymbol{x}_i^\top \sum_{\substack{j\in s \\ j\neq i}} \boldsymbol{x}_j Y_j \left(\sum_{j\in s} \boldsymbol{x}_j \boldsymbol{x}_j^\top\right)^{-1} | s, Y_i = y_i\right\}\right) \\
&= \frac{w_i - 1}{1-h_{ii}}\left(y_i - h_{ii} y_i - \boldsymbol{x}_i^\top \sum_{\substack{j\in s \\ j\neq i}} \boldsymbol{x}_j \boldsymbol{x}_j^\top \boldsymbol{\beta} \left(\sum_{j\in s} \boldsymbol{x}_j \boldsymbol{x}_j^\top\right)^{-1}\right) \\
&= \frac{w_i - 1}{1-h_{ii}}\left(y_i - h_{ii} y_i - \boldsymbol{x}_i^\top \sum_{j\in s} \boldsymbol{x}_j \boldsymbol{x}_j^\top \boldsymbol{\beta} \left(\sum_{j\in s} \boldsymbol{x}_j \boldsymbol{x}_j^\top\right)^{-1}\right. \\
&\quad \left. + \boldsymbol{x}_i^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\beta} \left(\sum_{j\in s} \boldsymbol{x}_j \boldsymbol{x}_j^\top\right)^{-1}\right) \\
&= \frac{w_i - 1}{1-h_{ii}}\left(y_i - h_{ii} y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} + h_{ii} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \\
&= \frac{w_i - 1}{1-h_{ii}}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)(1 - h_{ii})
\end{aligned}$$

$$= (w_i - 1)\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)$$
$$= B_i^{BLUP}.$$

A.2. Graphs

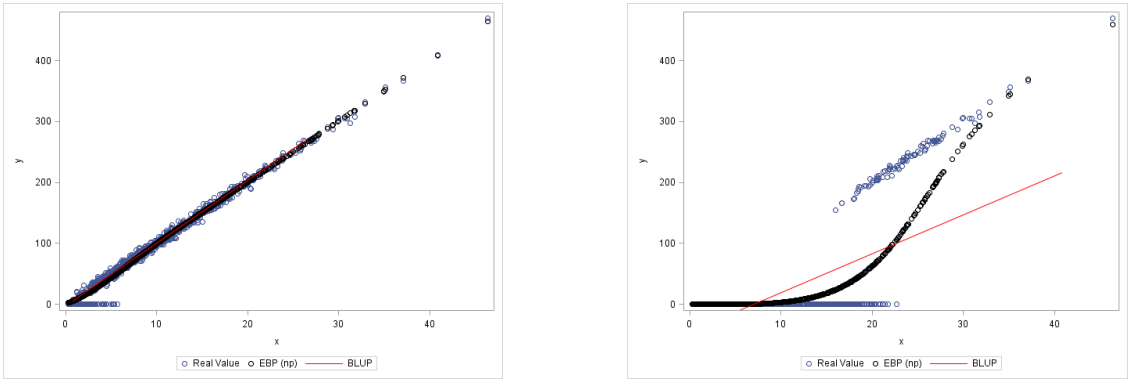


Fig. A.1. Example of population with Mechanism 1 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

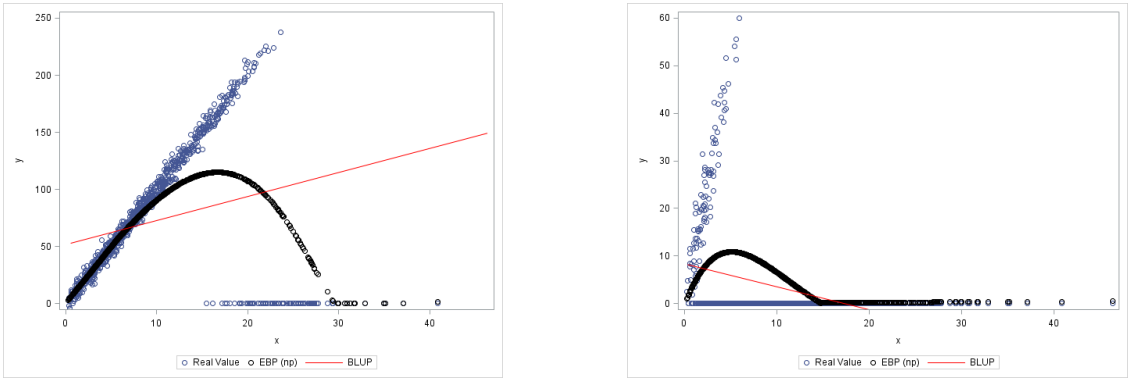


Fig. A.2. Example of population with Mechanism 2 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

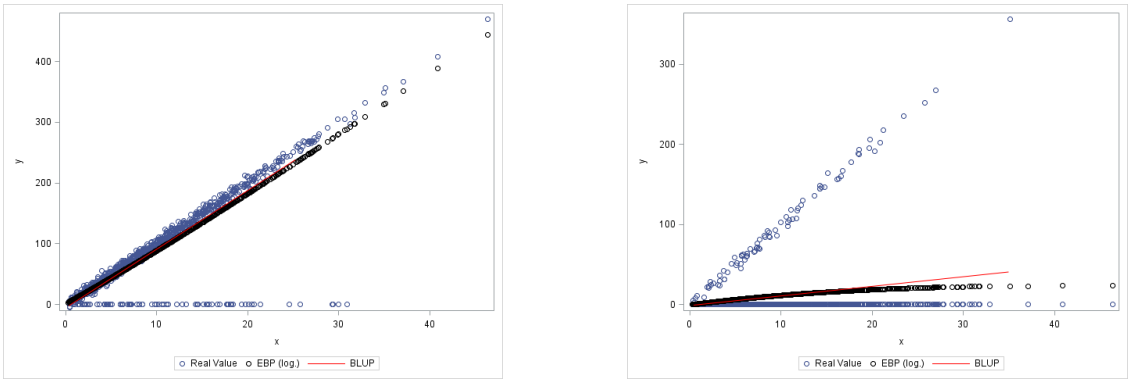


Fig. A.3. Example of population with Mechanism 3 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

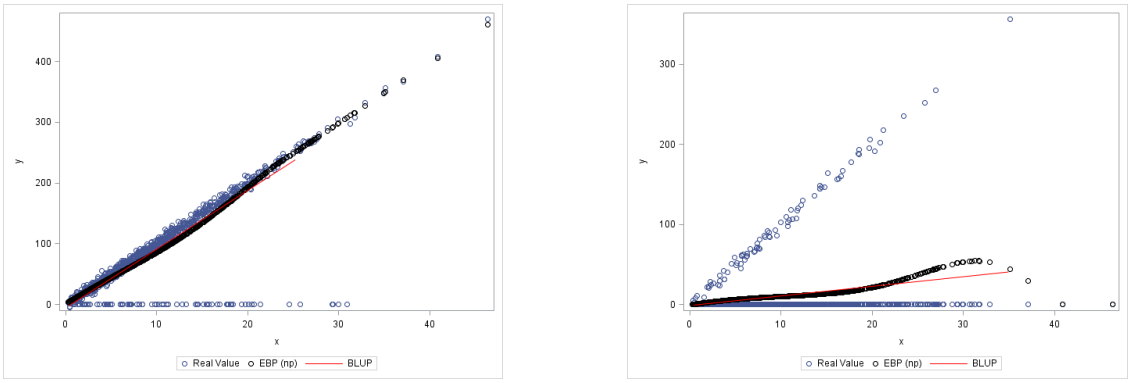


Fig. A.4. Example of population with Mechanism 3 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

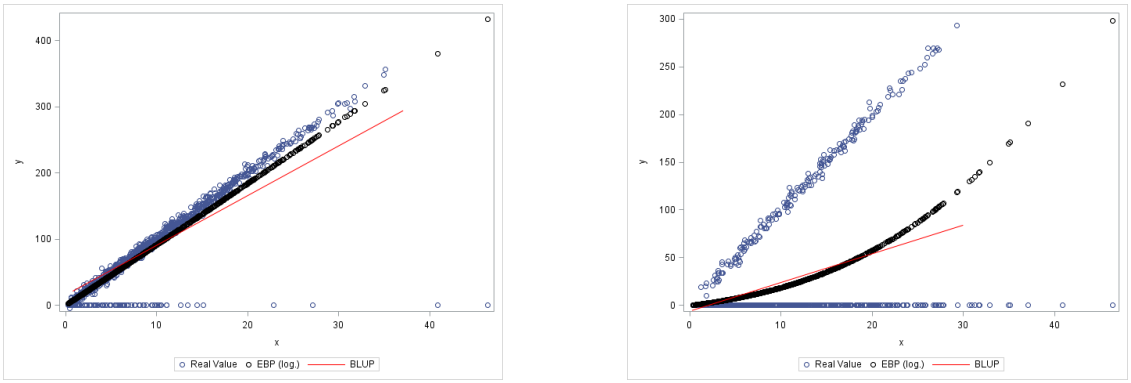


Fig. A.5. Example of population with Mechanism 4 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

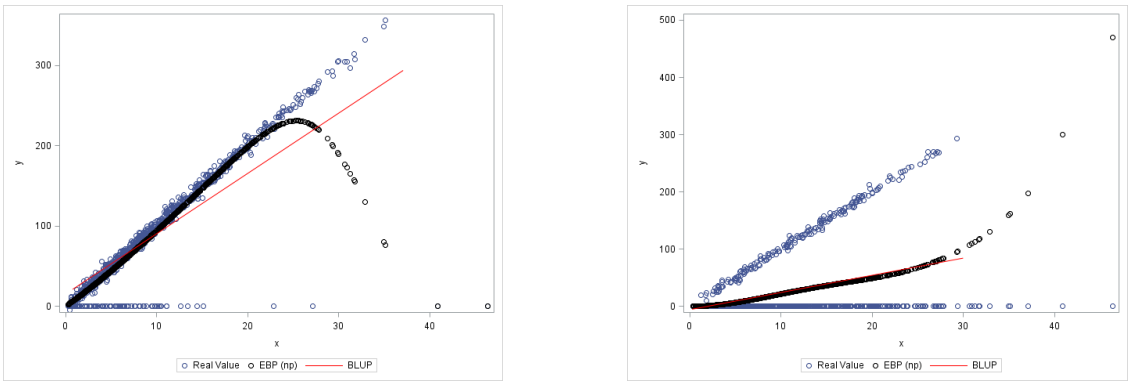


Fig. A.6. Example of population with Mechanism 4 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

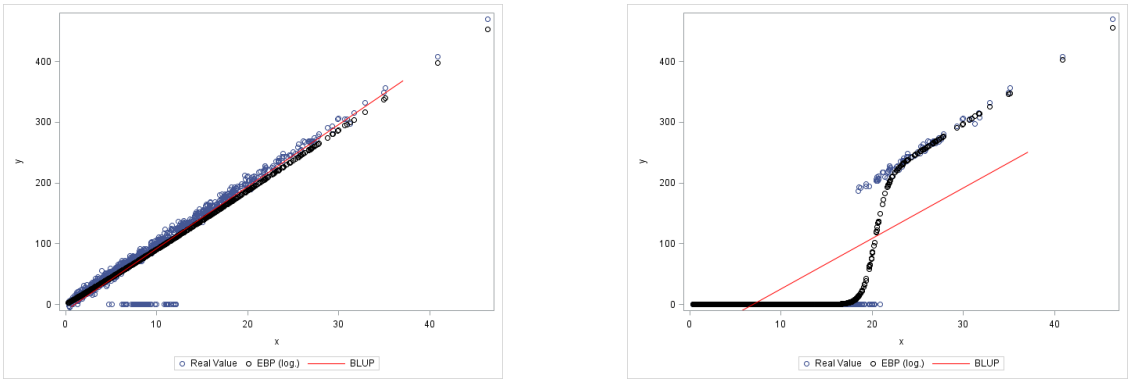


Fig. A.7. Example of population with Mechanism 5 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

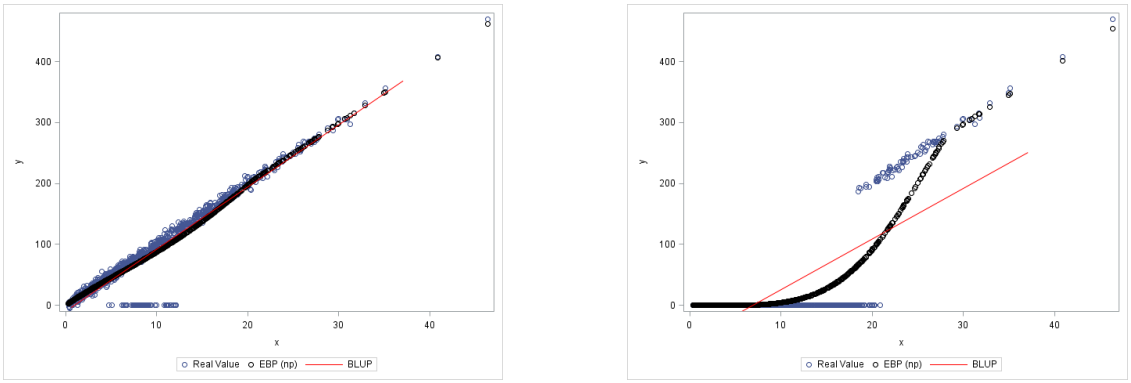


Fig. A.8. Example of population with Mechanism 5 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

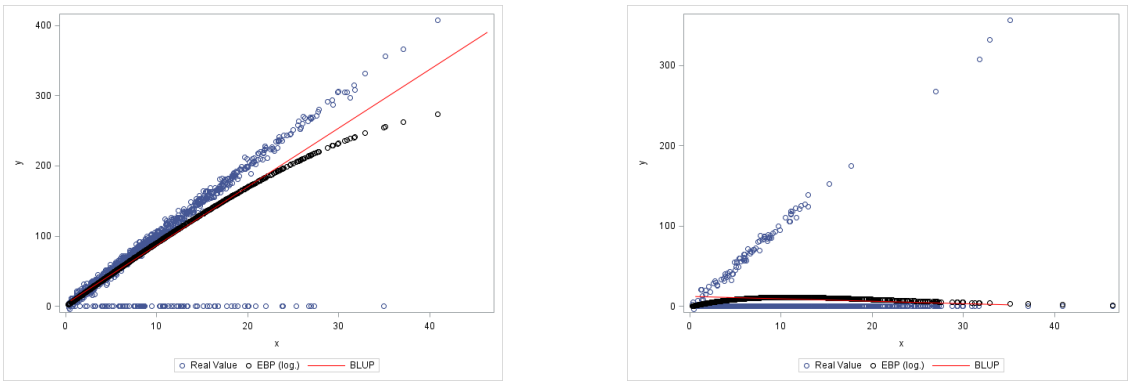


Fig. A.9. Example of population with Mechanism 6 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

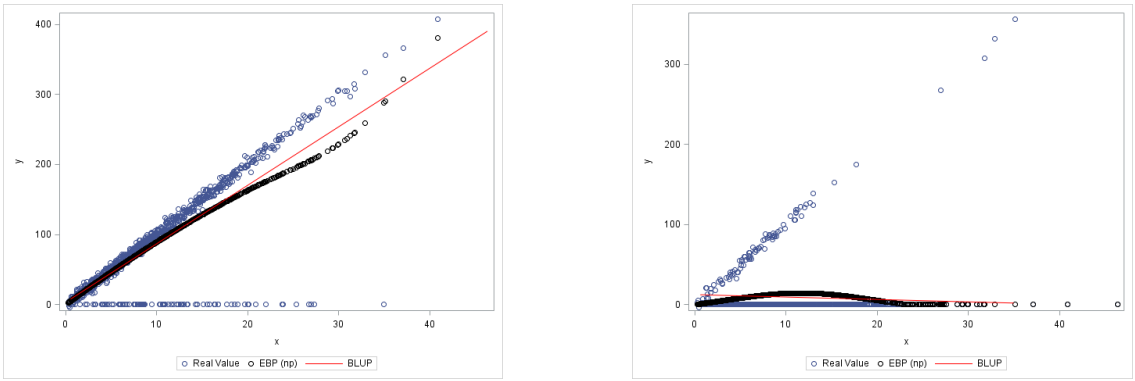


Fig. A.10. Example of population with Mechanism 6 and its predictions for BLUP and EBP with $p_0 = 0.1$ on the left and $p_0 = 0.9$ on the right.

A.3. Parameters used in simulations

	Design-based	Model-based						
		p_0	γ_0	γ_1	γ_2	γ_3	γ_4	γ_5
Mechanism 1	All combinations of of $\gamma_0 = 1$ to 9 and $\gamma_1 = 0.1$ to 0.8	0.1	-2.3	1				
		0.3	-5.5	1				
		0.5	-8.6	1				
		0.7	-12.3	1				
		0.9	-18.1	0.9				
Mechanism 2	All combinations of of $\gamma_0 = 1$ to 9 and $\gamma_1 = 0.1$ to 0.8	0.1	-18.1	0.9				
		0.3	-12.3	1				
		0.5	-8.6	1				
		0.7	-5.5	1				
		0.9	-2.3	1				
Mechanism 3		0.1	0.9					
		0.3	0.7					
		0.5	0.5					
		0.7	0.3					
		0.9	0.1					

Tab. A.1. Parameters of p_i for the first part of simulation for both design-based and model-based (Mech. 1 to 3).

	Design-based	Model-based						
		p_0	γ_0	γ_1	γ_2	γ_3	γ_4	γ_5
Mechanism 4		0.1	250	11	-0.3			
		0.3	180	11	-0.3			
		0.5	110	11	-0.3			
		0.7	40	11	-0.3			
		0.9	2	11	-0.3			
Mechanism 5		0.1	8	-1.6	0.1	4.1	0.2	-1.9
		0.3	5.3	-1.6	0.1	4.1	0.2	-1.9
		0.5	3.5	-1.6	0.1	4.1	0.2	-1.9
		0.7	1.4	-1.6	0.1	4.1	0.2	-1.9
		0.9	-8.8	-1.6	0.1	4.1	0.2	-1.9
Mechanism 6		0.1	0.36	0.01				
		0.3	0.7	0.01				
		0.5	0.95	0.01				
		0.7	1.16	0.01				
		0.9	1.38	0.01				

Tab. A.2. Parameters of p_i for the first part of simulation for both design-based and model-based (Mech. 4 to 6). Note that for Mechanism 4, the final p_i is found by dividing the resulting p_i the maximum value of the p_i.

For the simulation for robust predictions:

- Population 1 to 4: $\mu_i = 100 + 10x_i$ and $\nu_i = 25x_i$.
- Population 5: $\mu_1 = 100 + 10x_i$, $\sigma_1^2 = 2000x_i$, $\mu_2 = 5\mu_1$ and $\sigma_2^2 = 10\sigma_1^2$.