Penalization and data reduction of auxiliary variables in survey sampling

Muhammad Ahmed Shehzad

UNIVERSITÉ DE BOURGOGNE

U.F.R. Sciences et Techniques

Institut de Mathématiques de Bourgogne

UMR 5584 du CNRS

THÈSE

pour l'obtention du grade de

DOCTEUR de l'Université de Bourgogne

Discipline : Mathématiques

par

Muhammad Ahmed Shehzad

Pénalisation et réduction de la dimension des variables auxiliaires en théorie de sondages

Directeur de thèse: **Hervé Cardot**

Co-directrice de thèse: **Camelia Goga**

Thèse soutenue le **12 Octobre 2012** devant le jury composé de:

Muhammad Hanif	NCBA and E Lahore	Rapporteur
Anne Ruiz-Gazen	Université Toulouse 1 Capitole	Rapportrice
Jean-Claude Deville	CREST-ENSAI	Examinateur
Guillaume Chauvet	CREST-ENSAI	Examinateur
Célestin Kokonendji	Université de Franche-Comté	Examinateur
Aurélie Vanheuverzwyn	Médiamétrie	Invitée

Penalization and data reduction of auxiliary variables in survey sampling

Muhammad Ahmed Shehzad
Institute of Mathematics, University of Burgundy

October 15, 2012

Acknowledgments

I would like to thank to all who accompanied me and contributed in my work somehow or other, knowingly or unknowingly, close or far from me during my all efforts of life.

My heart is filled with utmost respect and deep gratitude for Dr. Camelia Goga and Dr. Hervé Cardot for their inspiring, tireless and dynamic guidance which made me able to complete my thesis work. My special thanks are for Dr. Camelia Goga who showed me way out in difficult times by her useful suggestions and noble ideas.

I am also really grateful to Dr. Muhammad Hanif Mian and Dr. Anne Ruiz-Gazen for accepting the evaluation of my thesis.

My inspirations has great part by Dr. Jean-Claude Deville whose ideas and work always remained helpful during my PhD thesis. I also thank him and Dr. Guillaume Chauvet, Dr. Célestin Kokonendji and Aurélie Vanheuverzwyn for providing the time to become jury members of my thesis.

I would also thank Dr. Muhammad Aslam whose appreciation motivated me to achieve my goals.

I thanks all of the members of IMB, specially to the Statistics and Probability team for being so nice with me. Also all the staff members including Caroline, Anissa, Francis and Pierre for giving a friendly working environment.

I pay my special thanks the Higher Education Commission (HEC) of Pakistan which provided me the key support in terms of granting me the PhD scholarship.

I would always remember my Pakistani fellows Sabir Hussain, Hamid Manzoor, Farrukh Azeem, Ahsan Mehmood, Amjad Ali, Ashfaq Ahmed Shah, Sajjad Haider, Muhammad Abid, Shamsheer, Farhan Hafeez, Muhammad Atif and Farasat for providing me the home-like environment in Dijon. Also, thanks to my friend Awais Rasheed who always supported me to boost my courage.

I am unable to describe the continuous support, prayers and guidance of my beloved parents who always ensured their helping hand in all weathers of life. Let

me also thank to my parents-like in laws for their love and affection. My thanks are also for my sincere brothers and sisters for their selfless cooperation. Finally, I dedicate my work to a special person in my life.

Résumé

Les enquêtes par sondage sont utiles pour estimer des caractéristiques d'une population telles que le total ou la moyenne. Cette thèse s'intéresse à l'étude de techniques permettant de prendre en compte un grand nombre de variables auxiliaires pour l'estimation d'un total.

Le premier chapitre rappelle quelques définitions et propriétés utiles pour la suite du manuscrit : l'estimateur de Horvitz-Thompson, qui est présenté comme un estimateur n'utilisant pas l'information auxiliaire ainsi que les techniques de calage qui permettent de modifier les poids de sondage de facon à prendre en compte l'information auxiliaire en restituant exactement dans l'échantillon leurs totaux sur la population.

Le deuxième chapitre, qui est une partie d'un article de synthèse accepté pour publication, présente les méthodes de régression ridge comme un remède possible au problème de colinéarité des variables auxiliaires, et donc de mauvais conditionnement. Nous étudions les points de vue "model-based" et "model-assisted" de la ridge régression. Cette technique qui fournit de meilleurs résultats en terme d'erreur quadratique en comparaison avec les moindres carrés ordinaires peut également s'interpréter comme un calage pénalisé. Des simulations permettent d'illustrer l'intérêt de cette technique par comparison avec l'estimateur de Horvitz-Thompson.

Le chapitre trois présente une autre manière de traiter les problèmes de colinéarité via une réduction de la dimension basée sur les composantes principales. Nous étudions la régression sur composantes principales dans le contexte des sondages. Nous explorons également le calage sur les moments d'ordre deux des composantes principales ainsi que le calage partiel et le calage sur les composantes principales estimées. Une illustration sur des données de l'entreprise Médiamétrie permet de confirmer l'intérêt des ces techniques basées sur la réduction de la dimension pour l'estimation d'un total en présence d'un grand nombre de variables auxiliaires.

Mots clés : sondage, colinéarité, régression ridge, calage pénalisé, estimateur

assisté par un modèle, estimateur basé sur un modèle, estimateur de Horvitz-Thompson, calage sur composantes principales.

Abstract

Survey sampling techniques are quite useful in a way to estimate population parameters such as the population total when the large dimensional auxiliary data set is available. This thesis deals with the estimation of population total in presence of ill-conditioned large data set.

In the first chapter, we give some basic definitions that will be used in the later chapters. The Horvitz-Thompson estimator is defined as an estimator which does not use auxiliary variables. Along with, calibration technique is defined to incorporate the auxiliary variables for sake of improvement in the estimation of population totals for a fixed sample size.

The second chapter is a part of a review article about ridge regression estimation as a remedy for the multicollinearity. We give a detailed review of the model-based, design-based and model-assisted scenarios for ridge estimation. These estimates give improved results in terms of MSE compared to the least squared estimates. Penalized calibration is also defined under survey sampling as an equivalent estimation technique to the ridge regression in the classical statistics case. Simulation results confirm the improved estimation compared to the Horvitz-Thompson estimator.

Another solution to the ill-conditioned large auxiliary data is given in terms of principal components analysis in chapter three. Principal component regression is defined and its use in survey sampling is explored. Some new types of principal component calibration techniques are proposed such as calibration on the second moment of principal component variables, partial principal component calibration and estimated principal component calibration to estimate a population total. Application of these techniques on real data advocates the use of these data reduction techniques for the improved estimation of population totals.

Keywords: Survey sampling, Multicollinearity, Ridge regression, Penalized calibration, Model-based estimator, Model-assisted estimator, Horvitz-Thompson estimator, Principal component calibration.

Contents

Introduction

Estimation of the statistical parameters such as population mean or population total is generally supposed to be made efficient by employing survey sampling techniques that are using extensively large auxiliary variables. However, the large data sets import some data problems and hence make the estimation rather faultier. In this thesis, the problems inherent in the dimension of data in the structure of data are solved by two different ways namely ridge regression and principal component regression.

The first chapter includes some basic definitions and the introduction to the regression and calibration estimators which serve equally in the estimation procedure.

The data problems such as multicollinearity and ill-conditioning in large data sets cause singular regression coefficient and potentially result in inefficient estimators and in calibration technique, resulting inappropriate weights, may be the worst circumstances faced by a survey statistician.

This thesis is an effort to establish those methods which can negotiate the above mentioned problems in the best possible way using large dimensional auxiliary variables. The ultimate goal of the whole exercise is to get improved estimators of the population total.

Among the reasons for the data problems, may be the non-response or recording errors which can be minimized but not totally eliminated. So, if the utilization of large amount of auxiliary information is quite attractive due to the improved estimators, the problems related to these extensive amount of data are also over-

whelming.

The objective is to achieve a compromise between the cost paid through these irresistible data problems and the gain attained via the use of the auxiliary variables.

The second chapter contains an article (Goga and Shehzad, 2011 (under review)), which in fact is a detailed overview of the ridge type of estimation both in model-based and model-assisted cases as a solution to the ill-conditioned data.

Although, the selection of a unique value for the ridge parameter remains an open problem, very popular and easy to calculate method is ridge trace (Hoerl and Kennard, 1970). Theobald (1974) gave a condition on the choice of ridge parameter for having means squared error of ridge estimator less than that of the least squared estimator. Several mathematical situations where ridge regression estimator can serve as a solution to the ill-conditioned data, are considered in chapter 2.

We discussed the ridge regression estimator as an estimator producing the smallest residual sum of squares compared to the ordinary least squares estimator. so, using the work of Hoerl and Kennard (1970), Marquardt and Snee (1975) and Izenman (2008), we can see the ridge estimator as a solution of a constrained minimization problem. Also, ridge regression coefficient estimator can be taken as the posterior mean of the unknown regression coefficient and search for its prior distribution. For a suitable prior and finding a posterior distribution of regression coefficient is searched with th mean as the ridge regression estimator.

The case of different error variances is is also discussed namely the problem of heteroscedasticity in case of ridge estimators.

Certain cases of ridge regression estimator in classical regression are gathered and results showing improved mean squared error of ridge estimator compared to the ordinary least square estimator is established.

Model-based and model-assisted estimators for the regression coefficient in ridge case presented in terms of the optimization problems and the relevant model-

based and model-assisted (GREG) estimators for the population total are calculated. Certain conditions on the ridge parameter are explored. The case of penalized calibration (or ridge calibration) is presented and the ridge calibration weights are calculated and a GREG-type estimator is obtained.

Another case of partial penalized calibration is stated and equivalence of two-type of partially penalized calibration estimators is shown. Deville (1999) however gave an estimator without need of any penalty but some sort of external source of information is pre-requisite for this method to hold. His method gets inspiration from the Bayesian estimator to estimated the regression coefficient. Statistical properties such as bias, variance, asymptotic variance and mean squared error under model-based and model-assisted cases are given. A small simulation study is done and superiority of ridge estimator over Horvitz-Thompson estimator is shown.

Chapter 3 comprises of the description of principal component analysis and its use in regression analysis namely principal component regression (PCR). The choice of number of principal components to be included in the estimation procedure depends upon the statisticians. However, Jolliffe (2002) gave a detailed overview of possible methods for the choice of principal components. Section 3.1.2 mentions the theorem by Gunst and Mason (1977) showing the mean squared error of principal component estimator inferior to that of the least squared estimator. In Section 3.2, we study the principal component regression in survey sampling and we formulate model-based and model-assisted properties of the principal component estimator. The convergence of Horvitz-Thompson type expression for asymptotic variance is developed and its estimator is also given. In section 3.2.4, we propose some new calibration techniques using principal components such as calibration on the second moment of the principal component variables, calibration using estimated principal components and partial principal component calibration for the estimation of population totals. These biased methods serve as an alternative to the ridge calibration for tackling the ill-conditioning present in the data.

Compared to the ridge estimators which are penalizing methods, the estimators based on principal components are the dimension reduction methods.

The principal component estimators have a certain advantage in the fact that each principal component is a linear combination of all original variables, so maximum information is in-hand while the reduction in dimension is also achieved. This however may not be possible to compute when the original variables are not known for whole population. We estimated principal components and used them in place of population principal components and simulation on the Mediametrie showed that both methods have similar performances. Finally, The comparison between other proposed methods is also made by using figures and tables. Finally, chapter 3 comprise of discussion about the results attained in this Phd work and some future perspectives are noted.

Chapter 1

Total estimation techniques in survey sampling

1.1 Introduction

Large dimensional data sets in survey sampling are often encountered in the estimation procedures. Several techniques have been found in literature both in model-based and design-based environments to deal with the complications such as multicollinearity and ill-conditioning related to the large dimensional data in survey sampling. The total estimation in survey sampling in different scenarios are discussed and several methods are designed both to cope with the above mentioned data problems and the largeness of the data dimensions. The proposed techniques are illustrated with some real data application.

After giving some basic notations in section 1, we present the unbiased estimator of population total Horvitz-Thompson (1952) estimator in section 2. General type of variance and estimator of variance expressions of the Horvitz-Thompson estimator are provided which will be used throughout this thesis work. We present the simple random sampling without replacement which is the sampling design used for the sample selection in chapter 2 and chapter 3 in the practical application of the proposed methods. Later, in the same section, we define the auxiliary

data sets to be used at the design and estimation stages. Section 3 contains the definition of the generalized difference and generalized regression (GREG) estimators (Cassel *et al*, 1976) which incorporate the auxiliary information for the improvement of the estimation procedure in terms of the smaller errors than Horvitz-Thompson estimator (Särndal *et al*, 1992). The expressions for the variance and its estimator are given for GREG estimator of the population total which will be used for the variance construction in chapter 2 and chapter 3 for ridge and PC type estimators. The section 4 describes the calibration estimation technique (Deville and Särndal ,1992) which does not depend on the superpopulation model and generates weights which are ultimately used for the estimation of population total t_y. This method serves as a back-up for the estimation procedure in case of model failure.

1.1.1 Notations

We consider a finite population U containing N elements such that

$$\mathcal{U} = \{a_1, \ldots, a_k, \ldots, a_N\} = \{1, \ldots, k, \ldots, N\}$$

with the supposition that the population units are identifiable uniquely by their label k (Cassel *et al*, 1977). Let $\mathbf{Y}$ be a variable of interest and y_k denotes the value of $\mathbf{Y}$ for the kth individual. The finite population parameter of the unknown variable of interest may be denoted as a vector, $\mathbf{Y} = (y_1, \ldots, y_N)$ and any real function of it is called *parametric function*. Making inferences about a parametric function like total or the mean for example, is the objective of the survey sampling. Any other complicated functions such as the mode, the various population quantiles and the population variances may also be the subject of interest in survey sampling.

A small part of population $\mathcal{U}$ named as sample s is used to make inference about a parametric function. The sample s is obtained from the population by a probabilistic selection method. Let $\mathcal{S}$ be the set of all possible subsets s of $\mathcal{U}$,

$s \in \mathcal{P}(\mathcal{U})$. The number of possible subsets is 2^N including ϕ and $\mathcal{U}$; a sample is an element of $\mathcal{S}$.

Let $p(s)$ be the probability of selecting $s \in \mathcal{S}$ given $\mathcal{U}$. Saying otherwise, the function $p(s)$ is called the sampling design satisfying the following conditions:

(a) $p(s) \geq 0 \ \forall s \in \mathcal{S}$

(b) $\sum_{s \in \mathcal{S}} p(s) = 1$

Cassel *et al* (1977) refers a sampling design $p(s)$ which is not a function of $\mathbf{Y}$ as a non-informative design. The sample size noted by n, denotes the number of elements in s, may be fixed or not for the samples $s \in \mathcal{S}$. The sample membership indicator (Deville and Särndal ,1992) is denoted by

$$I_k = \mathbf{1}_{(k \in s)} \ \forall k \in \mathcal{U}$$

where the random variable I_k is a Bernoulli variable indicating if the kth unit belongs to the sample or not. Assuming that the sampling design has been fixed, the probabilities of inclusion may be defined as follows:

(I). π_k: is the probability that the kth element is included in a sample. That is, $\pi_k = \sum_{s \ni k} p(s)$, for $k \in \mathcal{U}$, and π_k is called the first order probability of inclusion

(II). π_{kl}: is the second order inclusion probability defined as the probability that the elements k and l will be included in a sample. That is, $\pi_{kl} = \sum_{s \ni \{k,l\}} p(s)$, for $k \in \mathcal{U}$ and $l \in \mathcal{U}$.

Result 1 (Properties of the indicator function I_k)**.** *For a sampling design $p(.)$, the indicator function I_k satisfies the following properties:*

(i). $E(I_k) = \pi_k$

(ii). $V(I_k) = \pi_k(1 - \pi_k)$

(iii). $Cov(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$, $k \neq l$, $\forall \ k, l \in \mathcal{U}$

Proof. The proof comes from the reality that I_k is a *Bernoulli variable*.

(i). $E(I_k) = P(I_k = 1) = \pi_k$ since $\pi_k = P(k \in s) = \sum_{k \in s} p(s)$.

(ii). Also since $\pi_{kl} = P(k, l \in s) = P(I_k I_l = 1) = \sum_{k,l \in s} p(s)$ and $\pi_{kl} = \pi_{lk}$ for k, l. This implies that when $k = l$,

$$\pi_{kl} = P(I_k^2 = 1) = P(I_k = 1) = \pi_k$$

Hence,

$$E(I_k^2) = \pi_k = E(I_k)$$

It follows,

$$V(I_k) = E(I_k^2) - (E(I_k))^2 = \pi_k - (\pi_k)^2 = \pi_k(1 - \pi_k)$$

(iii). Moreover, $P(I_k I_l = 1) = \pi_{kl}$ if and only if both k and l are members of s. Thus

$$E(I_k I_l) = P(I_k I_l = 1) = \pi_{kl}$$

which leads us to the quantity,

$$Cov(I_k I_l) = E(I_k I_l) - E(I_k)E(I_l) = \pi_{kl} - \pi_k \pi_l = \Delta_{kl},\ k \neq l\ \forall k, l \in \mathcal{U}$$

Note that for all $k = l$

$$Cov(I_k I_l) = V(I_k).$$

□

For sake of simplicity in notations, we define the Δ-quantities as:

$$\begin{aligned} \Delta_{kl} &= \pi_{kl} - \pi_k \pi_l \\ \check{\Delta}_{kl} &= \frac{\Delta_{kl}}{kl}\ \forall k, l \in \mathcal{U}. \end{aligned}$$

We suppose from here onwards that $\pi_k > 0$ for all $k \in \mathcal{U}$, namely each unit in the population has a chance to be in the sample.

1.2 Estimation of population total

Let us consider the finite population total,

$$t_y = \sum_{\mathcal{U}} y_k.$$

For the section below we shall restrict our study in the context of the fixed population approach so the only randomness is due to the sampling design, $p(.)$. Consequently, definitions of expectation, variance, and mean square error of an estimator $\mathcal{T}$ of t_y can be formulated for a given design $p(s)$. For example, the expectation of $\mathcal{T}$ is

$$E(\mathcal{T}) = \sum_{s \in S} p(s)\mathcal{T}(s)$$

1.2.1 The Horvitz-Thompson Estimator

Among the class of linear estimators, we consider the one proposed by Horvitz and Thompson (1952). It is called Horvitz-Thompson estimator or π estimator for the total t_y because of the first order inclusion probabilities appearing in its formula,

$$\hat{t}_{y\pi} = \hat{t}_{HT} = \sum_{s} \frac{y_k}{\pi_k}.$$

The equivalent expression of the π estimator for the total t_x can be written as,

$$\hat{t}_{x\pi} = \sum_{s} \frac{\mathbf{x}_k}{\pi_k}.$$

Properties of the Horvitz-Thompson Estimator

Result 2. *The π estimator $\hat{t}_{y\pi}$ of the population total t_y has the following properties:*

i. $\hat{t}_{y\pi}$ is design unbiased for $t_y = \sum_U y_k$.

ii. The variance of $t_{y\pi}$ can be written as,

$$V(\hat{t}_{y\pi}) = \sum_{U} \sum_{U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

iii. If $\pi_{kl} > 0$ for all $k, l \in U$, an unbiased estimator for $V(t_{y\pi})$ is,

$$\hat{V}(\hat{t}_{y\pi}) = \sum_s \sum_s \check{\Delta}_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Proof. The proof is entirely based on the use of the indicator variable I_k:

i. We have

$$\hat{t}_{y\pi} = \sum_s \frac{y_k}{\pi_k} = \sum_U \frac{y_k}{\pi_k} I_k,$$

so

$$\begin{aligned} E(\hat{t}_{y\pi}) &= \sum_U \frac{y_k}{\pi_k} E(I_k) \\ &= \sum_U y_k \end{aligned}$$

so $\hat{t}_{y\pi}$ is unbiased for t_y.

ii. The variance has the expression,

$$\begin{aligned} V(\hat{t}_{y\pi}) &= V\left(\sum_U \frac{y_k}{\pi_k} I_k\right) \\ &= \sum_U \frac{y_k^2}{\pi_k^2} V(I_k) + \sum_{k \in U} \sum_{l \in U, l \neq k} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} Cov(I_k I_l), \end{aligned}$$

recalling the properties of I_k, $V(I_k) = \pi_k(1 - \pi_k)$ and $Cov(I_k I_l) = \pi_{kl} - \pi_k \pi_l$, we get,

$$\begin{aligned} V(\hat{t}_{y\pi}) &= \sum_U \frac{y_k^2}{\pi_k^2} (\pi_k(1 - \pi_k)) + \sum_{k \in U} \sum_{l \in U, l \neq k} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} (\pi_{kl} - \pi_k \pi_l), \\ &= \sum_U \sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}. \end{aligned}$$

iii. Since the estimator of variance has the expression,

$$\begin{aligned} \hat{V}(\hat{t}_{y\pi}) &= \sum_s \sum_s \check{\Delta}_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \\ &= \sum_U \sum_U \check{\Delta}_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} I_k I_l. \end{aligned}$$

We apply expectation on both sides,

$$\begin{aligned} E(\hat{V}(\hat{t}_{y\pi})) &= \sum_U \sum_U \breve{\Delta}_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} E(I_k I_l) \\ &= \sum_U \sum_U \breve{\Delta}_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \pi_{kl} \\ &= \sum_U \sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \\ &= V(\hat{t}_{y\pi}). \end{aligned}$$

So, $\hat{V}(\hat{t}_{y\pi})$ is unbiased estimator of the $\hat{t}_{y\pi}$

□

Yates and Grundy (1953) and Sen (1953) argued that equivalent formulas can be obtained for the variance and variance estimator of $\hat{t}_{y\pi}$ for a sampling design of fixed size, $n_s = n$.

Remark 1. *(Yates and Grundy (1953) and Sen (1953))*

If $p(s) > 0$ is of fixed sample size, then $V(\hat{t}_{y\pi})$ and $\hat{V}(\hat{t}_{y\pi})$ have the following expression,

i.

$$V(\hat{t}_{y\pi}) = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

ii. If $\pi_{kl} > 0$ for all $k, l \in U$, then

$$\hat{V}(\hat{t}_{y\pi}) = -\frac{1}{2} \sum_s \sum_s \breve{\Delta}_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

These results can be proved using the properties of indicator function and inclusion probabilities.

1.2.2 Simple Random Sampling Without Replacement (SRSWOR)

In this scheme of sampling, we select the first element of the sample with equal probability $\frac{1}{N}$ from the population and the selected element is kept away during the following selections. Again, we select another unit with equal probability from the remaining $N-1$ entities of the population and we repeat the procedure again and again until the required sample of size n is acquired. The design has the probability function expression as follows,

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

For SRSWOR, $\pi_k = \frac{n}{N}$ and $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$ are the expressions for the first and second order inclusion probabilities respectively. Lohr (1999) describes that for large populations it is the size of sample taken instead of the percentage of the population sampled, which determines the precision of the estimator: a sample of size 100 from a population of 100,000 units has almost the same precision compared to a sample of size 100 from a population of 100 million units.

Result 3. *Under the simple random sampling without replacement we have,*

1. $\hat{t}_{y\pi,SRSWOR} = N\bar{y}_s = \frac{1}{f}\sum_s y_k$ *where* $f = \frac{n}{N}$ *is called the sampling rate.*

2. *The variance of* $\hat{t}_{y\pi,SRSWOR}$ *is,*

$$Var_{SRSWOR}(\hat{t}_{y,\pi}) = N^2\frac{1-f}{n}S^2_{yU}$$

   where $S^2_{yU} = \frac{1}{N-1}\sum_U (y_k - \bar{y}_U)^2$ *and* $\bar{y}_u = \frac{1}{N}\sum_U y_k$.

3. *An unbiased estimator of* $Var_{SRSWOR}(\hat{t}_{y,\pi})$ *is,*

$$\hat{Var}_{SRSWOR}(\hat{t}_{y,\pi}) = N^2\frac{1-f}{n}S^2_{ys}$$

   where $S^2_{ys} = \frac{1}{n-1}\sum_s (y_k - \bar{y}_s)^2$ *and* $\bar{y}_s = \frac{1}{N}\sum_s y_k$.

1.2.3 Use of Auxiliary Information

A desirable characteristic of the survey sampling is the use of the auxiliary information for improving the precision of the estimators. Design and estimation in the sampling survey coordinate each other to make use of the information about the study population to construct the efficient procedures. The estimation goal can be to combine the in-hand information about the population with sample data to generate good representations of characteristics of interest. The in-hand information may be regarded as the auxiliary information.

Sometimes, the sampling frames contain one or more auxiliary variables, or any information that simply can be transferred into auxiliary variables. That is, the frame provides identification characteristics of the units with the each unit attached with the value of one or more auxiliary variables.

Three distinct situations are identified by Fuller (2002) with respect to the nature of the availability of the auxiliary information.

1. The values of the auxiliary vector that are known for each element in the population at the time of sample selection. That is, the value of the variable, say $\mathbf{X}_1$, is known for each of the N population elements so that the values $X_{11}, \ldots, X_{N1}$ are at our disposal prior to sampling. An auxiliary variable assists in designing the sample selection procedure and can be used in the estimation of the study variable. The goal is to obtain an estimator with increased accuracy.

2. All values of the vector are known, but a particular value cannot be associated with a particular element until the sample is observed. In this case, auxiliary information cannot be used in design, but a wide range of estimation options are available once the observations are available.

3. Only the population mean of $\mathbf{X}$ is known, or known for a large sample. In this case, the auxiliary information cannot be used in design and the estimation options are limited.

Two estimation situations can also be confronted.

a. A single variable and a parameter, or a very small number of parameters, is under consideration. The analyst has a well formulated population model, and is prepared to support the estimation procedure on the basis of the reasonableness of the model.

b. A large number of analyses of a large number of variables is anticipated. No single model is judged adequate for all variables.

We now assume that one or more auxiliary variables are present. The auxiliary information can be used at the design stage of a survey to create a sampling design that increases the precision of the Horvitz-Thompson estimator or at the estimation stage.

One approach is πps sampling, that is, to make the inclusion probabilities $\pi_1, ..., \pi_N$ of the design proportional to known, positive values $x_1, ..., x_N$ of an auxiliary variable. The π estimator will then have a small variance if x is more or less proportional to y , the study variable. However, πps sampling is sometimes found difficult to be carried out. Another approach is to use auxiliary information to construct the strata such that the π estimator for a stratified simple random sampling design,

$$\hat{t}_{y\pi} = \sum_{h=1}^{H} \boldsymbol{N_h} \bar{y}_{s_h}$$

obtains a small variance. However, the stratification that is efficient for one study may be inefficient for another. One of the important procedures that use population information from a large sample is regression estimation. The regression estimators are classified as linear estimators. We shall use the auxiliary information explicitly at the estimation stage i-e into the estimator formula, for the given π_k. That is, for a given sampling design, we construct estimators that utilize information from auxiliary variables and bring considerable variance reduction compared to the π-estimator. The basic assumption behind the use of auxiliary variables is that they covary with the study variable and thus carry information

about the study variable. Such covariation is used advantageously in the regression estimator.

1.2.4 Model Definition

A model ξ defines a class of distributions of $\mathbf{Y} = (y_1, \ldots, y_N)$. In other words by a superpopulation model or simply a model we mean specified set of conditions that define a class of distributions of $\mathbf{Y} = (y_1, \ldots, y_N)$ (Cassel *et al*, 1977). This class of distribution may perform a crude formulation and may also prescribe some certain features including means, variances and the covariances of ξ. There may also be a situation when ξ assumes a highly detailed specification. In this case, we shall treat $\mathbf{Y} = (y_1, \ldots, y_N)$ as a random quantity in addition to the randomness of the sampling design $p(.)$. This new randomness is subject to the uncertainty introduced by the probabilistic model. Cochran (1939, 1946), Deming and Stephan (1947) and Madow and Madow (1944) are few from the long listed history of initial users of the superpopulation model. A superpopulation model may also be defined as a mathematical device which is used to make theoretical derivations. Cassel *et al* (1977) however classify the superpopulation inference into being the non-Bayesian (ξ is assumed to contain unknown parameters which are necessary to be estimated first) and the Bayesian (a prior distribution is assigned to the unknown model parameters) inference tools.

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ be an N×p matrix of regressors. Let we have a superpopulation model as,

$$\xi : \mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \tag{1.1}$$

where, $\mathbf{y}$ is the N×1 vector of observations random variable; $\mathbf{X} = (\mathbf{x}'_k)_{k\in U}$:N×p matrix of regressors and $\boldsymbol{\varepsilon}$ is N×1 vector of random residuals. We take into account some common assumptions. These assumptions include that $\mathbf{X}$ is non-stochastic matrix of regressors, $\mathbf{X}'\mathbf{X}$ is a full rank matrix (i.e rank of $\mathbf{X}$ is p), with $V(\varepsilon_k) = \sigma^2 v_k^2, \forall k = 1, \ldots, N$, and the independence between different residual terms, i.e. $Cov(\varepsilon_k, \varepsilon_l) = 0, \forall k \neq l = 1, ..., N$.

1.3 Generalized difference estimator and generalized regression estimator

For the above defined model (1.1), the generalized difference estimator as suggested by Cassel *et al* (1976) can be defined as,

$$\hat{t}_{DIFF} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{x}'_k}{\pi_k} - \sum_U \mathbf{x}'_k \right) \boldsymbol{\beta}. \tag{1.2}$$

This estimator contains unknown regression coefficients $\boldsymbol{\beta}$ which makes it difficult to compute. The Horvitz-Thompson estimator $\hat{t}_{y\pi}$ is model biased with its bias given as

$$Bias_\xi(\hat{t}_{y\pi}) = E_\xi(\hat{t}_{y\pi} - t_y) = \left(\sum_s \frac{\mathbf{x}'_k}{\pi_k} - \sum_U \mathbf{x}'_k \right) \boldsymbol{\beta}.$$

So, we can see that the generalized difference estimator $\hat{t}_{DIFF}$ is clearly $\hat{t}_{y\pi}$ minus its ξ-bias, that is,

$$\hat{t}_{DIFF} = \hat{t}_{y\pi} - Bias_\xi(\hat{t}_{y\pi}). \tag{1.3}$$

So, the $\hat{t}_{DIFF}$ can be taken as an attempt to improve the basic Horvitz-Thompson estimator $\hat{t}_{y\pi}$. The unknown $\boldsymbol{\beta}$ is estimated by a two-step procedure

1. at the population level:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{1.4}$$

   where $\mathbf{V} = diag(v_j^2)$, $j = 1, \ldots, p$ and

2. at the sample level:

$$\hat{\boldsymbol{\beta}}_\pi = \left(\mathbf{X}'_s\mathbf{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}'_s\mathbf{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{y}_s \tag{1.5}$$

   assuming that $\mathbf{X}'_s\mathbf{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s$ is invertible.

If the unknown $\boldsymbol{\beta}$ are replaced by the estimated $\hat{\boldsymbol{\beta}}_{\boldsymbol{\pi}}$ in $\hat{t}_{DIFF}$, we obtain a new estimator called generalized regression (GREG) estimator given as,

$$\hat{t}_{GREG} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{x}_k'}{\pi_k} - \sum_U \mathbf{x}_k' \right) \hat{\boldsymbol{\beta}}_{\boldsymbol{\pi}}, \tag{1.6}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\pi}}$ is given by (1.5). Like the generalized difference estimator $\hat{t}_{DIFF}$, the GREG estimator $\hat{t}_{GREG}$ is also the addition of the Horvitz-Thompson estimator $\hat{t}_{y\pi}$ and adjustment term. For the large sample and the strong linear relationship between $\mathbf{Y}$ and $\mathbf{X}$, $\hat{t}_{GREG}$ produces smaller error than $\hat{t}_{y\pi}$ (Särndal et al, 1992). The regression estimator has been extensively used in previous few decades since it came into existence by Jessen (1942) and Cochran (1942). The use of regression in survey was given by Cochran (1942) and he showed that it works well even when the model fails. A substantial amount of work on the regression estimator in survey samples was done in the $1970's$ $1980's$ to improve the compatibility of model prediction in the design environment (Fuller, 2002) . Large sample properties of a regression coefficient vector obtained via a sample survey are given by Fuller (1973, 1975). Both model and design principals in the construction of an estimator were used by Cassel *et al* (1976) and named *generalised regression estimator* (GREG) for the consistent estimators of the form given in 1.6.

1. The variance of $\hat{t}_{GREG}$ can be approximated as follows,

$$V(\hat{t}_{GREG}) \simeq \sum_U \sum_U \Delta_{kl} \frac{y_k - \mathbf{x}_k' \hat{\boldsymbol{\beta}}_{GLS}}{\pi_k} \frac{y_l - \mathbf{x}_l' \hat{\boldsymbol{\beta}}_{GLS}}{\pi_l}.$$

2. If $\pi_{kl} > 0$ for all $k, l \in U$, an unbiased estimator for $V(\hat{t}_{GREG})$ is,

$$\hat{V}(\hat{t}_{GREG}) = \sum_s \sum_s \check{\Delta}_{kl} \frac{y_k - \mathbf{x}_k' \hat{\boldsymbol{\beta}}_{\pi}}{\pi_k} \frac{y_l - \mathbf{x}_l' \hat{\boldsymbol{\beta}}_{\pi}}{\pi_l}.$$

1.4 Calibration Technique

The calibration technique derived by Deville and Särndal (1992) with the motive of obtaining an estimator of the population total using some sample weights called

calibrated weights. These weights are obtained by minimizing the distance to the Horvitz-Thompson weights ($d_k = \frac{1}{\pi_k}$) with the additional condition on the calibration equations to be satisfied. The resulting sample weights will be function of the auxiliary variables. If the weights exactly satisfy the calibration equations, it would mean the exact estimation of the auxiliary variables. Särndal (2007) precises the scope of the calibration technique by saying that it takes into account the following points.

1. Finding a new set of weights w_k which minimize the distance between w_k and d_k using the auxiliary information in terms of the calibration equations. The additional condition on finding these weights is that the calibration equations

$$\sum_s w_k \mathbf{x}_k = t_x$$

   are satisfied.

2. The weights are then used to compute different types of linear weighted estimators of the parameter including totals.

3. The computation of nearly design unbiased estimates in the absence of non-response and other non-sampling errors.

The similar type of desirable properties of the calibration technique are described by Singh and Mohl (1996, p. 107).

Deville and Särndal (1992) considered a nonnegative distance function $G_k(w, d)$, such that the weights w_k are chosen by the minimization of this distance function from the basic design weights.

i. $G_k(w, d)$ is nonnegative, its derivative with respect to w exists, strictly convex, defined on an interval $D_k(d)$ which contains d;

ii. $G_k(d, d) = 0$, i.e. the distance function between the same design weights is zero.

iii. The derivative $g_k(w,d) = \frac{\partial G_k(w,d)}{\partial w}$ is continuous and the interval $D_k(d)$ is mapped onto an interval $\mathbf{I}_k(d)$ by a one-to-one function.

The minimization of the average distance $E_p\left[\sum_k G_k(w,d)\right]$ in fact offers the closeness between the requested weights w_k and the design weights d_k. The method of Lagrange multipliers is used by Deville and Särndal (1992) to find the unique solution of weights w_k called calibration weights if exists, given as,

$$w_k = d_k F_k(\mathbf{x}_k'\boldsymbol{\lambda})$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_j, \ldots, \lambda_p)$ is the vector of Lagrange multipliers, $d_k F_k$ is the reciprocal mapping of $g_k(., d_k)$ and $F_k(0) = 1$, with $q_k = F_k(0) > 0$. We have

$$\hat{t}_{xw} = \sum_s w_k \mathbf{x}_k = \sum_s d_k F_k(\mathbf{x}_k'\boldsymbol{\lambda})\mathbf{x}_k = \sum_U \mathbf{x}_k. \tag{1.7}$$

Certain conditions are given by Deville and Särndal (1992) to ensure that 1.7 yields a unique solution belonging to a convex domain $\mathbf{C} = \bigcap_{k\in\mathcal{U}} [\boldsymbol{\lambda} : \mathbf{x}_k'\boldsymbol{\lambda} \in \mathbf{I}_{m_k}(d_k)]$. Once $\boldsymbol{\lambda}$ determined,the calibration estimator of t_y can be written as,

$$t_{yw} = \sum_s w_k y_k = \sum_s d_k F_k(\mathbf{x}_k'\boldsymbol{\lambda}) y_k. \tag{1.8}$$

Deville and Särndal (1992) gave the discussion on how the difference in the choice of distance function leads to different estimators. The case when $F_k(u) = 1 + q_k u$ where $u = \mathbf{x}_k'\boldsymbol{\lambda}$ and $\boldsymbol{\lambda} = \left(\sum_U \mathbf{x}_k - \sum_s d_k x_k\right)' \left(\sum_s d_k q_k x_k x_k'\right)^{-1}$ (Särndal, 2007) yields the generalized regression estimator,

$$\hat{t}_{yreg} = \sum_s w_k y_k = \hat{t}_{y\pi} + \left(t_x - \hat{t}_{x\pi}\right)' \hat{\boldsymbol{\beta}}_\pi \tag{1.9}$$

where $\hat{t}_{y\pi}$ and $\hat{t}_{x\pi}$ are the π-estimators for the population total of y_k and x_k respectively, and $\hat{\boldsymbol{\beta}}_\pi = \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_s d_k q_k \mathbf{x}_k y_k$. So, for $q_k = \frac{1}{v_k^2}$, this calibration method and the regression technique (Särndal, 1980) lead to the same estimator. For some other distance function, Deville and Särndal (1992) prove that the calibration estimator $\hat{t}_{yw}$ is asymptotically equivalent to $\hat{t}_{yreg}$.

For some values of k, it is possible that $F_k(u)$ is negative which is undesirable (Singh and Mohl, 1996). Some optimal/desirable properties about the sample weights in estimation are given by Lohr (2007). The minimization of *MSE* is one of the core property of weighted estimators. The negative or more precisely undesirable weights can hammer the optimality of the calibrated estimators. Changes in the choice of right distance function can guaranty that the weights are neither too large nor too small. This change in the distance function will however have a little influence on the variance of the calibration estimator despite of the small sample size (Särndal, 2007).

Chapter 2

Ridge regression in survey sampling

Regression techniques are widely used in practice due to their large and easy applicability. They are often based on ordinary least squares method. Nevertheless, in presence of multicollinearity of data, the ordinary least squares estimator of the regression estimator can have extremely large variance even if it has the desirable property of being the minimum variance estimator in the class of linear unbiased estimators (the Gauss-Markov theorem). Biased estimators have been suggested to cope with that problem and the class of ridge estimators is one of them. Hoerl and Kennard (1970) suggest in a seminal paper the ridge estimator of the regression coefficient which depends on a penalty parameter that controls the trade-off between the bias and the variance. They show that for suitable values of the penalty parameter, the ridge estimator has smaller mean squared error than that of the ordinary least squares estimator. The method has been applied in many fields such as agriculture, engineering (Marquardt and Snee, 1975) and astrophysics (Matthews and Newman, 2012) among others. The book of Vinod

This chapter contains the article, Camelia Goga and Muhammad Ahmed Shehzad (2011), Overview of ridge regression estimators in survey sampling. Mathematical population studies (under review).

and Ullah (1981) gives a comprehensive description on this topic as well as many examples.

In a survey sampling setting, weighted estimators using auxiliary information are built in order to give precise estimations about parameters of interest such as totals, means, ratios and so on. Usually, these weighted estimators are equivalent to regression estimators but it happens that, in the presence of a large amount of information, the weights are very unstable, negative or very large (Deville and Särndal, 1992, page 378). Moreover, data may contain many zeros or, the sample sizes may be smaller than the number of auxiliary variables (for example, in the case of estimation for small domains), which may entail in certain situations problems of matrix invertibility.

In Section 2 we recall the construction of the ridge estimator for the regression coefficient as introduced by Hoerl and Kennard (1970) in a classical regression setting. At this occasion, we give the equivalent interpretations of this estimator such as the constrained minimization problem and the Bayesian point of view. We recall briefly the ridge trace as a method to find the penalty parameter. Section 3 gives a detailed presentation of the application of the ridge principle in survey sampling. This presentation includes the derivation of penalized estimators under the model-based approach given in section 2.3.1 as well as under the calibration approach, section 2.3.2. The geometry of penalized weights is given in section 2.3.2. Section 2.3.3 exhibits the partial calibration or balancing. When we attribute a prior on previous estimations, we may use the Bayesian interpretation to construct ridge regression type estimators. Deville (1999, page 208) considered it as a calibration on an uncertain source. We describe the method in section 2.3.4. Finally, section 2.3.5 gives the statistical properties of the class of penalized estimators and we finish with concluding remarks and some further work.

2.1 Ridge Regression in an i.i.d setting

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ be a $n \times p$ matrix of standardized known regressors i. e. $\mathbf{X}_j = (X_{kj})_{k=1}^n$ for all $j = 1, \ldots, p$. Consider the following linear model,

$$\mathbf{y} = \mathbf{1}_n \beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y} = (y_k)_{k=1}^n$ is the $n \times 1$ vector of observations and $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k=1}^n$ is the $n \times 1$ vector of errors. We assume that $\mathbf{X}$ is a non-stochastic matrix of regressors with $\mathbf{X}'\mathbf{X}$ of full rank matrix (i.e the rank of $\mathbf{X}$ is p). We suppose also that the errors ε_k are independent with zero mean and variance $\mathrm{Var}(\varepsilon_k) = \sigma^2$ for all $k = 1, \ldots, n$. The ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ minimizes the error sum of squares (ESS),

$$ESS = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

yielding the following estimator,

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ is unbiased under the model ξ, i.e.

$$E(\hat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta}$$

with the variance of $\hat{\boldsymbol{\beta}}_{OLS}$ given by,

$$Var(\hat{\boldsymbol{\beta}}_{OLS}) = E(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})' = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

However, the calculation of the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ solely depends upon the existence of the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ which may not be possible if the data matrix $\mathbf{X}$ is ill-conditioned.

2.1.1 Multicollinearity, ill-conditioning and consequences on the OLS estimator

Zero or no dependence among the explanatory variables is one of the assumptions of classical linear regression model. The subject of multicollinearity is widely

referred to the situation where there is either exact or approximately exact linear relationship among the explanatory variables (Gujarati, 2003).

Gunst and Mason (1977) discriminate between the existence and the degree of the multicollinearity found in the auxiliary variables. They state that *"the closer the linear combinations between the columns of* $\mathbf{X}$ *are to zero, the stronger are the multicolinearities and the more damaging are their effects on the least squares estimator"*. It should be kept in mind while detecting the multicollinearity that the question should be of the degree/intensity of multicollinearity and not of kind of the multicollinearity. Small eigenvalues and their corresponding eigenvectors help to identify the multicollinearities. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $\mathbf{X}'\mathbf{X}$ in decreasing order,

$$\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p = \lambda_{min} > 0$$

and their corresponding eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$. If we write (Gunst and Mason, 1977),

$$\lambda_j = \mathbf{a}_j'\mathbf{X}'\mathbf{X}\mathbf{a}_j = (\mathbf{X}\mathbf{a}_j)'(\mathbf{X}\mathbf{a}_j),\ j = 1, ..., p$$

we obtain that for small eigenvalues λ_j of $\mathbf{X}'\mathbf{X}$,

$$(\mathbf{X}\mathbf{a}_j)'(\mathbf{X}\mathbf{a}_j) \approx 0\ \Rightarrow \mathbf{X}\mathbf{a}_j \approx 0$$

which means that there is an approximately linear relationship between the columns of $\mathbf{X}$. The elements of the corresponding eigenvector $\mathbf{a}_j$ allow to identify the coefficients used in the linear dependency.

The multicollinearity is one form of ill-conditioning. More general, a measure of ill-conditioning is the *conditioning number* K given by $K = \sqrt{\lambda_{max}/\lambda_{min}}$. For $\lambda_{min} \to 0$, we have $K \to \infty$, and so, a large K implies an ill-conditioned matrix $\mathbf{X}$.

The multicollinearity or the ill-conditioning of $\mathbf{X}$ have serious consequences on the OLS estimator. The mean square error (MSE) of any estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})).$$

Then, the MSE of the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ becomes

$$\text{MSE}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2 \text{Trace}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}. \tag{2.2}$$

The above expression implies that the smaller the eigenvalues are, the greater are the variance of $\hat{\boldsymbol{\beta}}_{OLS}$ and the average value of the squared distance from $\hat{\boldsymbol{\beta}}_{OLS}$ to $\boldsymbol{\beta}$. This results in wider confidence intervals and therefore leads to accept more often the *Null Hypothesis* (i.e. the true population coefficient is zero). Moreover, in case of ill-conditioning, the OLS solution is unstable meaning that the regression coefficients are sensitive to small changes in the $\mathbf{y}$ or $\mathbf{X}$ data (see Marquardt and Snee, 1975 and Vinod and Ullah, 1981). Round-off errors tend to occur into least square calculations while the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ is computed and they may be important in presence of non-orthogonal data. Hoerl and Kennard (1970) discuss the case when the least square coefficients can be both too large in absolute value and incorrect with respect to sign.

Methods dealing with such data consist in (1) using a *priori* information (Bayesian approach), (2) omitting highly collinear variables, (3) obtaining additional or new data and (4) using biased regression methods. These methods can be used individually or together depending upon the encountered situation. Our discussion however remains limited towards the fourth case and ridge regression which is an important tool to deal with multicollinearity.

2.1.2 Definition of the ridge estimator

Ridge regression was first used by Hoerl (1962) and then by Hoerl and Kennard (1970) as a solution to the biased estimation for nonorthogonal data problems. As a purpose to control instability linked to the least squares estimates, Hoerl (1962) and Hoerl and Kennard (1968) suggested an alternative estimate of the regression coefficient as obtained by adding a positive constant κ to the diagonal elements of the least square estimator $\hat{\boldsymbol{\beta}}_{OLS}$,

$$\hat{\boldsymbol{\beta}}_{\kappa} = (\mathbf{X}'\mathbf{X} + \kappa \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}, \tag{2.3}$$

where $\mathbf{I}_p$ is the p-dimensional identity matrix. Since the constant κ is arbitrary, we obtain a class of estimators $\hat{\boldsymbol{\beta}}_\kappa$ for the regression coefficient $\boldsymbol{\beta}$ rather than a unique estimator. For $\kappa = 0$, we obtain the OLS estimator and as $\kappa \to \infty$, $\hat{\boldsymbol{\beta}}_\kappa \to 0$, we obtain the null vector.

The relationship between the ridge estimator and the OLS estimator is given by (Hoerl and Kennard, 1970),

$$\hat{\boldsymbol{\beta}}_\kappa = (\mathbf{I}_p + \kappa(\mathbf{X}'\mathbf{X})^{-1})^{-1}\hat{\boldsymbol{\beta}}_{OLS}.$$

Let us consider again the latent roots of $(\mathbf{X}'\mathbf{X})$, $\lambda_1, \ldots, \lambda_p$ with the corresponding eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$. Hence, the OLS estimator may be written as

$$\hat{\boldsymbol{\beta}}_{OLS} = \sum_{j=1}^{p} \frac{\mathbf{a}_j'\mathbf{X}'\mathbf{y}}{\lambda_j}\mathbf{a}_j. \tag{2.4}$$

The fact of adding a small constant to the diagonal of $\mathbf{X}'\mathbf{X}$ will have as consequence the increase of its eigenvalues with the same quantity and dramatically decrease in this way the conditioning number K. So, the matrix $\mathbf{X}'\mathbf{X} + \kappa\mathbf{I}$ has eigenvalues $\lambda_1 + \kappa, \ldots, \lambda_p + \kappa$ with the same eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$ and the ridge estimator may be written as follows

$$\hat{\boldsymbol{\beta}}_\kappa = \sum_{j=1}^{p} \frac{\mathbf{a}_j'\mathbf{X}'\mathbf{y}}{\lambda_j + \kappa}\mathbf{a}_j. \tag{2.5}$$

The effect of the smallest eigenvalues may not be entirely eliminated by this estimator $\hat{\boldsymbol{\beta}}_\kappa$ but their effect on the parameter estimates are significantly lessened. By this construction, the ridge estimator $\hat{\boldsymbol{\beta}}_\kappa$ is more stable than the OLS estimator to perturbations of data (Vinod and Ullah, 1981). Hoerl and Kennard (1970) show also that for $\kappa \neq 0$, the length of the ridge estimator $\hat{\boldsymbol{\beta}}_\kappa$ is shorter than that of $\hat{\boldsymbol{\beta}}_{OLS}$, namely $\hat{\boldsymbol{\beta}}_\kappa'\hat{\boldsymbol{\beta}}_\kappa < \hat{\boldsymbol{\beta}}'_{OLS}\hat{\boldsymbol{\beta}}_{OLS}$.

Let recall briefly the statistical properties of the ridge estimator. It is important to note that the ridge estimator $\hat{\boldsymbol{\beta}}_\kappa$ is a biased estimator of $\boldsymbol{\beta}$ unless $\kappa = 0$.

The bias of $\boldsymbol{\beta}$ can be obtained as,

$$\begin{aligned}
& \hat{\boldsymbol{\beta}}_\kappa - \boldsymbol{\beta} \\
= \; & (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta} \\
= \; & (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \\
= \; & (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\
= \; & (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}((\mathbf{X}'\mathbf{X} + \kappa\mathbf{I}) - \kappa\mathbf{I})\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\
= \; & (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})\boldsymbol{\beta} - \kappa(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\
\hat{\boldsymbol{\beta}}_\kappa - \boldsymbol{\beta} = \; & -\kappa(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\boldsymbol{\varepsilon},
\end{aligned}$$

and applying expectation on both sides,

$$E(\hat{\boldsymbol{\beta}}_\kappa) - \boldsymbol{\beta} = -\kappa(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}).$$

Because $E(\boldsymbol{\varepsilon}) = 0$. So, the bias is given by

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}_\kappa) - \boldsymbol{\beta} &= -\kappa(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\boldsymbol{\beta} && (2.6) \\
&= -\kappa\sum_{j=1}^{p}\frac{(\mathbf{a}_j'\boldsymbol{\beta})\mathbf{a}_j}{\lambda_j + \kappa}. && (2.7)
\end{aligned}$$

We can see from the above that the bias depends on the unknown $\boldsymbol{\beta}$ and on κ.

Consider again the equation (2.5). We can write

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_\kappa &= \sum_{j=1}^{p}\frac{\mathbf{a}_j'\mathbf{X}'\mathbf{y}}{\lambda_j + \kappa}\mathbf{a}_j \\
&= \sum_{j=1}^{p}\frac{\mathbf{a}_j'\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})}{\lambda_j + \kappa}\mathbf{a}_j \\
&= \sum_{j=1}^{p}\frac{\mathbf{a}_j'(\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\boldsymbol{\varepsilon})}{\lambda_j + \kappa}\mathbf{a}_j,
\end{aligned}$$

To calculate the variance of $\hat{\boldsymbol{\beta}}_\kappa$ in matrix form, consider,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_\kappa - E(\hat{\boldsymbol{\beta}}_\kappa) &= (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) + \kappa(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\boldsymbol{\beta} - \boldsymbol{\beta} \\
&= (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} + \kappa(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\boldsymbol{\beta} - \boldsymbol{\beta} \\
&= (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}((\mathbf{X}'\mathbf{X} + \kappa\mathbf{I}) - \kappa\mathbf{I})\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} + \kappa(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\boldsymbol{\beta} - \boldsymbol{\beta} \\
\hat{\boldsymbol{\beta}}_\kappa - E(\hat{\boldsymbol{\beta}}_\kappa) &= (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\boldsymbol{\varepsilon},
\end{aligned}$$

and therefore the $Var(\hat{\boldsymbol{\beta}}_\kappa)$ is given by,

$$E(\hat{\boldsymbol{\beta}}_\kappa - E(\hat{\boldsymbol{\beta}}_\kappa))(\hat{\boldsymbol{\beta}}_\kappa - E(\hat{\boldsymbol{\beta}}_\kappa))' = (\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1},$$

with $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\sigma}^2\mathbf{I}$, we get,

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}_\kappa) &= E(\hat{\boldsymbol{\beta}}_\kappa - E(\hat{\boldsymbol{\beta}}_\kappa))(\hat{\boldsymbol{\beta}}_\kappa - E(\hat{\boldsymbol{\beta}}_\kappa))' \\ &= \boldsymbol{\sigma}^2(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I})^{-1}. \end{aligned} \tag{2.8}$$

It appears that $\hat{\boldsymbol{\beta}}_\kappa$ can be used to improve the mean square error of the OLS estimator, and the magnitude of this improvement increases with an increase in spread of the eigenvalue spectrum. The ridge regression comes up with the objective of developing *stable* set of coefficient estimators which will do a reasonable job for predicting future observations. Conniffe and Stone (1973) however criticized the $\hat{\boldsymbol{\beta}}_\kappa$ since its properties depend on the non-stochastic choice of κ. Hoerl and Kennard (1970) and Hoerl, Kennard and Baldwin (1975) show that an improvement of the MSE can be obtained using $\hat{\boldsymbol{\beta}}_\kappa$. Consider for that the MSE of $\hat{\boldsymbol{\beta}}_\kappa$,

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\beta}}_\kappa) &= \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+\kappa)^2} + \kappa^2\sum_{j=1}^{p}\frac{(\mathbf{a}_j'\boldsymbol{\beta})^2}{(\lambda_j+\kappa)^2} \\ &= \text{Trace}(\text{Var}(\hat{\boldsymbol{\beta}}_\kappa)) + (\text{Bias}(\hat{\boldsymbol{\beta}}_\kappa))'(\text{Bias}(\hat{\boldsymbol{\beta}}_\kappa)) \\ &= \mathcal{A}(\kappa) + \mathcal{B}(\kappa). \end{aligned} \tag{2.9}$$

Hoerl and Kennard (1970) gave an existence theorem to show that such value of $\kappa > 0$ when added into the diagonal of the ill-conditioned matrix $\mathbf{X}'\mathbf{X}$, significant reductions in variance are found with a little charge of bias and an admirable improvement in the MSE of the estimation of the regression coefficient $\boldsymbol{\beta}$.

Theorem 1. *(existence theorem, Hoerl and Kennard, 1970) There always exists* $\kappa > 0$ *such that*

$$MSE(\hat{\boldsymbol{\beta}}_\kappa) < MSE(\hat{\boldsymbol{\beta}}) = \sigma^2\sum_{j=1}^{p}\frac{1}{\lambda_j}.$$

Moreover, the above inequality is valid for all $0 < \kappa < \kappa_{max} = \frac{\sigma^2}{\alpha_{max}^2}$ *where* α_{max} *is the largest value of* $(\mathbf{a}_1, \ldots, \mathbf{a}_p)\boldsymbol{\beta}$.

The proof is based on the fact that the variance term $\mathcal{A}(\kappa)$ from relation (2.9) is a continuous, monotonically decreasing function of κ and the squared bias term $\mathcal{B}(\kappa)$ is a continuous, monotonically increasing function of κ. Their first derivatives are always non-positive and non-negative, respectively. Moreover, the first derivative of $\mathcal{A}(\kappa)$ is negative as $\kappa \to 0^+$ and the first derivative of $\mathcal{B}(\kappa)$ is equal to zero as $\kappa \to 0^+$. Thus, there exists a positive κ in a neighborhood of the origin, such that the first derivative of $\text{MSE}(\hat{\boldsymbol{\beta}}_\kappa)$ is non-positive. In fact, this happens for all $0 < \kappa < \sigma^2/\alpha^2_{max}$.

It is important to notice that the features of $\mathcal{A}(\kappa)$ and $\mathcal{B}(\kappa)$ lead to the fact that moving from the origin to a positive κ, we introduce a little bias but we drastically reduce the variance and thereby, we improve the mean square error of the estimator.

However, Theobald (1974) criticized the MSE criteria used by Hoerl and Kennard (1970) and suggested a more general criteria. Theobald (1974) suggested minimizing the weighted mean square error (WMSE) defined by

$$WMSE(\hat{\boldsymbol{\beta}}) = E\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right),$$

for any non-negative definite matrix $\mathbf{W}$. For $\mathbf{W} = \mathbf{I}_p$ the identity matrix, we obtain the MSE criteria. He showed that minimizing the WMSE, for all non-negative definite matrix $\mathbf{W}$ is equivalent to minimizing the mean square error matrix (MMSE) defined by

$$MMSE(\hat{\boldsymbol{\beta}}) = E\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right)$$

and he obtained a range for the ridge parameter κ which guarantees that $\hat{\boldsymbol{\beta}}_\kappa$ is better than $\hat{\boldsymbol{\beta}}_{OSL}$ from the WMSE point of view.

Theorem 2. *(Theobald, 1974) The ridge estimator $\hat{\boldsymbol{\beta}}_\kappa$ is better than $\hat{\boldsymbol{\beta}}_{OSL}$ in the sense that $MMSE(\hat{\boldsymbol{\beta}}_{OSL}) - MMSE(\hat{\boldsymbol{\beta}}_\kappa)$ is a positive-definite matrix for*

$$0 < \kappa < \tilde{\kappa}_{max} = \frac{2\sigma^2}{\boldsymbol{\beta}'\boldsymbol{\beta}}.$$

Proof. Since by expressions 2.6 and 2.8 we have,

$$MMSE(\hat{\boldsymbol{\beta}}_\kappa) \;=\; \boldsymbol{\sigma}^2(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1} + Bias(\hat{\boldsymbol{\beta}}_\kappa)Bias(\hat{\boldsymbol{\beta}'}_\kappa)$$

and,

$$MMSE(\hat{\boldsymbol{\beta}}_{OLS}) = Var(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

hence,

$$\begin{aligned}
& MMSE(\hat{\boldsymbol{\beta}}_{OLS}) - MMSE(\hat{\boldsymbol{\beta}}_\kappa) \\
= \;& \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \boldsymbol{\sigma}^2(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1} - \kappa^2(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}\boldsymbol{\beta}\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1} \\
= \;& \sigma^2[(X^{'}X)^{-1} - (\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}] - \kappa^2(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}\boldsymbol{\beta}\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1} \\
= \;& (\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}(\sigma^2[(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I}) - (\mathbf{X}'\mathbf{X})] - \kappa^2\boldsymbol{\beta}\boldsymbol{\beta}')(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1} \\
= \;& (\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}(\sigma^2[(\mathbf{X}'\mathbf{X})^{-1}((\mathbf{X}'\mathbf{X})^2 + 2\kappa(\mathbf{X}'\mathbf{X}) + \kappa^2\mathbf{I}) - (\mathbf{X}'\mathbf{X})] - \kappa^2\boldsymbol{\beta}\boldsymbol{\beta}')(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1} \\
= \;& (\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}(\sigma^2[2\kappa\mathbf{I} + \kappa^2(\mathbf{X}'\mathbf{X})^{-1}] - \kappa^2\boldsymbol{\beta}\boldsymbol{\beta}')(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}
\end{aligned}$$

So, we finally get

$$\begin{aligned}
& MMSE(\hat{\boldsymbol{\beta}}_{OLS}) - MMSE(\hat{\boldsymbol{\beta}}_\kappa) \\
= \;& \kappa(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}(\sigma^2[2\mathbf{I} + \kappa(\mathbf{X}'\mathbf{X})^{-1}] - \kappa\boldsymbol{\beta}\boldsymbol{\beta}')(\mathbf{X}'\mathbf{X}+\kappa\mathbf{I})^{-1}
\end{aligned}$$

for $\kappa > 0$, this is positive definite.

$$\Leftrightarrow \quad 2\mathbf{I} + \kappa[(\mathbf{X}'\mathbf{X})^{-1} - \frac{\boldsymbol{\beta}\boldsymbol{\beta}'}{\sigma^2}] \quad \text{is a positive definite matrix.}$$

$$\text{If} \quad \lambda_{min}((\mathbf{X}'\mathbf{X})^{-1} - \frac{\boldsymbol{\beta}\boldsymbol{\beta}'}{\sigma^2}) \geq 0 \quad \text{this is true for all } \kappa > 0.$$

$$\text{If } \lambda_{min}((\mathbf{X}'\mathbf{X})^{-1} - \frac{\boldsymbol{\beta}\boldsymbol{\beta}'}{\sigma^2}) < 0 \quad \text{this is true if and only if } 0 < \kappa < -\frac{2}{\lambda_{min}}.$$

This is also true if

$2\sigma^2\mathbf{I} - \kappa\boldsymbol{\beta}\boldsymbol{\beta}'$ is positive definite.

Since the latent roots of $\kappa\boldsymbol{\beta}\boldsymbol{\beta}'$ are zero (with multiplicity p-1) and $\kappa\boldsymbol{\beta}'\boldsymbol{\beta}$, it follows that the roots of $2\sigma^2\mathbf{I} - \kappa\boldsymbol{\beta}\boldsymbol{\beta}'$ are $2\sigma^2$ and $2\sigma^2 - \kappa\boldsymbol{\beta}'\boldsymbol{\beta}$. Thus a sufficient condition is,

$$\kappa < \frac{2\sigma^2}{\boldsymbol{\beta}'\boldsymbol{\beta}}. \tag{2.10}$$

□

Vinod and Ullah (1981) give a different proof for the Theobald's result. Note that this condition is sufficient for the superiority of $\hat{\boldsymbol{\beta}}_\kappa$ but not necessary. A necessary and sufficient condition is given by the following theorem.

Theorem 3. *(Swindel and Chapman, 1973). A necessary and sufficient condition for* $MMSE(\hat{\boldsymbol{\beta}}_{OSL}) - MMSE(\hat{\boldsymbol{\beta}}_\kappa)$ *to be a positive-definite matrix is* $\kappa > 0$ *if* $\eta \geq 0$ *and*

$$0 < \kappa < -\frac{2}{\eta}, \quad \text{if } \eta < 0,$$

where η *is the minimum eigenvalue of* $(\mathbf{X}'\mathbf{X})^{-1} - (\boldsymbol{\beta}\boldsymbol{\beta}'/\sigma^2)$.

2.1.3 The ridge trace

We can remark that the $\hat{\boldsymbol{\beta}}_\kappa$ depends upon the unknown parameter κ which makes it impossible to calculate. Hoerl and Kennard (1970) suggested *the ridge trace* method to acquire the suitable value for the ridge parameter κ providing $\hat{\boldsymbol{\beta}}_\kappa$ with smaller MSE than that of the least squares solution $\hat{\boldsymbol{\beta}}_{OLS}$. The *ridge trace* is a graphical tool that plots the components of the ridge regression coefficient $\hat{\boldsymbol{\beta}}_\kappa$ versus κ. This plot will have one curve per coefficient and it can help to see which coefficients are sensitive to the data. High correlations among regressors imply that the components of $\hat{\boldsymbol{\beta}}_\kappa$ will change rapidly for small values of κ and will gradually stabilize at larger values of κ. A suitable value for κ may be chosen such that all the coefficients are stabilized. Marquardt and Snee (1975) consider the ridge trace as one of the major advantages of the ridge regression. It is clear that this method do not yield a single automatic solution to the estimation problem, but rather, a family of solutions. However, Conniffe and Stone (1973) doubt the lack of improvement of the least squares estimator via any particular choice of κ. Instead of it, they recommend direct examination of eigenvalues. Some other rules have been suggested in the literature for choosing κ (see Vinod and Ullah, 1981).

2.2 Other interpretations of the ridge regression estimator

2.2.1 The ridge regression estimator as a solution of a constrained minimization problem

The ridge estimator can also be seen as a solution of a constrained optimization problem. Hoerl and Kennard (1970) consider the error sum of squares due to any estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$,

$$\begin{aligned} ESS(\tilde{\boldsymbol{\beta}}) &= (\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= ESS(\hat{\boldsymbol{\beta}}_{OLS}) + (\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{OLS})'\mathbf{X}'\mathbf{X}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{OLS}) \end{aligned}$$

which achieves its minimum only when $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$. Relation (2.2) proves that the average of the distance between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}_{OLS}$ increases greatly in the presence of ill-conditioning in $\mathbf{X}'\mathbf{X}$ but without an appreciable increase in the error sum of squares. Hoerl and Kennard (1970) therefore, require finding the estimator $\tilde{\boldsymbol{\beta}}$ of minimum length that belongs to the hyperellipsoid centered at the OLS estimator and defined by the equation $\left(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{OLS}\right)' \mathbf{X}'\mathbf{X} \left(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{OLS}\right) = \Phi =$ constant. Figure 2.1 illustrate the geometry of the ridge regression when $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ is a two-dimensional parameter (Marquart and Snee, 1975). We can remark that $\hat{\boldsymbol{\beta}}_{\kappa}$ is the shortest vector that gives a residual sum of squares as small as the Φ value anywhere on the small ellipse.

In an equivalent way, we may minimize $ESS(\tilde{\boldsymbol{\beta}})$ for a fixed length of $\tilde{\boldsymbol{\beta}}$ say r. This is equivalent to finding the ellipse contour that is as close as possible to the circle centered in zero of ray equal to r. Using the Lagrangian principle (Izenman, 2008), the optimization problem may be presented as

$$\min_{\tilde{\boldsymbol{\beta}}}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}) + \kappa(\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}} - r^2),$$

or equivalently,

$$\min_{\tilde{\boldsymbol{\beta}}:||\tilde{\boldsymbol{\beta}}||^2 \leq r^2}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}), \tag{2.11}$$

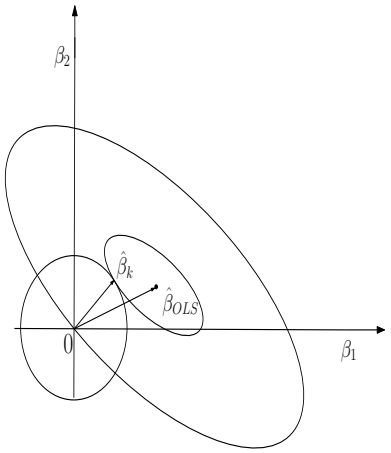


Figure 2.1: Geometry of ridge regression

where $||\cdot||$ is the Euclidean norm. In order to attribute the same influence of the constraint from (2.11), it is advisable to standardize the regressors. With no-standardized variables, one may use some other norm (Kapat and Goel, 2010) or the generalized ridge regression when each diagonal element of $\mathbf{X}'\mathbf{X}$ is modified differently (Hoerl and Kennard, 1970).

2.2.2 Bayesian or Mixed Regression Interpretation of Ridge Coefficients

The Bayesian approach treats the parameter $\boldsymbol{\beta}$ as a random variable with a prior probability density which may be based on some subjective prior information about $\boldsymbol{\beta}$. The goal is to determine the posterior probability density of $\boldsymbol{\beta}$ which is done by combining the prior probability density with the sample information given by the likelihood function. A ridge estimator can be seen also as a Bayes estimator when $\boldsymbol{\beta}$ takes a suitable normal prior distribution with mean $\boldsymbol{\beta}_0$ and variance covariance matrix $\sigma^2_{\beta}\boldsymbol{\Omega}$ (Vinod and Ullah, 1981, Izenman, 2008). Vinod and Ullah (1981) advocate that the Bayesian interpretation of the ridge regression coefficient $\hat{\boldsymbol{\beta}}_{\kappa}$ implies deriving the prior distribution of $\boldsymbol{\beta}$ for which $\hat{\boldsymbol{\beta}}_{\kappa}$ is the posterior mean.

They also state that the Bayesian methods imply that the posterior mean is the optimal estimator when using the MSE as expected loss. We consider the model given in (2.1) with the following supplementary assumptions: the errors $\boldsymbol{\varepsilon}$ are normally distributed with mean zero and variance covariance matrix $\sigma^2\mathbf{I}_p$ with σ^2 a known constant and $\mathbf{I}_p$ is the p dimensional identity matrix. In other words, $\mathbf{y}$ is normally distributed $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_p)$. We suppose that the prior normal distribution of $\boldsymbol{\beta}$ is also normal with known mean $\boldsymbol{\beta}_0$ and known variance $\sigma^2_\beta\boldsymbol{\Omega}$. The posterior density of $\boldsymbol{\beta}$ is therefore normal with mean $\boldsymbol{\beta}^*$ given by

$$\begin{aligned}\boldsymbol{\beta}^* &= (\mathbf{X}'\mathbf{X} + \alpha\boldsymbol{\Omega}^{-1})^{-1}(\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} + \alpha\boldsymbol{\Omega}^{-1}\boldsymbol{\beta}_0) &(2.12)\\ &= \boldsymbol{\beta}_0 + (\mathbf{X}'\mathbf{X} + \alpha\boldsymbol{\Omega}^{-1})^{-1}\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0) &(2.13)\end{aligned}$$

where $\alpha = \sigma^2/\sigma^2_\beta$. The variance-covariance matrix of $\boldsymbol{\beta}$ is given by $\sigma^2\boldsymbol{\Omega}^* = \sigma^2(\mathbf{X}'\mathbf{X} + \alpha^2\boldsymbol{\Omega}^{-1})^{-1}$. Relations (2.12) or (2.13) show that if the prior information is useless, i.e. $\sigma^2_\beta \to \infty$, then $\alpha \to 0$ and $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}_{OLS}$. On the other hand, for $\sigma^2_\beta \to 0$, we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. Vinod and Ullah (1981) remark that the estimator $\boldsymbol{\beta}^*$ given by formula (2.12) may be written as a weighted matrix combination of the OLS or the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{OLS}$ and the prior mean $\boldsymbol{\beta}_0$,

$$\boldsymbol{\beta}^* = \mathbf{H}\hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{I}_p - \mathbf{A})\boldsymbol{\beta}_0, \tag{2.14}$$

where $\mathbf{H}$ is given by

$$\begin{aligned}\mathbf{H} &= \left(\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})^{-1} + \alpha\mathrm{Var}(\boldsymbol{\beta}_0)^{-1}\right)^{-1}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})^{-1} &(2.15)\\ &= \mathbf{I}_p - \mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})\left(\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS}) + \alpha^{-1}\mathrm{Var}(\boldsymbol{\beta}_0)\right)^{-1} &(2.16)\end{aligned}$$

So, the normalized weights of $\hat{\boldsymbol{\beta}}_{OLS}$ and $\boldsymbol{\beta}_0$ are their precision matrix. The same result is obtained if one desires to compute the best estimator from the minimum variance point of view of $\boldsymbol{\beta}$ being a matrix combination of $\hat{\boldsymbol{\beta}}_{OLS}$ and $\boldsymbol{\beta}_0$ namely,

$$\mathbf{H} = \mathrm{argmin}_{\tilde{\mathbf{A}}}\mathrm{Var}\left(\tilde{\mathbf{H}}\hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{I}_p - \tilde{\mathbf{H}})\boldsymbol{\beta}_0\right).$$

One can remark from (2.12), that for $\alpha\boldsymbol{\Omega}^{-1} = k\mathbf{I}_p$ and $\boldsymbol{\beta}_0 = 0$, we get the ordinary

ridge estimator $\hat{\boldsymbol{\beta}}_\kappa$ given by (2.3). As Vinod and Ullah (1981) remarked, some Bayesians feel that this prior is unrealistic and a non null prior mean should be used, but in absence of prior knowledge on $\boldsymbol{\beta}_0$, one may shrink towards the zero vector. When a prior knowledge about $\boldsymbol{\beta}_0$ exists, then one shrinks the ridge estimator toward this known prior. Nevertheless, the drawback is that different choices of the prior lead to different ridge estimators.

It is worth mentioning that the Bayes estimator of $\boldsymbol{\beta}$ given by (2.13) corresponds to the estimator of the regression coefficient for the mixed regression model (Vinod and Ullah, 1981),

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}, \\ \boldsymbol{\beta}_i &= \boldsymbol{\beta}_0 + \boldsymbol{\eta}_i, \end{aligned}$$

with $E(\boldsymbol{\eta}) = 0$ and $\text{Var}(\boldsymbol{\eta}) = \sigma_\beta \boldsymbol{\Omega}$. Conditionally on $\boldsymbol{\beta}_0$, the value of $\boldsymbol{\beta}^*$ given by (2.12) is then obtained by minimization with respect to $\boldsymbol{\beta}$ of

$$\frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{1}{\sigma_\beta^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\boldsymbol{\Omega}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

Even if the two approaches lead to the same solution, the goals are different. In the Bayesian model, $\boldsymbol{\beta}$ is a random variable, whereas in the mixed effect models associated to a prior information the randomness of $\boldsymbol{\beta}$ allows to consider models that vary from one unit to another.

2.2.3 Ridge regression for heteroscedastic regression errors

For a linear regression model such that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $E(\boldsymbol{\varepsilon}) = 0$ and $Var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V}$ where $\mathbf{V}$ is the known sample positive definite covariance matrix and describes the pattern of heteroscedasticity.

The assumption of homoscedasticity claims that the regression errors have a constant variance i.e. $Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$. The violation of this assumption means

heteroscedasticity in the data when the regression errors $\hat{\varepsilon}_j$, $j = 1, ..., p$ do not have a common variance, i.e. $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{V}$, where $\boldsymbol{V} = \text{diag}(v_1^2, v_2^2, \ldots, v_n^2)$ that is, the variance changes with the change of variable in model (Gujarati, 2002). Trenkler (1984) discusses the performance of biased estimators in the linear regression model in the violation of homoscedasticity assumption. Let $\mathbf{G}$ be a non-singular matrix such that $\mathbf{G}'\mathbf{G} = \boldsymbol{V}^{-1}$. If we premultiply the above model by $\mathbf{G}$, we get

$$\begin{aligned} \mathbf{G}\mathbf{y} &= \mathbf{G}\mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\varepsilon} \\ \mathbf{y}^* &= \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \end{aligned}$$

where, $\mathbf{y}^* = \mathbf{G}\mathbf{y}$, $\mathbf{X}^* = \mathbf{G}\mathbf{X}$ and $\boldsymbol{\varepsilon}^* = \mathbf{G}\boldsymbol{\varepsilon}$; Now $E(\boldsymbol{\varepsilon}^*) = 0$ and $Var(\boldsymbol{\varepsilon}^*) = \sigma^2\mathbf{I}$. The generalized least square (GLS) estimator $\hat{\boldsymbol{\beta}}_{GLS}$ is obtained by applying OLS on the new transformed model $\mathbf{y}^*$ and get,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{2.17}$$

where $\mathbf{V} = diag(v_j^2)$, $j = 1, \ldots, p$ and $\mathbf{X}'^*\mathbf{X}^* = \mathbf{X}'\boldsymbol{V}^{-1}\mathbf{X}$. If $\mathbf{X}'\boldsymbol{V}^{-1}\mathbf{X}$ is ill-conditioned, then ridge estimator may be one of the solution to the ill-conditioned $\mathbf{X}'\boldsymbol{V}^{-1}\mathbf{X}$ as,

$$\begin{aligned} \hat{\boldsymbol{\beta}}^*_{\boldsymbol{\kappa}} &= (\mathbf{X}'^*\mathbf{X}^* + \kappa\mathbf{I})^{-1}\mathbf{X}'^*\mathbf{y}^* \\ \hat{\boldsymbol{\beta}}^*_{\boldsymbol{\kappa}} &= (\mathbf{X}'\boldsymbol{V}^{-1}\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \end{aligned}$$

and the variance expression for $\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}$ estimator is given by,

$$Var\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}\right) = \boldsymbol{\sigma}^2(\mathbf{X}'\boldsymbol{V}^{-1}\mathbf{X})^{-1}$$

similarly for

$$Var(\hat{\boldsymbol{\beta}}_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V\mathbf{X}(\mathbf{X}\mathbf{X})^{-1}$$

Until here we discussed the case where the $\boldsymbol{V}$ is known. But if $\mathbf{V}$ is unknown then $\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}$ and $Var(\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}})$ are not feasible. A way for the estimation of $\boldsymbol{V}$ is given by Vinod and Ullah (1981). For this purpose let

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{D}\boldsymbol{\varepsilon}$$

be the OLS residual estimator vector where $\mathbf{D} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{D}^2$ be the $T - p$ matrix. Then $E(\hat{\boldsymbol{\epsilon}}) = 0$, $Var(\hat{\boldsymbol{\epsilon}}) = \mathbf{D}\boldsymbol{V}\mathbf{D}$. Let us write,

$$E(\bar{\hat{\boldsymbol{\varepsilon}}}) = \bar{\mathbf{D}}\omega,$$

where $\bar{\hat{\boldsymbol{\varepsilon}}} = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$, $\bar{\mathbf{D}} = \mathbf{D}'\mathbf{D}$ and $\omega = [\sigma_1, \sigma_2, \ldots, \sigma_p]$ be the vector estimating $Diag(\mathbf{V})$. We can also write, $\gamma = \bar{\hat{\boldsymbol{\varepsilon}}} - E(\bar{\hat{\boldsymbol{\varepsilon}}})$ which implies,

$$\bar{\hat{\boldsymbol{\varepsilon}}} = \bar{\mathbf{D}}\omega + \gamma$$

which is regression of squared errors on the matrix $\bar{\mathbf{D}}$ and

$$\begin{aligned} \hat{\omega}_{OLS} &= (\bar{\mathbf{D}}'\bar{\mathbf{D}})^{-1}\bar{\hat{\boldsymbol{\varepsilon}}} \\ \hat{\omega}_{OLS} &= \bar{\mathbf{D}}^{-1}\bar{\hat{\boldsymbol{\varepsilon}}} \end{aligned}$$

For a singular $\mathbf{D}$, the ridge estimator of ω can be written as,

$$\hat{\omega}_\kappa = (\bar{\mathbf{D}}'\bar{\mathbf{D}} + \kappa\mathbf{I})^{-1}\bar{\hat{\boldsymbol{\varepsilon}}}$$

Finally, we have the general estimate of $\boldsymbol{\beta}$ is constructed by replacing the estimated $\hat{\mathbf{V}}$ of $\mathbf{V}$,

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \\ \text{and} & \\ Var(\tilde{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}\mathbf{X}(\mathbf{X}\mathbf{X})^{-1}, \end{aligned}$$

where we get $\tilde{\boldsymbol{\beta}}_{OLS}$ by using $\hat{\omega}_{OLS}$ and $\tilde{\boldsymbol{\beta}}_\kappa$ by using $\hat{\omega}_\kappa$.

In what concerns the standardization of the regressors, the problem is more delicate and it is not always very obvious when one should standardize the $\mathbf{X}$-variables. The standardization is not necessary for most theoretical results (Vinod and Ullah, 1981). However, it is advisable to standardize data before computing the ridge estimator specially when there are large variations between regressors and they are measured in different scales. An additional advantage of the standardization is that it makes the numerical magnitude of the components of $\boldsymbol{\beta}$ comparable with each other. As Kapat and Goel (2010) remarked, different solutions for the ridge

estimator $\hat{\boldsymbol{\beta}}_k$ may be obtained depending on the nature of the regressors, standardized or not, and on the constrained norm. Thus, it is important to distinguish between the solutions of these problems in order to avoid confusion.

2.3 Use of the ridge principle in surveys

In this section, we address the use of the ridge principle in a survey sampling setting. Under this setting, the main goal is not to make inference on the vector $\mathbf{y}$, but on either a function of $\mathbf{y}$ or the regression coefficient $\boldsymbol{\beta}$. We consider the simplest case of estimating the finite population total

$$t_y = \sum_{k \in U} y_k$$

of the variable of interest $\mathcal{Y}$ of values y_k. Here, U denotes a finite population containing N elements,

$$U = \{a_1, \ldots, a_k, \ldots, a_N\} = \{1, \ldots, k, \ldots, N\}$$

with the assumption that a population unit is uniquely identifiable by its label k. Furthermore, a sample s of size n is selected from U according to a sampling design and the vector $\mathbf{y}$ is known only on the sample individuals. Usually, the finite population total t_y is estimated by a weighted estimator $\hat{t}_w$,

$$\hat{t}_w = \sum_s w_k y_k \tag{2.18}$$

where the weights w_k are derived usually using auxiliary information by means of a superpopulation model (model-based or model-assisted approach) or by calibration. Usually, with multipurpose surveys, weights should not depend on the study variable in order to estimate means or totals of a very large number of variables. They should also be positive and depend only on the auxiliary information. The weights necessarily should produce internally consistent estimators and if they are suitably chosen, these weights will produce estimators with smaller variance than the estimators without using the weights.

The idea of ridge estimation was used for the first time in a survey sampling framework in order to eliminate negative or extremely large weights obtained when a too restrictive condition of unbiasedness was imposed. Latter situations may cause inefficient results rather than improving the estimators. So, weights are crucial in survey sampling theory. From (2.18), the weights vector $\mathbf{w}_s = (w_k)_{k\in s}$ is the unknown parameter to be found. The role of $\boldsymbol{\beta}$ is taken now by $\mathbf{w}_s$. In sections 2.3.1 and 2.3.2 we give in detail the derivation of ridge weights in survey sampling as solutions of constrained optimization problems as described in section 2.2. The same estimators may be obtained by using a superpopulation linear model depending on a parameter estimated using ridge regression and the class of model-based or model-assisted estimators for the finite population totals. This way of computing ridge estimators in survey sampling is the direct application of ridge principle from the classical regression described in section 2.1.2 and we present it below. When we attribute a prior on previous estimations, we may use the Bayesian interpretation to construct ridge regression type estimators. Deville (1999) considered it as a calibration on an uncertain source. We describe the method in section 2.3.4.

Suppose that the relationship between the variable of interest $\mathcal{Y}$ and the auxiliary variables $\mathcal{X}_1, \ldots, \mathcal{X}_p$ is given by a superpopulation model denoted by ξ in the survey literature:

$$\xi: \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.19}$$

The explicative variables are not standardized now. In order to distinguish the population from the sample, let $\mathbf{y} = (y_1, \ldots, y_N)'$ be a $N \times 1$ vector of and let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ be the $N \times p$ matrix with $\mathbf{x}_k' = (X_{k1}, \ldots, X_{kp})$ as rows. The errors ε_k, for all $k \in U$ are independent one of each other, of mean zero and variance $\mathrm{Var}(\varepsilon_k) = \sigma^2 v_k^2$. Let $\mathrm{Var}_\xi(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$ with $\mathbf{V} = \mathrm{diag}(v_k^2)_{k\in U}$ and v_k are positive known constants.

Some further notations are needed. Let $\mathbf{X}_s = (\mathbf{x}_k')_{k\in s}$, respectively $\mathbf{y}_s = (y_k)_{k\in s}$, be the restriction of $\mathbf{X}$, respectively of $\mathbf{y}$, on the sample s. Let also $\mathrm{Var}_\xi(\boldsymbol{\varepsilon}_s) = \sigma^2\mathbf{V}_s$

be the variance of $\boldsymbol{\varepsilon}_s$, the restriction of $\boldsymbol{\varepsilon}$ on the sample s, and $\mathrm{Var}_\xi(\boldsymbol{\varepsilon}_{\bar{s}}) = \sigma^2\mathbf{V}_{\bar{s}}$ be the variance of $\boldsymbol{\varepsilon}_{\bar{s}}$, the restriction of $\boldsymbol{\varepsilon}$ on $\bar{s} = U - s$. The population variance $\mathbf{V}$ may be written as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_s & \mathbf{0}_{n\times(N-n)} \\ \mathbf{0}_{(N-n)\times n} & \mathbf{V}_{\bar{s}} \end{pmatrix}.$$

Without auxiliary information, t_y is estimated by the Horvitz and Thompson (1952) (see also Narain, 1951) estimator given by

$$\hat{t}_{y,d} = \sum_s d_k y_k = \sum_s \frac{y_k}{\pi_k}, \tag{2.20}$$

where $\pi_k = P(k \in s)$ is the first order inclusion probability of the individual $k \in U$. The auxiliary information given by $\mathbf{X}_1, \ldots, \mathbf{X}_p$ may be used to improve the estimation of $\hat{t}_{y,d}$.

Using the model ξ, one estimate the regression parameter $\boldsymbol{\beta}$ and after, plugs-in a *model based estimator*, abbreviated as MB below,

$$\hat{t}_{MB} = \sum_s y_k + \sum_{U-s} \mathbf{x}_k'\boldsymbol{\beta}, \tag{2.21}$$

or in a *generalized difference estimator*, abbreviated as DIFF below,

$$\hat{t}_{DIFF} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{x}_k'}{\pi_k} - \sum_U \mathbf{x}_k'\right)\boldsymbol{\beta}. \tag{2.22}$$

This means that $\hat{t}_{MB}$ and $\hat{t}_{DIFF}$ rely on the estimation of the regression coefficient $\boldsymbol{\beta}$: best linear unbiased estimator of $\boldsymbol{\beta}$ for the MB estimator (Royall, 1976) and the best design-based estimator of $\boldsymbol{\beta}$ for the DIFF estimator (Särndal, 1980).

In a model-based setting and using the generalized least squares (GLS) estimation under the model ξ, the estimator of the regression coefficient $\boldsymbol{\beta}$ is obtained as solution of the optimization problem

$$\textbf{(P1)}: \qquad \hat{\boldsymbol{\beta}}_{GLS} = \mathrm{argmin}_{\boldsymbol{\beta}}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}), \tag{2.23}$$

yielding the estimator $\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s$ assuming that $(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}$ exists. Plugging $\hat{\boldsymbol{\beta}}_{GLS}$ in (2.21), yields the best linear unbiased estimator (BLUE)

of t_y from the ξ-variance point of view (Royall, 1976),

$$\hat{t}_{BLUE} = \sum_s y_k + \sum_{U-s} \mathbf{x}_k' \hat{\boldsymbol{\beta}}_{GLS}. \tag{2.24}$$

If the matrix $\mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{X}_s$ has eigenvalues close to zero, then it is advisable to perturb its diagonal before inverting it. We obtain the ridge estimator of $\boldsymbol{\beta}$ as follows

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{MBR} &= \text{argmin}_{\boldsymbol{\beta}} (\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta})' \mathbf{V}_s^{-1} (\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta}) + \boldsymbol{\beta}' \mathbf{C}^{-1} \boldsymbol{\beta} \\ &= \left(\mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{X}_s + \mathbf{C}^{-1} \right)^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{y}_s, \end{aligned}$$

where $\mathbf{C}$ is a $p \times p$ diagonal matrix with positive quantities on the diagonal. The *ridge MB* estimator is obtained by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}_{MBR}$ in (2.21),

$$\hat{t}_{MBR} = \sum_s y_k + \left(\sum_{U-s} \mathbf{x}_k' \right) \hat{\boldsymbol{\beta}}_{MBR}. \tag{2.25}$$

A similar reasoning may be used in a design-based approach. The design-based estimator $\hat{\boldsymbol{\beta}}_\pi$ of the regression coefficient $\boldsymbol{\beta}$ is the solution of the following optimization problem (Särndal, 1980),

$$\textbf{(P2)}: \qquad \hat{\boldsymbol{\beta}}_\pi = \text{argmin}_{\boldsymbol{\beta}} (\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta})' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} (\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta})$$

where $\boldsymbol{\Pi}_s = \text{diag}(\pi_k)_{k \in s}$. This optimization problem yields the following estimator for $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}_\pi = \left(\mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s \right)^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s \tag{2.26}$$

assuming that $\mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s$ is invertible. The total t_y is then estimated by the well known GREG estimator (also known as model-assisted (MA) estimator) obtained by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}_\pi$ in (1.3),

$$\hat{t}_{GREG} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{x}_k'}{\pi_k} - \sum_U \mathbf{x}_k' \right) \hat{\boldsymbol{\beta}}_\pi. \tag{2.27}$$

The ridge estimator of $\boldsymbol{\beta}$ is obtained as solution of the following penalized optimization problem,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\pi,R} &= \text{argmin}_{\boldsymbol{\beta}} (\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta})' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} (\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta}) + \boldsymbol{\beta}' \tilde{\mathbf{C}}^{-1} \boldsymbol{\beta} \\ &= \left(\mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s + \tilde{\mathbf{C}}^{-1} \right)^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s \end{aligned} \tag{2.28}$$

for some positive diagonal matrix $\tilde{\mathbf{C}}$. Plugging-in (1.3), we obtain *the ridge GREG* estimator,

$$\hat{t}_{GREG,R} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{x}'_k}{\pi_k} - \sum_U \mathbf{x}'_k \right) \hat{\boldsymbol{\beta}}_{\pi,R}. \tag{2.29}$$

The ridge estimators of $\boldsymbol{\beta}_{MBR}$ and $\boldsymbol{\beta}_{\pi,R}$ are ξ-biased and taking into account the discussion given in the previous section, we may affirm that they are more stable in presence of multicollinearity.

2.3.1 Ridge regression under the model-based approach

Bardsley and Chambers (1984) explored the relationship between the unbalanced samples and multicollinearity. A balanced sample is a sample for which the following relation is satisfied

$$\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k.$$

On the opposite situation, we have an unbalanced sample. As Bardsley and Chambers (1984) stated, in multipurpose sample surveys for which a large number of finite population totals or means are to be estimated, it is very difficult or even impossible to have a fully specified model underlying each study variable. In such situations, balanced sampling may protect from model misspecification (Royall and Herson, 1973).

In the model-based setting for unbalanced sampling, exclusion of variables may increase the bias and inclusion of too many variables may result in a overspecified model and the estimates will be unstable and inefficient even if they are unbiased. Also these variables can linearly be related with each other, and hence can cause multicollinearity. The strategy suggested by Bardsley and Chambers (1984) is to consider as many variables as they exist but to relax the balancing condition which is in fact the unbiasedness condition of the estimator under the model. This is equivalent to deriving a biased estimator but with a smaller prediction error and this is why, it leads naturally to a ridge type estimator.

Bardsley and Chambers (1984) suggest finding the weights $\mathbf{w}_s = (w_k)_{k\in s}$ such that the prediction error $\hat{t}_w - t_y = \sum_s w_k y_k - \sum_U y_k$ has minimum ξ-mean squared error among the class of bounded biased estimators,

$$\textbf{(P3)}: \quad \mathbf{w}_{MB,R} = \text{argmin}_{\mathbf{w}_s}(\mathbf{w}_s - \mathbf{1}_s)'\mathbf{V}_s(\mathbf{w}_s - \mathbf{1}_s) + \mathbf{B}'\mathbf{C}\mathbf{B}, \tag{2.30}$$

where $\mathbf{B} = \sum_s w_k \mathbf{x}_k - \sum_U \mathbf{x}_k$, $\mathbf{C}$ is some diagonal cost matrix and $\mathbf{1}_s$ is the n-dimensional vector of ones. The optimization problem **(P3)** results from the fact that the ξ-variance of $\hat{t}_w - t_y$ is equal to $\sigma^2(\mathbf{w}_s - \mathbf{1}_s)'\mathbf{V}_s(\mathbf{w}_s - \mathbf{1}_s)$ plus a term not depending on $\mathbf{w}_s$ and respectively, the ξ-bias is equal to $\mathbf{B}'\boldsymbol{\beta}$. The equality $\mathbf{B} = 0$ means that the estimator $\hat{t}_w$ is ξ-unbiased or that the design is *exactly balanced*. In the latter case, the solution of **(P3)** yields the BLUE estimator given by (2.24). The weights obtained by solving **(P3)** may be seen as the weights that explain the best the vector $\mathbf{1}_s$ according to a specific metric and such that the weighted estimator is not very far away from the true total. The metric employed here uses the sample variance $\mathbf{V}_s$ as we are in the case of a model-based approach.
The minimization problem from above can also be written as a constrained optimization problem

$$\textbf{(P3')}: \quad \mathbf{w}_{MB,R} = \text{argmin}_{\mathbf{w}_s, ||\mathbf{B}||_C^2 \leq r^2}(\mathbf{w}_s - \mathbf{1}_s)'\mathbf{V}_s(\mathbf{w}_s - \mathbf{1}_s)$$

for the norm $||\mathbf{B}||_C^2 = \mathbf{B}'\mathbf{C}\mathbf{B}$ which means that we penalize large values of $\mathbf{B}$. Solving **(P3)** or **(P3')**, we obtain (see proof of Proposition 2)

$$\mathbf{w}_{MB,R} = \mathbf{1}_s - \mathbf{V}_s^{-1}\mathbf{X}_s\left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \mathbf{C}^{-1}\right)^{-1}(\mathbf{1}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})' \tag{2.31}$$

leading to the *ridge MB* estimator $\hat{t}_{MBR}$ given by (2.55),

$$\hat{t}_{MBR} = \mathbf{w}_{MB,R}'\mathbf{y}_s = \sum_s y_k + \left(\sum_{U-s} \mathbf{x}_k'\right)\hat{\boldsymbol{\beta}}_{MBR} \tag{2.32}$$

with $\hat{\boldsymbol{\beta}}_{MBR} = \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \mathbf{C}^{-1}\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s$. We have mentioned earlier that the vector of weights $\mathbf{w}_s$ performs the similar role to the regression coefficient $\boldsymbol{\beta}$. We have seen in section 2.1.2 that adding a constant to the diagonal of $\hat{\boldsymbol{\beta}}_{OLS}$, reduced

its length. The same result is true for the weight vector $\mathbf{w}_{MB,R}$ (Bardsley and Chambers, 1984). Consider for that, the particular case $\mathbf{V}_s = \mathbf{I}_n$ and $\mathbf{C}^{-1} = \kappa \mathbf{I}_p$,

$$\begin{aligned} \mathbf{w}'_{MB,R}\mathbf{w}_{MB,R} &\simeq \mathbf{1}'_U \mathbf{X}(\mathbf{X}'_s\mathbf{X}_s + \kappa\mathbf{I}_p)^{-1}\mathbf{X}'_s\mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s + \kappa\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{1}_U \\ &= \sum_{i=1}^{p} \eta_i^2 \frac{\lambda_i}{(\lambda_i + \kappa)^2}, \end{aligned}$$

where λ_i, $i = 1, \ldots, p$ are the eigenvalues of $\mathbf{X}'_s\mathbf{X}_s$, $\boldsymbol{\eta} = (\eta_i)_{i=1}^p = \mathbf{A}\mathbf{X}'\mathbf{1}_U$ and $\mathbf{A}$ is the matrix of eigenvectors associated to the eigenvalues of $\mathbf{X}'_s\mathbf{X}_s$. Let $\mathbf{w}_{MB}$ be the weights giving the BLUE estimator $\hat{t}_{BLUE}$ exhibited in relation (2.24). More exactly, $\mathbf{w}_{MB} = \mathbf{1}_s - \mathbf{V}_s^{-1}\mathbf{X}_s \left(\mathbf{X}'_s\mathbf{V}_s^{-1}\mathbf{X}_s\right)^{-1} \mathbf{X}'_s\mathbf{V}_s^{-1}\mathbf{y}_s$ and they are obtained from $\mathbf{w}_{MB,R}$ for $\kappa = 0$. Following the same arguments as above, we obtain that

$$\mathbf{w}'_{MB}\mathbf{w}_{MB} \simeq \mathbf{1}'_U \mathbf{X}(\mathbf{X}'_s\mathbf{X}_s)^{-1}\mathbf{X}'\mathbf{1}_U = \sum_{i=1}^{p} \eta_i^2 \frac{1}{\lambda_i}.$$

Since for any $\kappa > 0$, we always have $\frac{1}{\lambda_i} > \frac{\lambda_i}{(\lambda_i+\kappa)^2}$, we get that $\mathbf{w}'_{MB,R}\mathbf{w}_{MB,R} < \mathbf{w}'_{MB}\mathbf{w}_{MB}$. This proves that the scatter of ridge weights is smaller and more stable under perturbation of $\mathbf{X}_s$ than that of BLUE weights. This is in concordance with the ridge principle. If $\lambda_{min} = \min_{i=1}^p \lambda_i$ is close to zero results in a large conditioning number $K = \sqrt{\lambda_{max}/\lambda_{min}}$. This fact may entail negative or extremely large calibration weights.

It is worth mentioning two extreme values of $\hat{t}_{MBR}$. As $\mathbf{C} \to \infty$ (i.e. infinite cost associated with the bias $\mathbf{B}$), we obtain that the ridge weights become the BLUE weights, $\mathbf{w}_{MB,R} = \mathbf{w}_{MB}$ and $\hat{t}_{MBR}$ is the minimum variance unbiased linear estimator $\hat{t}_{BLUE}$ (Royall, 1970). This means that the constraint $\mathbf{B} = 0$ is exactly satisfied. On the opposite case, as $\mathbf{C} \to 0$, we obtain that $\mathbf{w}_{MB,R} = \mathbf{1}_s$ and $\hat{t}_{MBR} = \sum_s y_k$ which is equivalent to removing the constraint from the optimization problem.

The derivation of the model-based ridge estimator depends on the cost matrix $\mathbf{C}$. Considering that $\mathbf{C} = \kappa^{-1}\mathbf{C}^*$, Bardsley and Chambers (1984) and Chambers (1996) use the ridge trace to determine the appropriate κ. $\mathbf{C}^*$ is a fixed cost matrix providing a correct relative weighting of the components of the relative bias vector

$(\text{diag}(\mathbf{X}'\mathbf{1}_U))^{-1}\mathbf{B}$. This transformation is needed because of the large differences in scale between the predictors in $\mathbf{X}$ and it is a kind of standardization of variables.

2.3.2 Ridge under the calibration approach or penalized calibration

Without assuming a superpopulation model, one can use the *calibration method* (Deville and Särndal, 1992) which consists in deriving a weighted estimator

$$\hat{t}_{yw} = \sum_s w_k y_k,$$

with weights minimizing a pseudo-distance, subject to calibration constraints (i.e. all the auxiliary variable totals are exactly estimated). Usually a chi-square distance is used, $\sum_s (w_k - d_k)^2/d_k q_k$, yielding the calibration weights $\mathbf{w}_s^c = (w_k^c)_{k\in s}$

$$\begin{aligned} \textbf{(P4)}: \quad \mathbf{w}_s^c &= \text{argmin}_{\mathbf{w}_s}(\mathbf{w}_s - \mathbf{d}_s)'\widetilde{\mathbf{\Pi}}_s(\mathbf{w}_s - \mathbf{d}_s) \\ &\text{subject to} \quad (\mathbf{w}_s)'\mathbf{X}_s = \mathbf{1}_U'\mathbf{X}, \end{aligned}$$

where $\widetilde{\mathbf{\Pi}}_s = \text{diag}(q_k^{-1}d_k^{-1})_{k\in s}$ and q_k are positive constants. Most of the times, we consider $q_k = 1$ for all k. The calibration weights thus get the following shape,

$$\mathbf{w}_s^c = \mathbf{d}_s - \widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s)^{-1}(\mathbf{d}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})'.$$

For $q_k = 1/(\sigma^2 v_k^2)$, the calibrated estimator $\hat{t}_{yw} = (\mathbf{w}_s^c)'\mathbf{y}_s$ is equal to the GREG estimator given by (1.6). Moreover, note that in this case we have $\widetilde{\mathbf{\Pi}}_s = \mathbf{V}_s\mathbf{\Pi}_s$, which means that the optimization problem **(P2)** uses the inverse of the weight matrix employed in the objective function from **(P4)**. For a more general distance function, Deville and Särndal (1992) show that under certain conditions the calibrated estimator is asymptotically equivalent to the model-assisted or GREG estimator $\hat{t}_{GREG}$. This equivalence is in the sense that $N^{-1}(\hat{t}_{yw} - \hat{t}_{GREG}) = O_p(n^{-1})$. This fact will consequently lead to the asymptotic equivalence of the variances of both estimators.

From a geometrical point of view, we search the weights w_k which explain the best the Horvitz-Thompson weights $d_k = 1/\pi_k$ and that lie in the constraint space

given by the kernel of the matrix $\mathbf{X}_s$. The constraint space is of dimension $n-p$, so increasing the number of auxiliary variables will decrease the number of degrees of freedom for w_k (Guggemos and Tillé, 2010). A similar reasoning given by Silva and Skinner (1997) proved that increasing the number of calibration variables after a certain number may increase the variance up to a harmful level. Guggemos and Tillé (2010) called it *over-calibration* and suggested not calibrating on those variables which are less correlated with the variables of interest.

Another issue with the calibration weights is the fact that they may not satisfy range restrictions (i.e. pre-specified lower and upper bounds) especially when the number of calibration or benchmark constraints is large. Satisfying such condition is desirable especially for avoiding the inflation of the sampling error of estimates in small to moderate domains (Beaumont and Bocci, 2008). Moreover, as Deville and Särndal (1992) stated, negative weights may occur when the chi-squared distance is employed. For the other distances used in their paper, the positiveness of weights is guaranteed but unrealistic or extreme weights may also occur. To cope with this issue, several modifications have been suggested in the literature. However, all these methods are iterative and may not yield a solution even if the range restriction is mild (Rao and Singh, 1997 and Beaumont and Bocci, 2008).

So, how to avoid negative or extremely large weights? Chambers (1996) and Rao and Singh (1997) answer this question by suggesting to relax the calibration constraints. Suppose we have non-negative constants C_j, $j=1,...,p$, representing the cost associated to the j-th calibration equation not to be satisfied and let $\mathbf{C}=\text{diag}(C_j)_{j=1}^{p}$. Relaxing the calibration constraints may be obtained by using a quadratic constraint function instead of the linear constraint function used in **(P4)**. With the chi-square distance, we want to find weights that verify

$$\begin{aligned}\textbf{(P5)}: \quad \mathbf{w}_{R,s}^{c} &= \text{argmin}_{\mathbf{w}_s}(\mathbf{w}_s-\mathbf{d}_s)'\widetilde{\mathbf{\Pi}}_s(\mathbf{w}_s-\mathbf{d}_s) \\ &\quad +\frac{1}{\lambda}(\mathbf{w}_s'\mathbf{X}_s-\mathbf{1}_U'\mathbf{X})\mathbf{C}(\mathbf{w}_s'\mathbf{X}_s-\mathbf{1}_U'\mathbf{X})'. \qquad (2.33)\end{aligned}$$

Rao and Singh (1997) consider the objective function without the constant λ. Writing the problem **(P5)** as a constrained optimization problem, puts into ev-

idence that we lessen the calibration equation corresponding to those variables which are somehow unable to satisfy the calibration constraints but not too much since we penalize the large values of $\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X}$. The absolute value of constraints $\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X}$ may be controlled by a tolerance matrix $\mathbf{\Delta} = diag(\delta_i)_{i=1}^p$ with $\delta_i \geq 0$ as described by Rao and Singh (1997),

$$|\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X}| \leq (\mathbf{1}_U'\mathbf{X})\mathbf{\Delta}.$$

The link between the tolerance matrix $\mathbf{\Delta}$ and the cost matrix $\mathbf{C}$ (Rao and Singh, 1997 and Beaumont and Bocci, 2008) may be used to find $\mathbf{C}$ that meets fixed tolerances δ_i for all $i = 1, \ldots, p$. In this way we eliminate the possibility of having very large or negative weights. Simply, we can say that the ridge estimator performs as a variable selection tool.

The weights verifying the optimization problem **(P5)** are given by

$$\mathbf{w}_{R,s}^c = \mathbf{d}_s - \widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s + \lambda\mathbf{C}^{-1})^{-1}(\mathbf{X}_s'\mathbf{d}_s - \mathbf{X}'\mathbf{1}_U), \tag{2.34}$$

which yield the ridge calibration estimator or the penalized calibration of the population total t_y,

$$\begin{aligned}\hat{\mathbf{t}}_{y,Rw} = (\mathbf{w}_{R,s}^c)'\mathbf{y}_s &= \mathbf{d}_s'\mathbf{y}_s - (\mathbf{X}_s'\mathbf{d}_s - \mathbf{X}'\mathbf{1}_U)'\hat{\boldsymbol{\beta}}_\lambda \\ &= \hat{t}_{y,d} - \left(\hat{\mathbf{t}}_{x,d} - \mathbf{t}_x\right)'\hat{\boldsymbol{\beta}}_\lambda, \end{aligned} \tag{2.35}$$

where $\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}_s'\widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s + \lambda\mathbf{C}^{-1})^{-1}\mathbf{X}_s'\widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s$ and $\hat{\mathbf{t}}_{x,d}$ is the Horvitz-Thompson estimator for the total $\mathbf{t}_x$. This approach is equivalent to construct a GREG estimator of population total with the regression coefficient estimated by a ridge estimator (Hoerl and Kennard, 1970). More precisely, $\hat{\boldsymbol{\beta}}_\lambda$ is in fact $\hat{\boldsymbol{\beta}}_{\pi,R}$ from (2.28) for $\lambda\mathbf{C}^{-1} = \tilde{\mathbf{C}}^{-1}$ and $\widetilde{\mathbf{\Pi}}_s^{-1} = \mathbf{V}_s^{-1}\mathbf{\Pi}_s^{-1}$.

The ridge estimator given by (2.35) can be written as a linear combination of the Horvitz-Thompson estimator and the GREG estimator (Rao and Singh, 1997) as follows,

$$\hat{t}_{y,Rw} = (1-\alpha)\hat{t}_{y,d} + \alpha\hat{t}_{GREG},$$

where $\hat{t}_{GREG} = \hat{t}_{y,d} - \left(\hat{\mathbf{t}}_{x,d} - \mathbf{t}_x\right)' \hat{\boldsymbol{\beta}}_\pi$ is the GREG estimator given by (1.6) and α is given by,

$$\begin{aligned}\alpha &= \mathbf{y}_s' \widetilde{\boldsymbol{\Pi}}_s^{-1} \mathbf{X}_s \left(\mathbf{X}'_s \widetilde{\boldsymbol{\Pi}}_s^{-1} \mathbf{X}_s + \lambda \mathbf{C}^{-1}\right)^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_{x,d}) \\ &\quad \left[\mathbf{y}_s' \widetilde{\boldsymbol{\Pi}}_s^{-1} \mathbf{X}_s \left(\mathbf{X}'_s \widetilde{\boldsymbol{\Pi}}_s^{-1} \mathbf{X}_s\right)^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_{x,d})\right]^{-1}.\end{aligned}$$

As for the model-based approach, the Horvitz-Thompson as well as the GREG estimator are two limit values of $\hat{t}_{y,Rw}$. More exactly, consider relation (2.35) for a fixed cost matrix $\mathbf{C}$ and let λ vary from 0 to ∞. The ridge calibration estimator is a continuous function of λ. For $\lambda = 0$, then $\alpha = 1$ and an infinite cost is attributed to all constraints meaning that they are all exactly satisfied. It implies that $\hat{t}_{y,Rw}$ is the GREG estimator which is ξ-unbiased for the population total t_y. Ridge weights with strictly positive biasing parameter λ means that the weights do not satisfy exactly the calibration equations. In this case, the estimator $\hat{t}_{y,Rw}$ is ξ-biased but the weights $\mathbf{w}_{R,s}^c$ are more stable (Chambers, 1996) and implied a reduction in *MSE* (Bardsley and Chambers, 1984). Values of λ producing weights larger or equal to 1 are accepted by Chambers (1996).

As $\lambda \to \infty$, $\alpha \to 0$ and the ridge calibrated estimator $\hat{t}_{y,Rw}$ goes to the Horvitz-Thompson estimator. In this case, we do not use any of the auxiliary variables for the estimation of the finite population total of the variable of interest.

It is of interest to see how $\hat{t}_{y,Rw}$ changes when a specific cost C_j varies from 0 to ∞. The zero cost $C_j = 0$ means that the constraint corresponding to the total t_{X_j} is discarded and the large or infinite cost $C_j = \infty$, that the corresponding calibration constraint is exactly satisfied. In the latter situation, the weights are computed using (2.34) with the cost matrix $\mathbf{C}^{-1}$ having 0 on the j-th diagonal element. Using matrix algebra, one can show the following result which is the equivalent of the constrained optimization problem (2.11) in a survey setting.

Proposition 1. *The weights $\boldsymbol{w}_{R,s}^c$ satisfying the optimization problem **(P5)** satisfy*

also the following optimization problem,

$$
\begin{aligned}
\textbf{\textit{(P6)}}: \quad \boldsymbol{w}_{R,s}^{c} &= argmin_{\mathbf{w}_s}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})\mathbf{C}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})' \\
&\quad +\lambda(\mathbf{w}_s - \mathbf{d}_s)'\widetilde{\mathbf{\Pi}}_s(\mathbf{w}_s - \mathbf{d}_s) \\
&= argmin_{\mathbf{w}_s, ||\mathbf{w}_s - \mathbf{d}_s||^2_{\widetilde{\mathbf{\Pi}}_s} \leq r^2}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})\mathbf{C}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})'.
\end{aligned}
$$

This results means that we want to find weights minimizing the distance between the weighted estimator $\mathbf{w}_s'\mathbf{X}_s$ and the total $\mathbf{1}_U'\mathbf{X}$ while lying at a given distance from the sampling weights. Figure 2.2 gives the geometric representation of the penalized weights for the two-dimensional case $\mathbf{w}_s = (w_1, w_2)'$. The interpretations are similar to those given by Hoerl and Kennard, (1970) in the case of classical regression (see section 2.2). More exactly, consider for simplicity that the auxiliary variables are centered, namely $\mathbf{1}_U'\mathbf{X} = 0$. Then, the optimization problem **(P6)** reduces to finding the minimum of $\mathbf{w}_s'\mathbf{X}_s\mathbf{C}\mathbf{X}_s'\mathbf{w}_s$ under the constraint that $||\mathbf{w}_s - \mathbf{d}_s||^2_{\widetilde{\mathbf{\Pi}}_s} \leq r^2$. Weights satisfying $\mathbf{w}_s'\mathbf{X}_s\mathbf{C}\mathbf{X}_s'\mathbf{w}_s = \Phi =$ constant lie on an ellipse centered in the origin. For the calibration weights $\mathbf{w}_s^c$, we get the minimum value of Φ, $\Phi_{min} = 0$ but the range restrictions are not necessarily satisfied. For the sampling weights $\mathbf{d}_s$, we get the maximum value of Φ, $\Phi_{max} = \mathbf{d}_s'\mathbf{X}_s\mathbf{C}\mathbf{X}_s'\mathbf{d}_s$. The penalized calibration weights are found in the following way. We start by fixing the constraint contour at r^2, namely $(\mathbf{w}_s - \mathbf{d}_s)'\widetilde{\mathbf{\Pi}}_s(\mathbf{w}_s - \mathbf{d}_s) = r^2$. This means that $\mathbf{w}_s$ lies on the ellipse centered in $\mathbf{d}_s$ (see figure 2.2). Next, we find the ellipse contour centered in the origin that is as close as possible to the ellipse centered in $\mathbf{d}_s$. The penalized calibration weights $\mathbf{w}_{R,s}^c$ is the vector at the first point where the ellipse contour hits the constraint region (see figure 2.2). A value r may be chosen such that all range restrictions as $L \leq w_k/d_k \leq U$ for all $k \in U$ are satisfied.

Knowing Φ_{min} and Φ_{max}, Beaumont and Bocci (2008) suggest the bisection algorithm to find λ that leads to the penalized calibration weights. Nevertheless, this algorithm may be time-consuming.

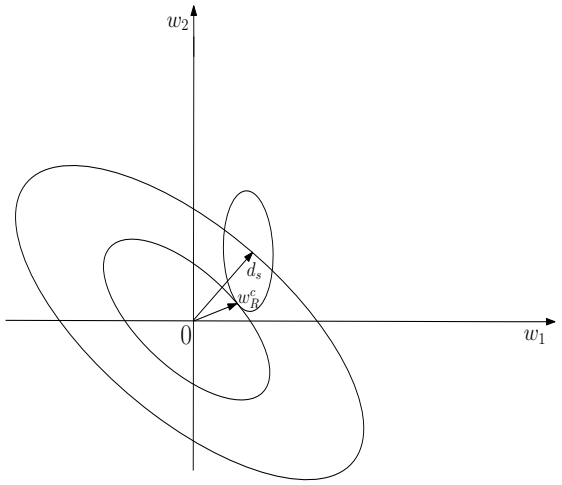


Figure 2.2: Geometry of penalized weights

2.3.3 Partially ridge regression or partially penalized calibration

In a model based approach, Bardsley and Chambers (1984) suggested to divide the p variables in the data matrix $\mathbf{X}$ into two sets of variables $\widetilde{\mathbf{X}}_1$ and $\widetilde{\mathbf{X}}_2$ based on the fact that variables in $\widetilde{\mathbf{X}}_1$ contain much more importance than the variables in $\widetilde{\mathbf{X}}_2$ in the sense that they can contribute more influentially in the estimation process. We may consider that the matrix $\mathbf{X}$ has the following expression after re-ordering the variables $\mathbf{X_1}, \ldots, \mathbf{X}_p$,

$$\mathbf{X} = \left(\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2\right),$$

where $\widetilde{\mathbf{X}}_1 = [\mathbf{X}_1, \ldots, \mathbf{X}_q]$ and $\widetilde{\mathbf{X}}_2 = [\mathbf{X}_{q+1}, \ldots, \mathbf{X}_p]$. The variables contained in $\widetilde{\mathbf{X}}_1$ may be related for example to socio-demographic criteria. Bardsley and Chambers (1984) attach the importance to the variables in terms of cost which are in fact penalties associated to the variables. Let $\mathbf{C}$ be the diagonal matrix of nonnegative costs which can measure the acceptable level of error while estimating the totals of variable from the $\mathbf{X}$ matrix,

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{0}_{(q,p-q)} \\ \mathbf{0}_{(p-q,p)} & \mathbf{C}_2 \end{pmatrix},$$

where $\mathbf{C}_1$, respectively $\mathbf{C}_2$, is the relative diagonal cost matrix of size $q \times q$ associated to $\widetilde{\mathbf{X}}_1$, respectively of size $(p-q) \times (p-q)$ associated to $\widetilde{\mathbf{X}}_2$.

As discussed in the above section, allowing an infinite cost C_j means that the associated constraint is exactly satisfied. Bardsley and Chambers (1984) consider the case when constraints corresponding to $\mathbf{X}_1, \ldots, \mathbf{X}_q$ are all exactly satisfied. This means $\mathbf{C}_1 = \infty$ and hence, weights may be derived using relation (2.31) with $\mathbf{C}_1^{-1} = \mathbf{0}_{(q \times q)}$. The weights using this *partially penalized ridge* regression and abbreviated as $\mathbf{w}_{ppr}$ below can be written as,

$$\begin{aligned} \mathbf{w}_{ppr} &= \mathbf{1}_s - \left(\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{1s}, \quad \mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{2s}\right) \begin{pmatrix} \widetilde{\mathbf{X}}_{1s}'\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{1s} & \widetilde{\mathbf{X}}_{1s}'\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{2s} \\ \widetilde{\mathbf{X}}_{2s}'\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{1s} & \widetilde{\mathbf{X}}_{2s}'\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{2s} + \mathbf{C}_2^{-1} \end{pmatrix}^{-1} \\ & \quad \begin{pmatrix} \widetilde{\mathbf{X}}_{1s}'\mathbf{1}_s - \widetilde{\mathbf{X}}_1'\mathbf{1}_U \\ \widetilde{\mathbf{X}}_{2s}'\mathbf{1}_s - \widetilde{\mathbf{X}}_2'\mathbf{1}_U \end{pmatrix}, \end{aligned} \tag{2.36}$$

where $\widetilde{\mathbf{X}}_{1s}$, respectively $\widetilde{\mathbf{X}}_{2s}$, is the sample restriction of $\widetilde{\mathbf{X}}_1$, respectively of $\widetilde{\mathbf{X}}_2$. In particular, we have $\mathbf{w}_{ppr}'\widetilde{\mathbf{X}}_{1s} = \mathbf{1}_U'\widetilde{\mathbf{X}}_1$. Using a calibration approach, the weights are derived using the above formula with $\mathbf{V}_s$ replaced by $\widetilde{\mathbf{\Pi}}_s$ and $\mathbf{1}_s$ by $\mathbf{d}_s$.

Now, if the cost matrix $\mathbf{C}_2$ also goes to infinity, then the constraints corresponding to variables in $\widetilde{\mathbf{X}}_2$ are also exactly satisfied. Hence, $\mathbf{w}_{ppr} = \mathbf{w}_{MB}$ and the estimator using the weights so derived is again nothing else than the best linear unbiased estimator $\hat{t}_{BLUE}$ given by (2.24) and derived under the model ξ that uses the whole matrix $\mathbf{X}$. Moreover, in the case $\mathbf{C}_2 \to \mathbf{0}_{(p-q,p-q)}$ the variables included in $\widetilde{\mathbf{X}}_2$ are discarded from the constraints and thus the model will include only the calibration variables from $\widetilde{\mathbf{X}}_1$,

$$\mathbf{w}_{ppr} \to \mathbf{w}_{ppr}^{(1)} = \mathbf{h} - \mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{1s}\left(\widetilde{\mathbf{X}}_{1s}'\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{1s}\right)^{-1}(\mathbf{1}_s'\widetilde{\mathbf{X}}_{1s} - \mathbf{1}_U'\widetilde{\mathbf{X}}_1)'.$$

The penalized estimator becomes the best unbiased estimator computed under the restricted model that uses only the matrix $\widetilde{\mathbf{X}}_1$. Since $\hat{t}_{BLUE}$ based on the whole model ξ as well as on the restricted model with $\widetilde{\mathbf{X}}_1$ are two extreme estimators as $\mathbf{C}_2$ varies from ∞ to 0, Bardsley and Chambers (1984) called the estimator

that uses weights $\mathbf{w}_{ppr}$ *an interpolated estimator* between the two extremes. So, the penalized ridge estimator may be considered as a trade-off between an over-specified model and an under-specified model.

Using matrix algebra, one can show the following result which shows that the partially penalized weights may be obtained as solution of two different optimization problems.

Proposition 2. *The ridge weights* $\mathbf{w}_{ppr}$ *verifying the optimization problem* ***(P3)*** *with the inverse matrix cost*

$$\mathbf{C}^{-1} = \begin{pmatrix} \mathbf{0}_{(q,q)} & \mathbf{0}_{(q,p-q)} \\ \mathbf{0}_{(p-q,p)} & \mathbf{C}_2^{-1} \end{pmatrix}, \tag{2.37}$$

may be obtained also as a solution of the following optimization problem

$$\begin{aligned} \textbf{(P7)}: \quad \mathbf{w}_{ppr} &= argmin_{\mathbf{w}}(\mathbf{w}_s - \mathbf{1}_s)'\mathbf{V}_s(\mathbf{w}_s - \mathbf{1}_s) \\ &\quad +(\mathbf{w}_s'\widetilde{\mathbf{X}}_{2s} - \mathbf{1}_U'\widetilde{\mathbf{X}}_2)\mathbf{C}_2(\mathbf{w}_s'\widetilde{\mathbf{X}}_{2s} - \mathbf{1}_U'\widetilde{\mathbf{X}}_2)' \\ &\quad subject\ to \quad \mathbf{w}_s'\widetilde{\mathbf{X}}_{1s} = \mathbf{1}_U'\widetilde{\mathbf{X}}_1. \end{aligned}$$

The partial penalized estimator for the total t_y *becomes*

$$\hat{t}_{ppr} = \mathbf{w}_{ppr}'\mathbf{y}_s = \mathbf{1}_s'\mathbf{y}_s - (\mathbf{1}_s'\widetilde{\mathbf{X}}_{1s} - \mathbf{1}_U'\widetilde{\mathbf{X}}_1)\hat{\mathbf{b}} + (\mathbf{1}_s'\widetilde{\mathbf{X}}_{2s} - \mathbf{1}_U'\widetilde{\mathbf{X}}_2)\hat{\mathbf{u}}$$

where $\hat{\mathbf{b}} = \left(\widetilde{\mathbf{X}}_{1s}'\mathbf{\Omega}_{ss}^{-1}\widetilde{\mathbf{X}}_{1s}\right)^{-1}\widetilde{\mathbf{X}}_{1s}'\mathbf{\Omega}_{ss}^{-1}\mathbf{y}_s$, $\mathbf{\Omega}_{ss} = \mathbf{V_s} + \widetilde{\mathbf{X}}_{2s}\mathbf{C}_2\widetilde{\mathbf{X}}_{2s}'$ *and* $\hat{\mathbf{u}} = \mathbf{C}_2\widetilde{\mathbf{X}}_{2s}'\mathbf{\Omega}_s^{-1}(\mathbf{y}_s - \widetilde{\mathbf{X}}_{1s}\hat{\mathbf{b}})$.

It may be written in a simple form as

$$t_{ppr} = \mathbf{1}_s'\mathbf{y}_s + \left(\sum_{U-s}\mathbf{x}_k'\right)\hat{\boldsymbol{\beta}}_{MBR},$$

where $\hat{\boldsymbol{\beta}}_{MBR} = \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2^{-1} \end{pmatrix}\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s.$

Proof. Let $\mathbf{w}_{ppc}^{(1)}$ be solution of the optimization problem **P3** is,

$$\mathbf{w}_{ppc}^{(1)} = \text{argmin}_{\mathbf{w}}\left[(\mathbf{w} - \mathbf{1}_s)'\mathbf{V}_s^{-1}(\mathbf{w} - \mathbf{1}_s) + (\mathbf{w}'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})\mathbf{C}(\mathbf{w}'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})'\right].$$

The loss function is

$$\mathcal{L}(\mathbf{w}) = (\mathbf{w} - \mathbf{1}_s)'\mathbf{V}_s^{-1}(\mathbf{w} - \mathbf{1}_s) + (\mathbf{w}'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})\mathbf{C}(\mathbf{w}'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})'.$$

We derive $\mathcal{L}(\mathbf{w})$ with respect to $\mathbf{w}$ and we solve $\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} = 0$ to obtain $\mathbf{w}_{ppc}^{(1)}$. We get,

$$\mathbf{w}_{ppc}^{(1)} = \mathbf{1}_s - (\mathbf{V}_s + \mathbf{X}_s\mathbf{C}\mathbf{X}_s')^{-1}\mathbf{X}_s\mathbf{C}(\mathbf{X}_s'\mathbf{1}_s - \mathbf{X}'\mathbf{1}_U)$$

and by the property of inverse of the sum of the matrices (Henderson and Searle, 1981),

$$(\mathbf{V}_s + \mathbf{X}_s\mathbf{C}\mathbf{X}_s')^{-1}\mathbf{X}_s\mathbf{C} = \mathbf{V}_s^{-1}\mathbf{X}_s(\mathbf{C}^{-1} + \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}$$

so our expression for the partially penalized calibration weights in this case becomes,

$$\begin{aligned} \mathbf{w}_{ppr}^{(1)} &= \mathbf{1}_s - \mathbf{V}_s^{-1}\mathbf{X}_s(\mathbf{C}^{-1} + \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}(\mathbf{X}_s'\mathbf{1}_s - \mathbf{X}'\mathbf{1}_U) \\ &= \mathbf{1}_s - \mathbf{V}_s^{-1}\mathbf{X}_s\mathbf{R}^{-1}(\mathbf{X}_s'\mathbf{1}_s - \mathbf{X}'\mathbf{1}_U) \end{aligned} \tag{2.38}$$

where

$$\mathbf{R}^{-1} = (\mathbf{C}^{-1} + \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}.$$

Now if the cost for the calibration variables $\mathbf{X}_1$ is infinity, then the inverse of cost matrix $\mathbf{C}$ will be

$$\mathbf{C}^{-1} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2^{-1} \end{pmatrix}.$$

Hence,

$$\begin{aligned} \mathbf{R}^{-1} &= \left(\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2^{-1} \end{pmatrix} + \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s \right)^{-1} \\ &= \begin{pmatrix} \tilde{\mathbf{X}}_{1s}'\mathbf{V}_s^{-1}\tilde{\mathbf{X}}_{1s} & \tilde{\mathbf{X}}_{1s}'\mathbf{V}_s^{-1}\tilde{\mathbf{X}}_{2s} \\ \tilde{\mathbf{X}}_{2s}'\mathbf{V}_s^{-1}\mathbf{X}_{1s} & \mathbf{C}_2^{-1} + \tilde{\mathbf{X}}_{2s}'\mathbf{V}_s^{-1}\tilde{\mathbf{X}}_{2s} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{L} \end{pmatrix}^{-1} \end{aligned} \tag{2.39}$$

with $\mathbf{A} = \tilde{\mathbf{X}}'_{1s}\mathbf{V}_s^{-1}\tilde{\mathbf{X}}_{1s}$, $\mathbf{B} = \tilde{\mathbf{X}}'_{1s}\mathbf{V}_s^{-1}\tilde{\mathbf{X}}_{2s}$ and $\mathbf{L} = \mathbf{C}_2^{-1} + \tilde{\mathbf{X}}'_{2s}\mathbf{V}_s^{-1}\tilde{\mathbf{X}}_{2s}$. We have (Rao, 1969),

$$\mathbf{R}^{-1} = \begin{pmatrix} \mathbf{H}^{-1} & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}' & \mathbf{E}^{-1} \end{pmatrix} \tag{2.40}$$

where $\mathbf{H}^{-1} = (\tilde{\mathbf{X}}'_{1s}\mathbf{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{1s})^{-1}$ with $\mathbf{\Omega}_{ss} = \mathbf{V}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}'_{2s}$ and its inverse as,

$$\begin{aligned} \mathbf{\Omega}_{ss}^{-1} &= (\mathbf{V}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}'_{2s})^{-1} \\ &= \mathbf{V}_s^{-1} - \mathbf{V}_s^{-1}\tilde{\mathbf{X}}_{2s}(\mathbf{C}_2^{-1} + \tilde{\mathbf{X}}_{2s}\mathbf{V}_s^{-1}\tilde{\mathbf{X}}'_{2s})^{-1}\tilde{\mathbf{X}}'_{2s}\mathbf{V}_s^{-1}, \end{aligned} \tag{2.41}$$

$\mathbf{E} = \mathbf{L} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}$ and $\mathbf{F} = \mathbf{A}^{-1}\mathbf{B}$ where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{L}$ are already defined above. For the above value of $\mathbf{R}^{-1}$, the partially penalized calibrated weights given in expression (2.38) become,

$$\begin{aligned} \mathbf{w}_{ppr}^{(1)} &= \mathbf{1}_s - \mathbf{V}_s^{-1}\mathbf{X}_s(\mathbf{C}^{-1} + \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}(\mathbf{X}_s'\mathbf{1}_s - \mathbf{X}'\mathbf{1}_U) \\ &= \mathbf{1}_s - \mathbf{V}_s^{-1}(\tilde{\mathbf{X}}_{1s}, \tilde{\mathbf{X}}_{2s}) \begin{pmatrix} \mathbf{H}^{-1} & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}' & \mathbf{E}^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{X}}'_{1s}\mathbf{1}_s - \tilde{\mathbf{X}}'_1\mathbf{1}_U \\ \tilde{\mathbf{X}}'_{2s}\mathbf{1}_s - \tilde{\mathbf{X}}'_2\mathbf{1}_U \end{pmatrix} \end{aligned}$$

where $\mathbf{\Omega}_{ss} = \mathbf{V}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}'_{2s}$. So the weights $\mathbf{w}_{ppr}^{(1)}$ can be written as,

$$\begin{aligned} \mathbf{w}_{ppr}^{(1)} = \mathbf{1}_s - \mathbf{V}_s^{-1}\Big[&\Big(\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1} - \tilde{\mathbf{X}}_{2s}\mathbf{L}^{-1}\mathbf{B}'\mathbf{H}^{-1}\Big)\Big(\tilde{\mathbf{X}}'_{1s}\mathbf{1}_s - \tilde{\mathbf{X}}'_1\mathbf{1}_U\Big) + \\ &\Big(-\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\mathbf{B}\mathbf{L}^{-1} + \tilde{\mathbf{X}}_{2s}\left(\mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{B}'\mathbf{H}^{-1}\mathbf{B}\mathbf{L}^{-1}\right)\Big)\Big(\tilde{\mathbf{X}}'_{2s}\mathbf{1}_s - \tilde{\mathbf{X}}'_2\mathbf{1}_U\Big)\Big], \end{aligned} \tag{2.42}$$

since $-\mathbf{E}^{-1}\mathbf{F}' = -\mathbf{L}^{-1}\mathbf{B}'\mathbf{H}^{-1}$ and $\mathbf{E}^{-1} = \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{B}'\mathbf{H}^{-1}\mathbf{B}\mathbf{L}^{-1}$. Consider now the optimization problem **P7**. Using the same idea as before we get,

$$\begin{aligned} \mathbf{w}_{ppr}^{(2)} = \mathbf{\Omega}_{ss}^{-1}\Bigg[&\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\Bigg(\tilde{\mathbf{X}}'_1\mathbf{1}_U - \underbrace{\tilde{\mathbf{X}}_{1s}\mathbf{\Omega}_{ss}^{-1}(\mathbf{V}_s\mathbf{1}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}'_2\mathbf{1}_U)}_{*}\Bigg) + \\ &\underbrace{\mathbf{V}_s\mathbf{1}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}'_2\mathbf{1}_U}_{**}\Bigg] \end{aligned} \tag{2.43}$$

Consider the above terms (*) and we use the fact that $\mathbf{V}_s = \mathbf{\Omega}_{ss} - \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}'_{2s}$

$$\begin{aligned} \tilde{\mathbf{X}}_{1s}\mathbf{\Omega}_{ss}^{-1}(\mathbf{V}_s\mathbf{1}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}'_2\mathbf{1}_U) &= \tilde{\mathbf{X}}_{1s}\mathbf{\Omega}_{ss}^{-1}((\mathbf{\Omega}_{ss} - \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}'_{2s})\mathbf{1}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}'_2\mathbf{1}_U) \\ &= \tilde{\mathbf{X}}'_{1s}\mathbf{1}_s + \tilde{\mathbf{X}}'_{1s}\mathbf{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{2s}\mathbf{C}_2\left(\tilde{\mathbf{X}}'_2\mathbf{1}_U - \tilde{\mathbf{X}}'_{2s}\mathbf{1}_s\right). \end{aligned} \tag{2.44}$$

Also consider (**),

$$\begin{aligned}\mathbf{V}_s\mathbf{1}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}_2'\mathbf{1}_U &= (\boldsymbol{\Omega}_{ss} - \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}_{2s}')\mathbf{1}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}_2'\mathbf{1}_U \\ &= \boldsymbol{\Omega}_{ss}\mathbf{1}_s - \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}_{2s}'\mathbf{1}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}_2'\mathbf{1}_U \\ &= \boldsymbol{\Omega}_{ss}\mathbf{1}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2(\tilde{\mathbf{X}}_2'\mathbf{1}_U - \tilde{\mathbf{X}}_{2s}'\mathbf{1}_s). \qquad (2.45)\end{aligned}$$

Using above equations (2.44 and 2.45), we get,

$$\begin{aligned}\mathbf{w}_{ppr}^{(2)} &= \boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\left[(\tilde{\mathbf{X}}_1'\mathbf{1}_U) - \tilde{\mathbf{X}}_{1s}'\boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{2s}\mathbf{C}_2(\tilde{\mathbf{X}}_2'\mathbf{1}_U - \tilde{\mathbf{X}}_{2s}'\mathbf{1}_s)\right] + \boldsymbol{\Omega}_{ss}^{-1}\boldsymbol{\Omega}_{ss}\mathbf{1}_s \\ &\quad + \boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{2s}\mathbf{C}_2(\tilde{\mathbf{X}}_2'\mathbf{1}_U - \tilde{\mathbf{X}}_{2s}'\mathbf{1}_s) \\ &= \mathbf{1}_s + \boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\left(\tilde{\mathbf{X}}_1'\mathbf{1}_U - \tilde{\mathbf{X}}_{1s}'\mathbf{1}_s\right) - \boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\tilde{\mathbf{X}}_{1s}'\boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{2s}'\mathbf{C}_2\left(\tilde{\mathbf{X}}_2'\mathbf{1}_U - \tilde{\mathbf{X}}_{2s}'\mathbf{1}_s\right) \\ &\quad + \boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{2s}\mathbf{C}_2\left(\tilde{\mathbf{X}}_2'\mathbf{1}_U - \tilde{\mathbf{X}}_{2s}'\mathbf{1}_s\right) \\ &= \mathbf{1}_s + \boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\left(\tilde{\mathbf{X}}_1'\mathbf{1}_U - \tilde{\mathbf{X}}_{1s}'\mathbf{1}_s\right) + (\boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{2s}\mathbf{C}_2 - \boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\tilde{\mathbf{X}}_{1s}'\boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{2s}'\mathbf{C}_2) \\ &\quad \left(\tilde{\mathbf{X}}_2'\mathbf{1}_U - \tilde{\mathbf{X}}_{2s}'\mathbf{1}_s\right). \qquad (2.46)\end{aligned}$$

Consider again 2.41, we get

$$\boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1} = \mathbf{V}_s^{-1}(\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1} - \tilde{\mathbf{X}}_{2s}\mathbf{L}^{-1}\mathbf{B}'\mathbf{H}^{-1}) \qquad (2.47)$$

and also,

$$\begin{aligned}\boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{2s}'\mathbf{C}_2 &= (\mathbf{V}_s + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}_{2s}')^{-1}\tilde{\mathbf{X}}_{2s}\mathbf{C}_2 \\ &= \mathbf{V}_s^{-1}\tilde{\mathbf{X}}_{2s}'(\mathbf{C}_2^{-1} + \tilde{\mathbf{X}}_{2s}\mathbf{C}_2\tilde{\mathbf{X}}_{2s}')^{-1} \\ &= \mathbf{V}_s^{-1}\tilde{\mathbf{X}}_{2s}'\mathbf{L}^{-1}. \qquad (2.48)\end{aligned}$$

Thus using 2.47 we get,

$$\begin{aligned}&\boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\tilde{\mathbf{X}}_{1s}'\boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{2s}'\mathbf{C}_2 \\ &= \boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\tilde{\mathbf{X}}_{1s}'\mathbf{V}_s^{-1}\tilde{\mathbf{X}}_{2s}'\mathbf{L}^{-1} \\ &= \boldsymbol{\Omega}_{ss}^{-1}\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\mathbf{B}\mathbf{L}^{-1} \\ &= \mathbf{V}_s^{-1}(\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\mathbf{B}\mathbf{L}^{-1} - \tilde{\mathbf{X}}_{2s}\mathbf{L}^{-1}\mathbf{B}'\mathbf{H}^{-1}\mathbf{B}\mathbf{L}^{-1}) \qquad (2.49)\end{aligned}$$

Hence using 2.47, 2.48 and 2.49, the weights in equation 2.46 become,

$$\begin{aligned}\mathbf{w}_{ppr}^{(2)} &= \mathbf{1}_s - \mathbf{V}_s^{-1}\left[\left(\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1} - \tilde{\mathbf{X}}_{2s}\mathbf{L}^{-1}\mathbf{B}'\mathbf{H}^{-1}\right)\left(\tilde{\mathbf{X}}_{1s}'\mathbf{1}_s - \tilde{\mathbf{X}}_1'\mathbf{1}_U\right) + \right.\\ &\quad \left(-\tilde{\mathbf{X}}_{1s}\mathbf{H}^{-1}\mathbf{B}\mathbf{L}^{-1} + \tilde{\mathbf{X}}_{2s}\left(\mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{B}'\mathbf{H}^{-1}\mathbf{B}\mathbf{L}^{-1}\right)\right)\\ &\quad \left.\left(\tilde{\mathbf{X}}_{2s}'\mathbf{1}_s - \tilde{\mathbf{X}}_2'\mathbf{1}_U\right)\right] \qquad (2.50)\end{aligned}$$

which are identical to the weights $\mathbf{w}_{ppc}^{(1)}$ given by 2.43.

□

This optimization problem **(P7)** is used by Park and Yang (2008) and Guggemos and Tillé (2010). Using the model ξ given by (2.19) with intercept, Park and Yang (2008) aim at estimating the mean $\bar{y}_U = \sum_U y_k/N$ of the variable of interest $\mathcal{Y}$ using a Hájek-type estimator. This means that they use a weighted estimator with weights that sum up to unity and being as close as possible to the Hájek (1971) weights,

$$\alpha_i = \frac{\pi_i^{-1}}{\sum_s \frac{1}{\pi_i}}.$$

This means that the optimization problem **(P7)** is used with $\mathbf{1}_s$ replaced by $\boldsymbol{\alpha}_s = (\alpha_i)_{i\in s}$. They build two partially penalized estimators. In the first case, $\widetilde{\mathbf{X}}_1 = \mathbf{1}_U$ and in the second case, $\widetilde{\mathbf{X}}_1 = (\mathbf{1}_U, \mathbf{X}_2, \ldots, \mathbf{X}_q)$. Weights may be derived using relation (2.36). Slightly simplified formulas are obtained since $\mathbf{1}_s'\boldsymbol{\alpha}_s - \mathbf{1}_U'\mathbf{1}_U/N = 0$. In a linear regression context, it is not very common to consider the penalty or the cost matrix $\mathbf{C}^{-1}$ given by (3.1.1). This is more likely to happen with a mixed model. Using a calibration approach, Guggemos and Tillé (2010) consider the following mixed model

$$\xi' : \quad \mathbf{y} = \widetilde{\mathbf{X}}_1\mathbf{b} + \widetilde{\mathbf{X}}_2\mathbf{u} + \boldsymbol{\eta},$$

where $\mathbf{u}$ is a random effect vector. We replace the matrix $\mathbf{V}_s$ by $\widetilde{\boldsymbol{\Pi}}_s$, and the vector $\mathbf{1}_s$ by $\mathbf{d}_s$ in the objective function **(P7)** from (2). Guggemos and Tillé (2010) consider also that the second term of the objective function **(P7)** depends on a penalty parameter and they suggest the Fisher scoring algorithm to compute it. The value of the penalty parameter is obtained at the convergence of the

Fisher scoring algorithm. They give also the application of the partially penalized calibration for the estimation of finite population totals in a small area context.

2.3.4 Calibration on uncertain auxiliary information

In presence of several external estimations which may be considered as uncertain, Deville (1999) suggested another construction which uses in fact the Bayesian interpretation of the ridge estimator given in section 2.2. Consider that another estimation $\hat{t}_{x^*,d} = \mathbf{d}_s'\mathbf{X}^*$ based on the auxiliary information $\mathbf{X}^*$ is available from external sources such as previous surveys. We have also the current estimation based on $\mathbf{X}$. We suppose that the variances of $\hat{t}_{x^*,d}$ and $\hat{t}_{x,d}$ are known and the covariance between the two sources is zero. We suppose also that the covariance between $\hat{t}_{x^*,d}$ and $\hat{t}_{y,d}$ is also zero. Deville looks for linear weighted estimators for t_y of the form

$$\hat{t}_w = \mathbf{d}_s'\mathbf{y}_s + (\mathbf{d}_s'\mathbf{X}_s^* - \mathbf{d}_s'\mathbf{X}_s)\boldsymbol{\beta} = \hat{t}_{y,d} + (\hat{t}_{x^*,d} - \hat{t}_{x,d})'\boldsymbol{\beta}. \tag{2.51}$$

The optimal value of the unknown parameter $\boldsymbol{\beta}$ is the one that minimizes the sampling variance of $\hat{t}_w$. We find

$$\boldsymbol{\beta}_{opt} = \big(\mathrm{Var}(\hat{t}_{x^*,d}) + \mathrm{Var}(\hat{t}_{x,d})\big)^{-1}\,\mathrm{Cov}(\hat{t}_{y,d}, \hat{t}_{x,d}),$$

and the same value may be derived by using a variance minimization criteria as in Montanari (1987) plus a penalty term, namely

$$\textbf{(P8)}: \quad \boldsymbol{\beta}_{opt} = \mathrm{argmin}_{\boldsymbol{\beta}}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\boldsymbol{\Delta}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{X}_s'^{*}\boldsymbol{\Delta}\mathbf{X}_s^*\boldsymbol{\beta}, \tag{2.52}$$

where $\boldsymbol{\Delta} = (\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}})_{k,l\in U}$. We remark that the penalty is now on the variance of $\hat{t}_{x^*,d}$.

The estimation of t_y given by (2.51) computed for $\boldsymbol{\beta} = \boldsymbol{\beta}_{opt}$ may be improved by replacing $\hat{t}_{x^*,d}$ with the best unbiased linear estimator of $\hat{t}_{x^*,d}$ and $\hat{t}_{x,d}$. This is equivalent to determine the posterior estimation knowing that the priori estimation given by the auxiliary information is $\hat{t}_{x^*,d}$ and the actual estimation is $\hat{t}_{x,d}$. One

may use relation (2.14) to find the posterior estimation as

$$\hat{t}^{opt}_{x,x^*} = (\mathbf{I}_p - \mathbf{A})\,\hat{t}_{x^*,d} + \mathbf{A}\hat{t}_{x,d},$$

where $\mathbf{A}$ is a squared p-dimensional matrix given by

$$\mathbf{A} = \mathbf{I}_p - \mathrm{Var}(\hat{t}_{x,d})\left(\mathrm{Var}(\hat{t}_{x,d}) + \mathrm{Var}(\hat{t}_{x^*,d})\right)^{-1}.$$

Then, one can derive the estimator $\hat{t}^{opt}_y$ of t_y from relation (2.51) with $\hat{t}_{x^*,d}$ replaced with $\hat{t}^{opt}_{x,x^*}$,

$$\hat{t}^{opt}_y = \hat{t}_{y,d} + (\hat{t}^{opt}_{x,x^*} - \hat{t}_{x,d})'(\mathrm{Var}(\hat{t}_{x,d}))^{-1}\mathrm{Cov}(\hat{t}_{y,d}, \hat{t}_{x,d}).$$

One can easily obtain that if $y_k = x_k$ for all $k \in U$, we obtain $\hat{t}^{opt}_{x,x^*}$ or equivalently, the estimator is calibrated on $\hat{t}^{opt}_{x,x^*}$. If the variance covariance $\mathrm{Cov}(\hat{t}_{y,d}, \hat{t}_{x,d})$ is estimated by the usual Horvitz-Thompson estimator, $\hat{t}^{opt}_y$ is a linear estimator in y_k with weights w_k given by

$$w_k = d_k + (\hat{t}_{x^*,d} - \hat{t}_{x,d})'(\mathrm{Var}(\hat{t}_{x,d}))^{-1} z_k d_k,$$

where $z_k = \sum_{l\in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{x_l}{\pi_l}$. The main advantage of Deville's construction is that it does not need to determine a penalty parameter as it was the case before. All we need is the variance of the external estimation. Deville (1999) also gives a practical implementation and generalization to the several external estimations.

2.3.5 Statistical properties of ridge estimators with survey data

Ridge-type estimators are biased estimators suggested in classical regression in order to diminish the model mean squared error. Both model-based and design-based penalized estimators given by (2.32) and (2.35) are biased under the model ξ. As for the partial penalized estimator, it is biased under the model ξ but it is unbiased under the model ξ' (Guggemos and Tillé, 2010). Bardsley and Chambers (1984) affirm that the model-based ridge estimator $\hat{t}_{MBR}$ has smaller prediction variance than the best linear unbiased estimator $\hat{t}_{BLUE}$ but they do not give a

rigorous proof. Bellhouse (1987) shows that a predictor $\hat{Y}^{(1)} = \sum_s y_k + (N-n)\hat{\mu}_s^{(1)}$ of the finite population total t_y is better than another predictor $\hat{Y}^{(2)} = \sum_s y_k + (N-n)\hat{\mu}_s^{(2)}$ with respect to the mean square error under the model ξ and the sampling design p if, for every sample s of fixed size n, $\hat{\mu}_s^{(1)}$ is better than $\hat{\mu}_s^{(2)}$ in the sense that

$$E_\xi(\hat{\mu}_s^{(1)} - \mu_{ns})^2 \leq E_\xi(\hat{\mu}_s^{(2)} - \mu_{ns})^2,$$

where μ_{ns} is the unknown prediction of the non sampled mean of $\mathcal{Y}$. Using this result and the same arguments as in Vinod and Ullah (1981), one can get that for any penalty constant κ satisfying $0 < \kappa < 2\sigma^2/\boldsymbol{\beta}'\boldsymbol{\beta}$,

$$E_\xi E_p(\hat{t}_{MBR} - t_y)^2 < E_\xi E_p(\hat{t}_{BLUE} - t_y)^2,$$

where $\hat{t}_{MBR}$ is the ridge model based estimator given by (2.32) for $\mathbf{C}^{-1} = \kappa \mathbf{I}_p$ and $\hat{t}_{BLUE}$ is the best linear unbiased estimator given by (2.24). A necessary and sufficient condition for the ridge estimator $\hat{t}_{MBR}$ to be more efficient than the BLUE estimator $\hat{t}_{BLUE}$ is given in theorem 3.

Dunstan and Chambers (1986) derived confidence intervals for finite population totals estimated using the ridge model-based procedure and robust model-based variance estimators.

In a design-based setting, Park and Yang (2008) determine also the optimal values of the penalty matrix $\mathbf{C_2}$ from optimization problem **(P7)**. Nevertheless, in a design-based framework, the concern was about asymptotic properties of $\hat{t}_{y,Rw}$ given by (2.35) with respect to the sampling design p. As Rao and Singh (1997) stated, *"an important requirement while relaxing benchmark constraints is that for given tolerance levels, the calibration method should ensure design consistency like the generalized regression method."* The asymptotic design unbiasedness and consistency of $\hat{t}_{y,Rw}$ are derived using the equivalence with GREG estimators even if $\hat{t}_{y,Rw}$ has been obtained as a solution of penalized calibration problems. Under broad assumptions (Fuller, 2002), the design-based ridge estimator $\hat{\boldsymbol{\beta}}_\lambda$ of $\boldsymbol{\beta}$ tends in probability to $\boldsymbol{\beta}_\lambda = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{C}^{-1})^{-1}\mathbf{X}'\mathbf{y}$ and the ridge estimator $\hat{t}_{y,Rw}$ is

asymptotically equivalent to

$$\hat{t}_{y,Rw} \simeq \mathbf{d}_s'\mathbf{y}_s - (\mathbf{X}_s'\mathbf{d}_s - \mathbf{X}'\mathbf{1}_U)'\boldsymbol{\beta}_\lambda = \mathbf{d}_s'(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}_\lambda) + \mathbf{1}_U'\mathbf{X}\boldsymbol{\beta}_\lambda,$$

which implies that $\hat{t}_{y,Rw}$ is asymptotically design unbiased and consistent under broad assumptions that provide the design unbiasedness and consistency of the Horvitz-Thompson estimators $\mathbf{d}_s'\mathbf{y}_s$ and $\mathbf{d}_s'\mathbf{X}_s$ (Rao and Singh, 1997 and Théberge, 2000). The asymptotic variance under the sampling design may thus be deduced as being the Horvitz-Thompson variance applied to residuals $y_k - \mathbf{x}_k'\boldsymbol{\beta}_\lambda$.

Statistical properties of model-based and model-assisted ridge estimators

Some of the results depicting the properties of model-based $\hat{t}_{MBR}$ (given by 2.32) and model-assisted $\hat{t}_{GREG,R}$ (given by 2.29) ridge estimator are given in form of results in the following.

We recall that

$$\hat{t}_{MBR} = \mathbf{w}_{MBR}'\mathbf{y}_s = \sum_s y_k + \left(\sum_{U-s} \mathbf{x}_k'\right)\hat{\boldsymbol{\beta}}_{MBR}$$

with $\hat{\boldsymbol{\beta}}_{MBR} = \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s$.

Consider the eigenvalues $\lambda_{1,s} \geq \ldots \geq \lambda_{p,s}$ of $\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s$ with the corresponding eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$ and $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$ be the $p \times p$ matrix of eigenvectors which satisfies $\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s = \mathbf{A}diag(\lambda_{i,s})_{i=1}^p\mathbf{A}' = \mathbf{A}\boldsymbol{\Lambda}_s\mathbf{A}'$ and $\mathbf{A}\mathbf{A}' = \mathbf{I}$ where $\boldsymbol{\Lambda}_s$ is the diagonal matrix of eigenvalues $\lambda_{1,s} \geq \ldots \geq \lambda_{p,s}$ of $\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s$.

Result 4. *1. The ξ-bias of $\hat{t}_{MBR}$ is given by,*

$$Bias_\xi\left(\hat{t}_{MBR}\right) = -\kappa\left(\sum_{U-s}\mathbf{x}_k'\right)\mathbf{A}diag\left(\frac{1}{\lambda_{i,s}+\kappa}\right)_{i=1}^p\mathbf{A}'\boldsymbol{\beta} \tag{2.53}$$

2. $\hat{\boldsymbol{\beta}}_{MBR}$ in function of $\hat{\boldsymbol{\beta}}_{GLS}$ can be written as,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{MBR} &= \mathbf{A}\Delta\mathbf{A}'\hat{\boldsymbol{\beta}}_{GLS} \\ &= \mathbf{A}(diag(\delta_i))_{i=1}^p\mathbf{A}'\hat{\boldsymbol{\beta}}_{GLS}\end{aligned} \tag{2.54}$$

Proof. We have that

$$\begin{aligned}
\hat{t}_{MBR} &= \sum_s y_k + \left(\sum_{U-s} \mathbf{x}'_k\right)\hat{\boldsymbol{\beta}}_{MBR} \text{ , so} \\
\hat{t}_{MBR} - t_y &= \sum_s y_k + \left(\sum_{U-s} \mathbf{x}'_k\right)\hat{\boldsymbol{\beta}}_{MB,R} - \sum_U y_k \\
&= \sum_s (\mathbf{x}'_k\boldsymbol{\beta} + \varepsilon_k) + \left(\sum_{U-s} \mathbf{x}'_k\right)\hat{\boldsymbol{\beta}}_{MBR} - \sum_U (\mathbf{x}'_k\boldsymbol{\beta} + \varepsilon_k) \\
&= \sum_s (\mathbf{x}'_k\boldsymbol{\beta}) + \sum_s \varepsilon_k + \left(\sum_{U-s} \mathbf{x}'_k\right)\hat{\boldsymbol{\beta}}_{MBR} - \sum_U \mathbf{x}'_k\boldsymbol{\beta} - \sum_U \varepsilon_k \\
&= \left(\sum_s \mathbf{x}'_k - \sum_U \mathbf{x}'_k\right)\boldsymbol{\beta} + \left(\sum_{U-s} \mathbf{x}'_k\right)\hat{\boldsymbol{\beta}}_{MBR} - \sum_{U-s} \varepsilon_k \\
&= \left(\sum_{U-s} \mathbf{x}'_k\right)\hat{\boldsymbol{\beta}}_{MBR} - \left(\sum_{U-s} \mathbf{x}'_k\right)\boldsymbol{\beta} - \sum_{U-s} \varepsilon_k
\end{aligned}$$

So,

$$\begin{aligned}
E_\xi(\hat{t}_{MBR} - t_y) &= \left(\sum_{U-s} \mathbf{x}'_k\right)\left(\hat{\boldsymbol{\beta}}_{MBR} - \boldsymbol{\beta}\right) - \sum_{U-s} \varepsilon_k \\
&= \left(\sum_{U-s} \mathbf{x}'_k\right) E_\xi\left(\hat{\boldsymbol{\beta}}_{MBR} - \boldsymbol{\beta}\right) \\
&= \left(\sum_{U-s} \mathbf{x}'_k\right) Bias_\xi\left(\hat{\boldsymbol{\beta}}_{MBR}\right) \qquad (2.55)
\end{aligned}$$

Here we use a transformation, $\mathbf{X}'_s\mathbf{V}_s^{-1}\mathbf{X}_s = \mathbf{A}(diag\lambda_i)_{i=1}^p\mathbf{A}' = \mathbf{A}\boldsymbol{\Lambda}_s\mathbf{A}'$ with $\mathbf{X}_s = \mathbf{V}_s^{\frac{1}{2}}\mathbf{O}\boldsymbol{\Lambda}_s^{\frac{1}{2}}\mathbf{A}'$ where $\mathbf{O}$ is an $n \times p$ matrix of coordinates of the observations along the principal axes of $\mathbf{X}_s$, standardized in the sense, $\mathbf{O}'\mathbf{O} = \mathbf{I}$ and $\mathbf{V}_s$ is the variance of $\boldsymbol{\varepsilon}_s$ as defined earlier. Then, $\hat{\boldsymbol{\beta}}_{GLS}$ given in 2.17 can be written as,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{GLS} &= (\mathbf{X}'_s\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{V}_s^{-1}\mathbf{y}_s \\
&= (\mathbf{A}\boldsymbol{\Lambda}_s\mathbf{A}')^{-1}\mathbf{A}\boldsymbol{\Lambda}_s^{\frac{1}{2}}\mathbf{O}'\mathbf{V}_s^{\frac{1}{2}}\mathbf{V}_s^{-1}\mathbf{y}_s \\
&= \mathbf{A}\boldsymbol{\Lambda}_s^{-1}\mathbf{A}'\mathbf{A}\boldsymbol{\Lambda}_s^{\frac{1}{2}}\mathbf{O}'\mathbf{V}_s^{-\frac{1}{2}}\mathbf{y}_s \\
&= \mathbf{A}\boldsymbol{\Lambda}_s^{-1}\boldsymbol{\Lambda}_s^{\frac{1}{2}}\mathbf{O}'\mathbf{V}_s^{-\frac{1}{2}}\mathbf{y}_s \\
&= \mathbf{A}\boldsymbol{\Lambda}_s^{-\frac{1}{2}}\mathbf{O}'\mathbf{V}_s^{-\frac{1}{2}}\mathbf{y}_s \qquad (2.56)
\end{aligned}$$

We can therefore write $\hat{\boldsymbol{\beta}}_{MBR}$ in function of $\hat{\boldsymbol{\beta}}_{GLS}$ as follows,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{MBR} &= \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s \\
&= \left(\mathbf{A}\boldsymbol{\Lambda}_s\mathbf{A}' + \kappa\mathbf{A}\mathbf{A}'\right)^{-1}\mathbf{A}\boldsymbol{\Lambda}_s^{\frac{1}{2}}\mathbf{O}'\mathbf{V}_s^{-\frac{1}{2}}\mathbf{y_s} \\
&= \mathbf{A}\left(\boldsymbol{\Lambda}_s + \kappa\mathbf{I}\right)^{-1}\boldsymbol{\Lambda}_s\mathbf{A}'\mathbf{A}\boldsymbol{\Lambda}_s^{-1}\boldsymbol{\Lambda}_s^{\frac{1}{2}}\mathbf{O}'\mathbf{V}_s^{-\frac{1}{2}}\mathbf{y_s} \\
&= \mathbf{A}\left(\boldsymbol{\Lambda}_s + \kappa\mathbf{I}\right)^{-1}\boldsymbol{\Lambda}_s\mathbf{A}'\mathbf{A}\boldsymbol{\Lambda}_s^{-\frac{1}{2}}\mathbf{O}'\mathbf{V}_s^{-\frac{1}{2}}\mathbf{y_s}.
\end{aligned}$$

Using equation(2.56), we get

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{MBR} &= \mathbf{A}\Delta\mathbf{A}'\hat{\boldsymbol{\beta}}_{GLS} \\
&= \mathbf{A}(diag(\delta_i))_{i=1}^{p}\mathbf{A}'\hat{\boldsymbol{\beta}}_{GLS} \qquad (2.57)
\end{aligned}$$

where $\Delta = diag\left(\delta_i\right)_{i=1}^{p}$ with $\delta_i = \frac{\lambda_{i,s}}{\lambda_{i,s}+\kappa}$ is a diagonal matrix of *shrinkage factors*.We have declining deltas for strictly positive κ and strictly declining eigenvalues which means that the so-called *shrinkage factor* shrinks the coefficient matrix for the declining eigenvalues given the fact that $\kappa \in (0, \infty)$.
For the bias of $\hat{\boldsymbol{\beta}}_{MBR}$ consider again,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{MBR} &= \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s \\
&= \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}(\mathbf{X}_s\boldsymbol{\beta} + \boldsymbol{\varepsilon}_s) \\
&= \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s\boldsymbol{\beta} + (\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I})^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\boldsymbol{\varepsilon}_s \\
&= \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}\right)^{-1}\left[(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}) - \kappa\mathbf{I}\right]\boldsymbol{\beta} + (\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I})^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\boldsymbol{\varepsilon}_s \\
\hat{\boldsymbol{\beta}}_{MBR} &= \boldsymbol{\beta} - \kappa\left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}\right)^{-1}\boldsymbol{\beta} + \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\boldsymbol{\varepsilon}_s,
\end{aligned}$$

applying ξ-expectation on both sides with $E(\boldsymbol{\varepsilon}_s) = 0$, we get the bias of $\hat{\boldsymbol{\beta}}_{MBR}$ as,

$$Bias_{\xi}(\hat{\boldsymbol{\beta}}_{MBR}) = E_{\xi}(\hat{\boldsymbol{\beta}}_{MBR}) - \boldsymbol{\beta} = -\kappa\left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}\right)^{-1}\boldsymbol{\beta} \qquad (2.58)$$

Referring to the equations (2.57 and 2.58), with $\boldsymbol{\beta}$ is the model parameter, we can

write the ξ-bias of $\hat{t}_{MBR}$ as,

$$\begin{aligned} E_\xi(\hat{t}_{MB,R}) - t_y &= \left(\sum_{U-s} \mathbf{x}_k'\right) Bias_\xi\left(\hat{\boldsymbol{\beta}}_{MBR}\right) \\ &= -\kappa \left(\sum_{U-s} \mathbf{x}_k'\right) \left(\mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{X}_s' + \kappa \mathbf{I}\right)^{-1} \boldsymbol{\beta}, \end{aligned}$$

which leads to

$$Bias_\xi\left(\hat{t}_{MBR}\right) = -\kappa \left(\sum_{U-s} \mathbf{x}_k'\right) \mathbf{A} diag\left(\frac{1}{\lambda_i + k}\right) \mathbf{A}' \boldsymbol{\beta}$$

□

Result 5. *The ξ-mean squared error of $\hat{t}_{MBR}$ is given by,*

$$\begin{aligned} E_\xi(\hat{t}_{MBR} - t_y)^2 &= \sigma^2 \left(\sum_{U-s} \mathbf{x}_k'\right) \left(\mathbf{A} diag\left(\frac{\lambda_i}{(\lambda_i + \kappa)^2}\right)_{i=1}^{p} \mathbf{A}'\right) \left(\sum_{U-s} \mathbf{x}_k'\right)' + \sigma^2 \sum_{\bar{s}} v_k \\ &\quad + \kappa^2 \left(\sum_{U-s} \mathbf{x}_k'\right) \mathbf{A} diag\left(\frac{1}{\lambda_i + \kappa}\right)_{i=1}^{p} \mathbf{A}' \boldsymbol{\beta} \boldsymbol{\beta}' \mathbf{A} diag\left(\frac{1}{\lambda_i + \kappa}\right)_{i=1}^{p} \mathbf{A}' \left(\sum_{U-s} \mathbf{x}_k'\right)' \end{aligned}$$

Proof. We have that

$$E_\xi(\hat{t}_{MBR} - t_y)^2 = \mathrm{Var}_\xi(\hat{t}_{MB,R} - t_y) + (E_\xi(\hat{t}_{MB,R} - t_y))^2$$

where

$$\begin{aligned} \mathrm{Var}_\xi(\hat{t}_{MBR} - t_y) &= \mathrm{Var}_\xi \left(\sum_s (\mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k) + \left(\sum_{U-s} \mathbf{x}_k'\right) \hat{\boldsymbol{\beta}}_{MBR} - \sum_U (\mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k)\right) \\ &= \mathrm{Var}_\xi \left(\left(\sum_{U-s} \mathbf{x}_k'\right) \hat{\boldsymbol{\beta}}_{MBR} - \sum_{U-s} (\mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k)\right) \\ &= \mathrm{Var}_\xi \left(\left(\sum_{U-s} \mathbf{x}_k'\right) \hat{\boldsymbol{\beta}}_{MBR} - \sum_{U-s} \mathbf{x}_k' \boldsymbol{\beta} - \sum_{U-s} \varepsilon_k\right) \\ &= \left(\sum_{U-s} \mathbf{x}_k'\right) \mathrm{Var}_\xi(\hat{\boldsymbol{\beta}}_{MBR}) \left(\sum_{U-s} \mathbf{x}_k'\right)' + \sigma^2 \sum_{U-s} v_k. \end{aligned}$$

Now, again consider the relation given earlier in (2.57),

$$\hat{\boldsymbol{\beta}}_{MBR} = \mathbf{A}(diag(\delta_i))_{i=1}^{p}\mathbf{A}'\hat{\boldsymbol{\beta}}_{GLS}.$$

It gives

$$\mathrm{Var}_\xi(\hat{\boldsymbol{\beta}}_{MBR}) = \mathbf{A}(diag(\delta_i))_{i=1}^{p}\mathbf{A}'\mathrm{Var}_\xi(\hat{\boldsymbol{\beta}}_{GLS})\mathbf{A}(diag(\delta_i))_{i=1}^{p}\mathbf{A}'$$

with,

$$\begin{aligned}\mathrm{Var}_\xi(\hat{\boldsymbol{\beta}}_{GLS}) &= \sigma^2(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1} = \sigma^2(\mathbf{A}(diag\lambda_i)\mathbf{A}')^{-1} \\ &= \sigma^2\left(\mathbf{A}diag\left(\frac{1}{\lambda_i}\right)_{i=1}^{p}\mathbf{A}'\right)\end{aligned}$$

so,

$$\begin{aligned}\mathrm{Var}_\xi(\hat{\boldsymbol{\beta}}_{MBR}) &= \sigma^2\mathbf{A}diag(\delta_i)_{i=1}^{p}\mathbf{A}'(\mathbf{A}diag(\lambda_i)_{i=1}^{p}\mathbf{A}')^{-1}\mathbf{A}(diag(\delta_i))_{i=1}^{p}\mathbf{A}' \\ &= \sigma^2\mathbf{A}diag\left(\frac{{\delta_i}^2}{\lambda_{i,s}}\right)_{i=1}^{p}\mathbf{A}' \\ &= \sigma^2\mathbf{A}diag\left(\frac{\lambda_{i,s}}{(\lambda_{i,s}+\kappa)^2}\right)_{i=1}^{p}\mathbf{A}'\end{aligned}$$

thus, $\mathrm{Var}_\xi(\hat{t}_{MBR} - t_y)$ becomes,

$$\mathrm{Var}_\xi(\hat{t}_{MBR} - t_y) = \sigma^2\left(\sum_{U-s}\mathbf{x}_k'\right)\left(\mathbf{A}diag\left(\frac{\lambda_{i,s}}{(\lambda_{i,s}+\kappa)^2}\right)_{i=1}^{p}\mathbf{A}'\right)\left(\sum_{U-s}\mathbf{x}_k'\right)' + \sigma^2\sum_{\bar{s}}v_k$$

Hence, the ξ-MSE is given by,

$$\begin{aligned}E_\xi(\hat{t}_{MBR} - t_y)^2 &= \sigma^2\left(\sum_{U-s}\mathbf{x}_k'\right)\left(\mathbf{A}diag\left(\frac{\lambda_{i,s}}{(\lambda_{i,s}+\kappa)^2}\right)_{i=1}^{p}\mathbf{A}'\right)\left(\sum_{U-s}\mathbf{x}_k'\right)' \\ &\quad +\sigma^2\sum_{\bar{s}}v_k + \kappa^2\left(\sum_{U-s}\mathbf{x}_k'\right)\mathbf{A}diag\left(\frac{1}{\lambda_{i,s}+\kappa}\right)_{i=1}^{p}\mathbf{A}'\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{A} \\ &\quad diag\left(\frac{1}{\lambda_{i,s}+\kappa}\right)_{i=1}^{p}\mathbf{A}'\left(\sum_{U-s}\mathbf{x}_k'\right)'\end{aligned}$$

□

Model-assisted

Let us describe some of the properties of the model-assisted estimator $\hat{t}_{GREG,R}$ given in the relation (2.29) for the particular case $\mathbf{C}^{-1} = \kappa \mathbf{I}_p$.

Result 6. *The p-bias of* $\hat{t}_{GREG,R}$ *is*

$$Bias_p(\hat{t}_{GREG,R}) = -Trace\left(Cov_p\left(\sum_s \frac{\mathbf{x}_k}{\pi_k}, \hat{\boldsymbol{\beta}}_{\pi,R}\right)\right)$$

Proof.

$$\begin{aligned} \text{Bias}_p(\hat{t}_{GREG,R}) &= E_p(\hat{t}_{GREG,R} - t_y) \text{ where} \\ \hat{t}_{GREG,R} - t_y &= \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{x}'_k}{\pi_k} - \sum_U \mathbf{x}'_k\right)\hat{\boldsymbol{\beta}}_{\pi,R} - \sum_U y_k \end{aligned}$$

Recalling the design properties of the Horvitz-Thompson estimator, we know that the Horvitz-Thompson estimator of the population total $\sum_U y_k$ is design unbiased. i.e. $E_p(\sum_s \frac{y_k}{\pi_k}) = \sum_U y_k$. Hence,

$$\begin{aligned} E_p(\hat{t}_{GREG,R} - t_y) &= -E_p\left(\left(\sum_s \frac{\mathbf{x}'_k}{\pi_k} - \sum_U \mathbf{x}'_k\right)\hat{\boldsymbol{\beta}}_{\pi,R}\right) \\ &= -\text{Trace}\left(\text{Cov}_p\left(\sum_s \frac{\mathbf{x}_k}{\pi_k}, \hat{\boldsymbol{\beta}}_{\pi,R}\right)\right) \end{aligned}$$

□

Result 7. *The* ξ*-bias of* $\hat{t}_{GREG,R}$ *is given by,*

$$Bias_\xi(\hat{t}_{GREG,R}) = -\kappa\left(\sum_U \mathbf{x}'_k - \sum_s \frac{\mathbf{x}'_k}{\pi_k}\right)\left(\mathbf{X}'_s\mathbf{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s + \kappa\mathbf{I}_p\right)^{-1}\boldsymbol{\beta}$$

Proof.

$$\begin{aligned}
\hat{t}_{GREG,R} - t_y &= \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{x}'_k}{\pi_k} - \sum_U \mathbf{x}'_k \right) \hat{\boldsymbol{\beta}}_{\pi,R} - \sum_U y_k \\
\hat{t}_{GREG,R} - t_y &= \left(\sum_s \frac{y_k}{\pi_k} - \sum_s \frac{\mathbf{x}'_k}{\pi_k} \hat{\boldsymbol{\beta}}_{\pi,R} \right) + \left(\sum_U \mathbf{x}'_k \hat{\boldsymbol{\beta}}_{\pi,R} - \sum_U y_k \right) \\
E_\xi(\hat{t}_{GREG,R} - t_y) &= E_\xi \left(\sum_s \frac{\mathbf{x}'_k \boldsymbol{\beta} + \varepsilon_k - \mathbf{x}'_k \hat{\boldsymbol{\beta}}_{\pi,R}}{\pi_k} \right) + E_\xi \left(\sum_U \mathbf{x}'_k \hat{\boldsymbol{\beta}}_{\pi,R} - \sum_U (\mathbf{x}'_k \boldsymbol{\beta} + \varepsilon_k) \right) \\
E_\xi(\hat{t}_{GREG,R} - t_y) &= E_\xi \left(\sum_U \mathbf{x}'_k \hat{\boldsymbol{\beta}}_{\pi,R} - \sum_U \mathbf{x}'_k \boldsymbol{\beta} \right) - E_\xi \left(\sum_s \frac{\mathbf{x}'_k \hat{\boldsymbol{\beta}}_{\pi,R} - \mathbf{x}'_k \boldsymbol{\beta}}{\pi_k} \right)
\end{aligned}$$

so the bias of $\hat{t}_{GREG,R}$ takes the shape,

$$\begin{aligned}
Bias_\xi(\hat{t}_{GREG,R}) &= \left(\sum_U \mathbf{x}'_k \right) E_\xi(\hat{\boldsymbol{\beta}}_{\pi,R} - \boldsymbol{\beta}) - \left(\sum_s \frac{\mathbf{x}'_k}{\pi_k} \right) E_\xi(\hat{\boldsymbol{\beta}}_{\pi,R} - \boldsymbol{\beta}) \\
&= \left(\sum_U \mathbf{x}'_k \right) Bias_\xi(\hat{\boldsymbol{\beta}}_{\pi,R}) - \left(\sum_s \frac{\mathbf{x}'_k}{\pi_k} \right) Bias_\xi(\hat{\boldsymbol{\beta}}_{\pi,R}) \\
&= \left(\sum_U \mathbf{x}'_k - \sum_s \frac{\mathbf{x}'_k}{\pi_k} \right) Bias_\xi(\hat{\boldsymbol{\beta}}_{\pi,R})
\end{aligned}$$

where

$$\text{Bias}_\xi(\hat{\boldsymbol{\beta}}_{\pi,R}) = -\kappa \left(\mathbf{X}'_s \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s + \kappa \mathbf{I}_p \right)^{-1} \boldsymbol{\beta}.$$

So we have finally,

$$\text{Bias}_\xi(\hat{t}_{GREG,R}) = -\kappa \left(\sum_U \mathbf{x}'_k - \sum_s \frac{\mathbf{x}'_k}{\pi_k} \right) \left(\mathbf{X}'_s \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s + \kappa \mathbf{I}_p \right)^{-1} \boldsymbol{\beta}.$$

□

Result 8. *Suppose that* $\hat{\boldsymbol{\beta}}_{\pi,R} - \hat{\boldsymbol{\beta}}_\kappa = o_p(1)$ *where* $\hat{\boldsymbol{\beta}}_\kappa = \left(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X} + \kappa \mathbf{I}_p \right)^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y}$. *The asymptotic variance of* $\hat{t}_{GREG,R}$,

$$\begin{aligned}
AV_p(\hat{t}_{MA,R}) &= Var_p \left(\sum_s \frac{y_k - \mathbf{x}'_k \hat{\boldsymbol{\beta}}_\pi}{\kappa} \right) \\
&= \sum_s \sum_s \Delta_{kl} \left(\frac{y_k - \mathbf{x}'_k \hat{\boldsymbol{\beta}}_\kappa}{\pi_k} \right) \left(\frac{y_l - \mathbf{x}'_l \hat{\boldsymbol{\beta}}_\kappa}{\pi_l} \right),
\end{aligned}$$

Proof. The proof of this result is inspired by the *variance* of Horvitz-Thompson estimator for the population total t_y. Since, the *asymptotic variance* is,

$$E_p(\hat{t}_{GREG,R} - t_y)^2 = Var_p(\hat{t}_{GREG,R} - t_y) + (Bias_p(\hat{t}_{GREG,R}))^2$$

Since $\frac{1}{N}(\hat{t}_{x\pi} - t_x) = O_p(\frac{1}{\sqrt{n}}) = o_p(1)$ (i.e. this converges in probability towards zero). Similarly, $\hat{\boldsymbol{\beta}}_{\pi,R} - \boldsymbol{\beta} = o_p(1)$. Hence, we have,

$$\begin{aligned}
\frac{1}{N}(\hat{t}_{GREG,R} - t_y) &= \frac{1}{N}(\hat{t}_{y\pi} - t_y) - \overbrace{\frac{1}{N}(\hat{t}_{x\pi} - t_x)}^{O_p\left(\frac{1}{\sqrt{n}}\right)} \overbrace{(\hat{\boldsymbol{\beta}}_{\pi,R} - \hat{\boldsymbol{\beta}}_{\kappa})}^{o_p(1)} - \overbrace{\frac{1}{N}(\hat{t}_{x\pi} - t_x)}^{O_p\left(\frac{1}{\sqrt{n}}\right)} \hat{\boldsymbol{\beta}}_{\kappa} \\
\frac{1}{N}(\hat{t}_{GREG,R} - t_y) &= \frac{1}{N}(\hat{t}_{y\pi} - \hat{t}_{x\pi}\hat{\boldsymbol{\beta}}_{\kappa}) - \frac{1}{N}(t_y - t_x\hat{\boldsymbol{\beta}}_{\kappa}) + o_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

$$\frac{1}{N}(\hat{t}_{GREG,R} - t_y) = \frac{1}{N}(\hat{t}_{y\pi} - t_y) - \frac{1}{N}(\hat{t}_{x\pi} - t_x)(\hat{\boldsymbol{\beta}}_{\kappa}) + o_p\left(\frac{1}{\sqrt{n}}\right)$$

and the *asymptotic* p-bias for the $\hat{\boldsymbol{\beta}}_{MA,R}$ becomes as follows,

$$\text{Bias}_p(\hat{t}_{GREG,R}) \simeq 0$$

So, the *asymptotic variance* becomes,

$$\begin{aligned}
E_p(\hat{t}_{GREG,R} - t_y)^2 &= Var_p(\hat{t}_{GREG,R} - t_y) \\
\text{AV}_p(\hat{t}_{GREG,R}) &= \sum_U \sum_U \Delta_{kl} \left(\frac{y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}_{\kappa}}{\pi_k}\right)\left(\frac{y_l - \mathbf{x}_l'\hat{\boldsymbol{\beta}}_{\kappa}}{\pi_l}\right).
\end{aligned}$$

□

2.4 Application to the Mediametrie Data

We verify in this section the suggested estimators on Médiamétrie data. The application here is about panel Mediamat data of 6 to 13 September 2010. The population consists of 9750 individuals aged of more than four year old watching a channel during this time period. The available information on sample and population are at two levels:

1. The variable describing the INSEE Region and Household: the agglomeration size of residence, age and socio-professional category of the Household Head, age and activity of the housekeeper/resident, number of persons per household, presence of children of less than 15 year old, number of televisions, mode/source of reception (satellite, ADSL cable, TNT, Analogical hertzien), contracted to CanalSat, contracted to Canal+, possession of mini-computer, access to Internet.

2. The variables describing the individuals: sex, age, socio-professional status, type of Employement.

The variables of interest are the Listening Duration of individuals by channel and by day.

We have performed a small simulation study to verify the performance of the principal component regression estimator and ridge estimator. We have considered the sample of 6-13 September 2010 as our study population from which we selected 1000 random samples without replacement of size 500. The considered variable of interest is the Listening Duration on a certain channel on Monday 13 of September considering as auxiliary variables the age, the socio-professional category, the geographic region, the sex and the Listening Duration of the same channel during the previous Monday. The $\boldsymbol{X}$ matrix is built of 19 columns and is ill-conditioned. The GREG estimator does not always work because the $\mathbf{X}_\mathbf{s}'\mathbf{\Pi}_s\mathbf{X}_\mathbf{s}$ matrix has the minimum eigen-value λ_{min} equal to zero for many samples.

We have therefore calculated ridge estimators on 1000 samples through the relative-

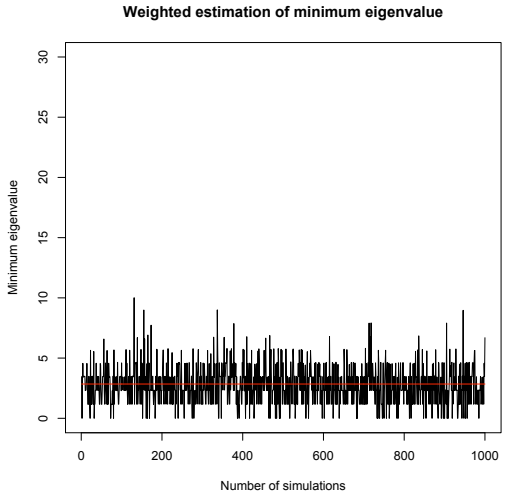


Figure 2.3: Minimum Eigenvalues of $\mathbf{X_s'\Pi}_s\mathbf{X_s}$, 1000 simulations

bias and the relation between the MSE of the proposed estimators and that of the Horvitz-Thompson estimator which does not take into account the auxiliary information. We trace in figure 2.4, the ratio between the MSE of $\hat{t}_{ridge}$ and the MSE of $\hat{t}_{HT}$ for many values of κ and for 10 repetitions of the simulation study. We can remark that for small values of k the gain is important (65%), while for large κ, the $\hat{t}_{ridge}$ estimator approaches to $\hat{t}_{HT}$.

2.5 Conclusion and extensions

In this section, we have undertaken an overview of the applications of ridge-type estimators in survey sampling theory. The paper of Bardsley and Chambers (1984) did not receive much attention at the beginning but during the last years, we remark an increasing interest on this subject. This is mostly due to the fact that nowadays, we face bulk of information. This kind of issue is now more often encountered in practice than before.

Broadly speaking, the ridge technique means solving an optimization problem under a quadratic constraint. We have presented in this section the application of

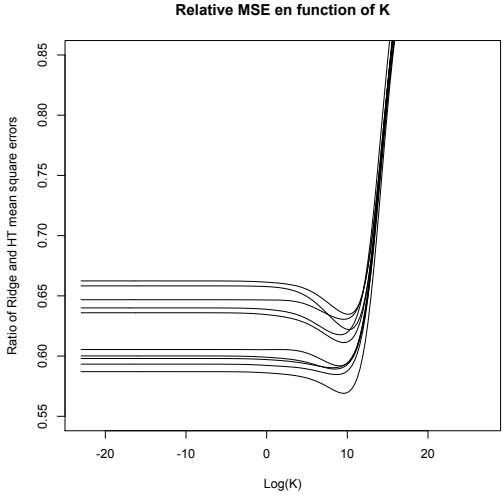


Figure 2.4: Ratio of the mean square errors between the ridge and the Horvitz-Thompson estimators

this principle in a model-based framework as well as in a design-based or calibration framework. It is established that in both approaches, weights are derived as solution of optimization problems under the constraints given by $\mathbf{w}_s'\mathbf{X}_s = \mathbf{1}_U'\mathbf{X}_U$. These constraints prove to be too restrictive if their number is too large leading to instability of weights. Using a quadratic constraint function leads to ridge-type weights that are more stable than those obtained under a linear constraint. The weighted estimators obtained in this way do not estimate exactly the finite population totals t_{X_j} of the calibration variables, but they are as close as possible to t_{X_j} while satisfying fixed weight range restrictions.

To use this class of estimators, two practical issues should be treated carefully. The first one is the computation of the penalty parameter. Several algorithms have been suggested in the literature such as the ridge trace (Bardsley and Chambers, 1984), the Fischer scoring algorithm (Guggemos and Tillé, 2010) or the bisection algorithm (Beaumont and Bocci, 2008) used before by Chen *et al.* (2002) for computing range restricted weights for a given tolerance matrix. Beaumont and Bocci (2008) show that it is better to fix the cost matrix $\mathbf{C}$ and to determine next the

ridge estimator than to fix first the tolerance matrix (Chen *et al.*, 2002). Nevertheless, it would be interesting to have a comparison between all these algorithms.

There is another important point that we would like to stress. All the papers dealing with ridge-type estimators in survey sampling give few details about the standardization of the auxiliary variables if any has been done. Or, as mentioned at the end of section 2.2, it is important to know what kind of standardization is used since different methods lead to different ridge estimators. The cost matrix used in the objective functions from the optimization problems **(P3)** and **(P5)** may be interpreted as a standardization matrix.

Finally, some other alternative methods for dealing with huge data sets must be investigated. We mention here the lasso methods which consist in considering a constraint with the absolute value instead of the euclidean norm. We are not aware of the existence of such application in survey sampling. The regression on principal component analysis is another interesting alternative. This method consists in considering the principal components of $\mathbf{X}'\mathbf{X}$ which reduce the number of auxiliary variables while keeping maximum of information. For huge survey data, in the next chapter we suggest calibration on the set of these new variables which is in general of much smaller dimension than the initial one.

Chapter 3

Dimension Reduction of Survey Data using Principal Components Analysis

The chapter 3 is divided into three main parts. In section 3.1, we give some general overview of the PCA technique as given by Pearson (1901), Hotelling (1933) and Dunteman (1989). The construction of principal components is described in section 3.1.1 and some general methods (Jolliffe, 2002) for selecting the number of principal components are also discussed. In section 3.1.2, the different uses of principal components in regression analysis is mentioned (Dunteman, 1989) and some underlying risks concerning the use of PCA in regression analysis are also given (Izenman, 2008). In section 3.2, we discussed in detail the use of principal components in survey sampling. Model-based and model-assisted estimators (section 3.2.1 and section 3.2.2) are given and the calibration using principal components is also discussed. Different type of calibrated estimators using principal components are proposed such as calibration on second moment (section 3.2.4), partially calibrated principal component estimator (section 3.2.5) and the estimated principal component estimator (section 3.2.6). Finally, a detailed simulation study is

conducted on Mediametrie data and the results are presented in the tabular and graphical form in the end.

3.1 General Background on PCA

Principal components analysis (PCA) is arguably one of the best multivariate technique in which we can systematically reduce a large number of dependent variables to a relatively more consistent or coherent smaller set of variables (Dunteman, 1989). Pearson (1901) introduced principal components analysis which was further explored and extended by Hotelling (1933). The basic idea behind PCA generally remains the reduction in the dimension of the data set. The method had restricted use until the modern age computers came into existence which made the computation easier to a great extent.

Certain reasons can be given to defend the use of PCA in the large data sets but most frequent and possibly the healthiest one is that we can save bulk of the cost and time if the given data set has large number of intercorrelated variables. PCA also guaranty the retention of the maximum possible variation of the initial data set into the new data set of reduced dimension. This is due to the fact that each of the new variable called principal component is in fact the linear combination of all original variables. Aftermath of this process results into a new set of variables, the principal components which are uncorrelated among them and ordered such that the first few principal components retain most of the information in terms of the variation available in the original variables. Johnson and Wichern (2002) describe data reduction and interpretation as the general objectives of PCA. They also declared that the direct concern of a PCA is to explain the variance-covariance structure via a few linear combination of the original variables. Talking in the same way, we can say that most of the statistical goals and objectives are subject to finding the relationship among the different individual points in a particular data set.

The property of uncorrelated principal components is important in a way that

it eliminates the interdependence available in the original data set. Certain type of relationships are not detected by any ordinary analysis (means, analysis of variance etc) and their interpretations are revealed only by the analysis of principal components.

An example also discussed by Johnson and Wichern (2002) advocating the use of PCA, is a study in which investigating the reaction of cancer patients to radiotherapy was investigated: 6 reaction variables for 98 patients were measured. All of the 6 reaction variables are difficult to interpret for the observations at the same time, so a rather easier and simpler measure of patients' response was of interest. PCA seems to be the suitable technique for the construction of a simple measure of patient response to radiotherapy, and still containing maximum of the available sample information.

3.1.1 Construction of Principal Components

The method for constructing the principal components is very simple. Algebraically speaking, principal components are linear combinations of the all p variables $\mathbf{X}_1, ..., \mathbf{X}_p$. If the covariance or correlation structure between the p variables and their variances are of interest then for large p, it will often not be sufficient to examine just the p variances or the $\frac{p(p-1)}{2}$ correlations or covariances. Another problem can be that if p is large, then it is difficult to construct covariance structure. An alternative way to examine the information contained by the p variables is to search for a less than p variables such that most of the information in the variances and correlations or covariances are preserved. Despite of the fact that PCA reduces dimension of the variables, it does not ignore the covariances or correlation but it concentrate the variances in a rather fewer number of variables which are the principal components. Geometrically speaking, the principal components are representing the selection of the new coordinate system obtained via rotation of the original system with $\mathbf{X}_1, ..., \mathbf{X}_p$ as the coordinate axes (Johnson and Wichern (2002)). The new system of axes shows the directions with maximum

variability and demonstrates a rather simple and easily interpretable description of the covariance structure. That is to say that dimension reduction via PCA does not endanger the potential information due to the fact that each of the principal component is the linear combination of all p auxiliary random variables. Qian *et al* (1994) stress on the importance of the measurement units or a particular coordinate system and say that the only case when principal components are meaningful is when all the variables are measured in the same units. If it is not the case, one should perform principal component analysis on the standardized observations of the variables. The principal components obtained from covariance matrix and correlation matrix differ from each other as illustrated by Johnson and Wichern (2002). The variables should be standardized if they are measured in different units. This thing helps in the construction of the principal components as the covariance and correlation matrices of the standardized variables are equal.

Suppose the correlation matrix $\mathbf{\Sigma}$ for the vector $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_p)$ with the variances on the diagonal and covariances between two different variables on the off-diagonal of the matrix. We consider the eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_r$ corresponding to the largest eigenvalues $\lambda_1 \geq \ldots \geq \lambda_r > 0$ of the matrix $\frac{1}{N}\mathbf{X}'\mathbf{X}$. Now the first principal component can be defined as,

$$\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1 = a_{11}\mathbf{X}_1 + a_{21}\mathbf{X}_2 + ... + a_{p1}\mathbf{X}_p$$

where $\mathbf{a}_1$ is the eigenvector corresponding to the largest eigenvalue λ_1 of $\mathbf{\Sigma}$. Similarly the second principal component is given as,

$$\mathbf{z}_2 = \mathbf{X}\mathbf{a}_2$$

where $\mathbf{A}_2$ is the eigenvector corresponding to the second largest eigenvalue λ_2 of $\mathbf{\Sigma}$. More generally the ith principal component can be written as,

$$\mathbf{z}_i = \mathbf{X}\mathbf{a}_i,\ i = 1, \ldots, p$$

The eigenvalues are in fact the variances of the principal components. That is, $var(\mathbf{z}_i) = \lambda_i$. Since $var(\mathbf{z}_i) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i$ and $cov(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_j,\ i \neq j$ where

$i, j = 1, ..., p$, so the objective is to find those uncorrelated linear combination $\mathbf{z}_1, ..., \mathbf{z}_p$ who maximize the relevant variances. Thus the first principal component PC1 maximizes $var(\mathbf{z}_1) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$ hence PC1 has maximum variability. This fact of maximizing the variation may easily be exploited by just multiplying any constant to the $var(\mathbf{z}_1)$ and therefore increasing the variance. So to eliminate this risk, we, for the sake of convenience, restrict the attention to coefficient vectors $\mathbf{v}_i$ of unit length. That is, we impose an additional condition that coefficient vectors are of unit length. Our objective function for the construction of the principal components therefore becomes,

$$\begin{aligned}
PC_1 = \mathbf{z}_1 &= \text{Linear combination } \mathbf{X}\mathbf{a}_1 \text{ that maximizes } var(\mathbf{z}_1) \\
&\quad \text{such that } \mathbf{a}_1'\mathbf{a}_1 = 1 \\
PC_2 = \mathbf{z}_2 &= \text{Linear combination } \mathbf{X}\mathbf{a}_2 \text{ that maximizes } var(\mathbf{z}_2) \\
&\quad \text{such that } \mathbf{a}_2'\mathbf{a}_2 = 1 \text{ and } cov(\mathbf{z}_1, \mathbf{z}_2) = 0 \\
&\quad . \\
&\quad . \\
&\quad . \\
PC_i = \mathbf{z}_i &= \text{Linear combination } \mathbf{X}\mathbf{a}_i \text{ that maximizes } var(\mathbf{z}_i) \\
&\quad \text{such that } \mathbf{a}_i'\mathbf{a}_i = 1 \text{ and } cov(\mathbf{z}_i, \mathbf{z}_j) = 0, \text{ for } j < i
\end{aligned}$$

Johnson and Wichern (2002) also consider a special case of the covariance matrix such that it is equal to,

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & \dots & \mathbf{0} \\ \dots & \sigma_i & \dots \\ \mathbf{0} & \dots & \sigma_p \end{pmatrix},$$

As apparent from its structure, the off-diagonal elements are zero, so the variables are already uncorrelated from each other. Thus, it would be a useless exercise to construct principal components which are the uncorrelated linear combinations of the auxiliary variables. Even the standardized variables will not change the situation as we will have $\rho = 1$ for the variables meaning that eigenvalues are

1 and thus the principal components will be equal to the original standardized variables and we gain nothing. So, in order to construct the principal components for a covariance matrix of standardized variables, they must be correlated among themselves.

Another geometrical implication of first PC given by Dunteman (1989) is that it is the line of closest fit to the N observations in the p dimensional variable space. Saying otherwise, sum of the squared distances of the N observations from the line in the variable space representing the first PC is minimized by this fit. A plane of closest fit to the cluster of points in the p dimensional variable space is defined by the first two principal components. The second PC, equivalently is a line of closest fit to the residuals from the first PC. A three dimensional plane, called a hyperplane, of closest fit is defined by the first three principal components and so on. The total number of principal components for p random variable can be less than p if there exist any dependencies between the variables and maximum number of principal components remains equal to p.
There are many methods regarding the decision of what number of the principal components should be selected. This number is usually denoted by r and is much smaller than the total number of variables p. A most frequent method used for this purpose is to choose those first r variables which comprise the maximum percentage of variation. Normally, this percentage ranges from 80% or 90%. That is, we select those first r which cover at least 80% or 90% of the total variation. This threshold is set up depending upon the sensibility of the problem encountered. If we have a very large p and many of the first principal components represent very few variation then setting a lower percentage for the selection of the r will be appropriate say $< 80\%$. On the other hand if first few (on or two) principal components assume bulk of the variation say more than 90%, then a suitable percentage for chosing r should be well more than 90% so that we can get maximum variation intact (for more discussion Jolliffe (2002)). We shall use this typical method for selection of principal components. This method serves equally well

for both situations whether the principal components are constructed using a covariance or correlation matrix.

3.1.2 Principal Component Regression

We can see the traces of using principal components in regression analysis back into Kendall (1957) and Hotelling (1957). The idea was to orthogonalize the regression problem by replacing the initial regressors variables by their principal components. By doing this, computation was made more stable and rather easier.

Dunteman (1989) suggested that there may exist several ways of using principal components in regression analysis. The variables with high correlation among them can be replaced by their independent uncorrelated principal components in the regression analysis. Detection of multicollinearity among the auxiliary variables and selection of a subset of auxiliary variables for the regression analysis can be the other potential uses of principal component analysis. The idea of using principal component regression emerges from the classical problem of multicollinearity with the usual least squares estimators. Since the principal components are uncorrelated with no multicollinearities, their use in the regression in place of the original auxiliary variables will make the regression calculation simple. The use of all principal components in the regression will result in the model equivalent to the least squares, so the variance inflation due to the multicollinearities is not removed. Jolliffe (1982) points out about a misconception about the rule of deciding the particular principal components into the analysis. The method initially developed by judging the principal components in the similar fashion as original predictor variables to decide whether they should be included into the regression analysis or not. However, the attention certainly shifted to the rule of inclusion of principal components based on the large variance and rejecting the principal components with small variances. The regression estimators using principal components are biased, but they can prove useful in large reduction of variances due to the multicollinearity in the regression coefficient estimators.

An example of using PCR is in the field of chemometrics where the interest may be the calibration of the fat concentration when the number of variables p may be much greater than the number of individuals N. Reduction of regression dimension can be done using PCR by deleting those variables that contribute to the collinearity (Martens and Naes, 1989).

The common practice pre-assumes that if the selection of the principal components is based on the variance (i.e. those principal components with smaller variances are deleted) then we loose marginal estimation power in regression analysis. This however, is not necessarily true every time as there may be the cases when the inclusion of the principal components should also depend on the relationship between the dependent variable. Also examination is important as any component with smaller variance may belong to the auxiliary variable (Jeffers (1967) and Hawkins (1973)). Certain examples advocating the inclusion of low-variance principal components can be found in the literature (Smith and Campbell (1980), Kung and Sharif (1980)). The deletion of the small variance PC's should be avoided until the negligible correlation of these PC's with the study variable $\mathbf{Y}$ are confirmed (Jolliffe, 2002). Izenman (2008) states that there are certain risks of PCR to collapse heavily. The first r PC's $\mathbf{z}_1, \ldots, \mathbf{z}_r$ used in the regression procedure have no apparent reason to be strongly correlated with the variable of interest $\mathbf{Y}$. On the contrary, the last few PC's (Jolliffe, 1982) or some times only the last PC (Hadi and Ling, 1998) may be strongly correlated with $\mathbf{Y}$ but are normally dropped from the analysis. So, the use of PCR should be with some caution and the contribution of each PC in the regression sum of squares should also be considered in addition to the consideration of the variance decomposition.

In ridge regression the trade-off between the bias and variance is handled by the optimal choice of the ridge parameter whereas, the compromise between bias and variance in principal component regression is achieved by selection of the right number of principal components to be used in the regression procedure (Jolliffe,

2002).

$$\xi : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{Z}\tilde{\boldsymbol{\eta}} + \boldsymbol{\varepsilon} \tag{3.1}$$

where $\mathbf{Z} = \mathbf{z}_1, \ldots, \mathbf{z}_p = \mathbf{X} \cdot \mathbf{A}$ with $\mathbf{A} = \mathbf{a}_1, \ldots, \mathbf{a}_p$. The PCR consists in reducing the space spanned by the columns of $\mathbf{X}$ and consider the regression model ξ' over the reduced space. Let the first r principal components denoted by

$$\mathbf{Z}_r = (\mathbf{z}_1, \ldots, \mathbf{z}_r). \tag{3.2}$$

The new model consists in regressing $\mathbf{y}$ on $\mathbf{Z}_r$

$$\xi' : \quad \mathbf{y} = \mathbf{Z}_r\boldsymbol{\eta} + \boldsymbol{\varepsilon}_r, \tag{3.3}$$

where $\boldsymbol{\varepsilon}_r$ is the restriction of $\boldsymbol{\varepsilon}$ The estimation of $\boldsymbol{\eta}$ is done by least squares,

$$\hat{\boldsymbol{\eta}} = (\mathbf{Z}_r'\mathbf{Z}_r)^{-1}\mathbf{Z}_r'\mathbf{y} \tag{3.4}$$

and the estimator of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}}_{PC} = (\mathbf{a}_1, \ldots, \mathbf{a}_r)\hat{\boldsymbol{\eta}}$. Let $\mathbf{A}_r = \mathbf{a}_1, \ldots, \mathbf{a}_r$, then

$$\hat{\boldsymbol{\beta}}_{PC} = \mathbf{A}_r\hat{\boldsymbol{\eta}} \quad \text{and} \tag{3.5}$$

$$\hat{\boldsymbol{\eta}} = \mathbf{A}_r'\hat{\boldsymbol{\beta}}_{PC}. \tag{3.6}$$

Gunst and Mason (1977) expressed the PC estimator $\hat{\boldsymbol{\beta}}_{PC}$ in function of the eigenvalues λ_i, $i = 1, \ldots, r$ as,

$$\hat{\boldsymbol{\beta}}_{PC} = \frac{1}{N}\sum_{i=1}^{r}\frac{\mathbf{a}_i'\mathbf{X}'\mathbf{y}\mathbf{a}_i}{\lambda_i}$$

or equivalently,

$$\hat{\boldsymbol{\beta}}_{PC} = \hat{\boldsymbol{\beta}}_{OLS} - \frac{1}{N}\sum_{i=r+1}^{p}\frac{\mathbf{a}_i'\mathbf{X}'\mathbf{y}\mathbf{a}_i}{\lambda_i}. \tag{3.7}$$

The expected value of $\hat{\boldsymbol{\beta}}_{PC}$ under the model ξ is given by,

$$\begin{aligned} E_\xi(\hat{\boldsymbol{\beta}}_{PC}) &= E_\xi(\hat{\boldsymbol{\beta}}_{OLS}) - \frac{1}{N}\sum_{i=r+1}^{p}\frac{\mathbf{a}_i'\mathbf{X}'E_\xi(\mathbf{y})\mathbf{a}_i}{\lambda_i} \\ &= \boldsymbol{\beta} - \frac{1}{N}\sum_{i=r+1}^{p}\frac{\mathbf{a}_i'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\mathbf{a}_i}{\lambda_i} \\ &= \boldsymbol{\beta} - \frac{1}{N}\sum_{i=r+1}^{p}\mathbf{a}_i'\boldsymbol{\beta}\mathbf{a}_i. \end{aligned}$$

The bias is given by,

$$E_\xi(\hat{\boldsymbol{\beta}}_{PC}) - \boldsymbol{\beta} = -\frac{1}{N}\sum_{i=r+1}^{p} \mathbf{a}_i'\boldsymbol{\beta}\mathbf{a}_i.$$

The bias of $\hat{\boldsymbol{\beta}}_{PC}$ involves unknown parameter $\boldsymbol{\beta}$. Substantial reduction in variance and mean square error is gained in presence of serious multicollinearity and the introduction of the small bias (Gunst and Mason (1977) and Jolliffe (2002)). Notably, $\hat{\boldsymbol{\beta}}_{PC}$ is the realization of $\hat{\boldsymbol{\beta}}_{OLS}$ (equation 2.4) over r terms. The ξ-mean squared error (MSE) of $\hat{\boldsymbol{\beta}}_{PC}$ is given by,

$$MSE(\hat{\boldsymbol{\beta}}_{PC}) = E_\xi\left[(\hat{\boldsymbol{\beta}}_{PC} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_{PC} - \boldsymbol{\beta})\right]$$

and is the trace of the variance-covariance matrix

$$MMSE(\hat{\boldsymbol{\beta}}_{PC}) = E_\xi\left[(\hat{\boldsymbol{\beta}}_{PC} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{PC} - \boldsymbol{\beta})'\right].$$

$MSE(\hat{\boldsymbol{\beta}}_{PC})$ becomes,

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}_{PC}) &= \boldsymbol{\sigma}^2\sum_{i=1}^{r}\frac{1}{\lambda_i} + \sum_{i=r+1}^{p}\left(\mathbf{a}_i'\boldsymbol{\beta}\right)^2 \\ &= Trace\left(Var_\xi(\hat{\boldsymbol{\beta}}_{PC})\right) + (Bias_\xi(\hat{\boldsymbol{\beta}}_{PC}))'(Bias_\xi(\hat{\boldsymbol{\beta}}_{PC})) \\ &= \mathcal{C}(r) + \mathcal{D}(r) \end{aligned} \tag{3.8}$$

Jolliffe (2002) provides criteria for choosing the number of principal components.

Result 9. *(Gunst and Mason, 1977)*
The MMSE of $\hat{\boldsymbol{\beta}}_{PC}$ under model ξ is smaller than that of $\hat{\boldsymbol{\beta}}_{OLS}$, that is,

$$MMSE(\hat{\boldsymbol{\beta}}_{OLS}) - MMSE(\hat{\boldsymbol{\beta}}_{PC})$$

is a positive-definite matrix if

$$\sum_{i=r+1}^{p}\left(\frac{\lambda_i}{\boldsymbol{\sigma}^2}\left(\mathbf{a}_i'\boldsymbol{\beta}\right)^2\right) \leq 1.$$

3.2 Principal Components Regression in Survey Sampling

We suppose without loss of generality that the auxiliary variables are standardized, namely $\mathbf{1}_U'\mathbf{X}_i = 0$ and $\mathbf{X}_i'\mathbf{X}_i = 1$ for all $i = 1, \ldots, p$ and $\mathbf{1}_U'$ is the N-dimensional vector of ones. We suggest a new class of GREG type estimators using principal component regression (PCR).

3.2.1 Model-assisted approach

Let $\mathbf{z}_i = \mathbf{X}\mathbf{a}_i = (z_{ki})_{k\in U}$ for $i = 1, \ldots, p$ with $\tilde{\boldsymbol{z}}_k' = (z_{k1}, \ldots, z_{kr})$ be the vector containing the values of the first r principal components for the i-th individual and $\mathbf{Z}_r = (\mathbf{z}_1, \ldots, \mathbf{z}_r) = (\tilde{\boldsymbol{z}}_k')_{k=1}^N$ given by (3.2). The estimator $\hat{\boldsymbol{\eta}}$ given by (3.4) can not be calculated since it contains the unknown population vector $\mathbf{y}$. The design-based estimator of $\hat{\boldsymbol{\eta}}$ is given by

$$\hat{\boldsymbol{\eta}}_\pi = (\mathbf{Z}_{r,s}'\boldsymbol{\Pi}_s^{-1}\mathbf{Z}_{r,s})^{-1}\mathbf{Z}_{r,s}'\boldsymbol{\Pi}_s^{-1}\mathbf{y}_s \tag{3.9}$$

where $\mathbf{Z}_{r,s}$ is the restriction of $\mathbf{Z}_r$ on the sample s, namely $\mathbf{Z}_{r,s} = (\tilde{\boldsymbol{z}}_k')_{k\in s}$ and $\boldsymbol{\Pi}_s = diag(\pi_k)_{k\in s}$ .We suggest to estimate the total t_y by

$$\hat{t}_{PC} = \hat{t}_{y,\pi} - \left(\hat{t}_{z,\pi} - t_z\right)' \hat{\boldsymbol{\eta}}_\pi \tag{3.10}$$

where $\hat{t}_{z,\pi} = \sum_s \dfrac{\tilde{\boldsymbol{z}}_k}{\pi_k}$ is the Horvitz-Thompson estimator of $t_z = \sum_U \tilde{\boldsymbol{z}}_k$. For standardized variables $\mathbf{X}_i$, $i = 1, \ldots, p$ we have that the principal components are of zero mean and this fact implies that $t_z = 0$. As a consequence, the estimator given by (3.10) becomes

$$\hat{t}_{PC} = \hat{t}_{y,\pi} - \hat{t}_{z,\pi}'\hat{\boldsymbol{\eta}}_\pi = \sum_s \frac{y_k - \tilde{\boldsymbol{z}}_k'\hat{\boldsymbol{\eta}}_\pi}{\pi_k} \tag{3.11}$$

which is the Horvitz-Thompson estimator for the sample fit residuals $y_k - \tilde{\boldsymbol{z}}_k'\hat{\boldsymbol{\eta}}_\pi$. We can remark that $\hat{t}_{PC}$ is a GREG type estimator for the vector of the first r principal components $\mathbf{Z}_r$ of $\mathbf{X}$. By its construction, we achieve a reduction in dimension of $\mathbf{X}$ by retaining maximum information. Nevertheless, this method

demands knowing $\mathbf{X}$ over the whole population in order to derive the eigenvalues and eigenvectors.

We have that $\mathbf{Z}_r = \mathbf{X}\mathbf{A}_r$ which implies that the restriction on the sample s is $\mathbf{Z}_{r,s} = \mathbf{X}_s\mathbf{A}_r$ where $\mathbf{A}_r = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$. The design-based estimator of $\hat{\boldsymbol{\beta}}_{PC}$ is

$$\hat{\boldsymbol{\beta}}_{PC,\pi} = \mathbf{A}_r\hat{\boldsymbol{\eta}}_\pi,$$

implying that,

$$\hat{\boldsymbol{\eta}}_\pi = \mathbf{A}_r\hat{\boldsymbol{\beta}}_{PC,\pi}.$$

So, the estimator $\hat{t}_{PC}$ given by (3.10) can be written in function of $\mathbf{X}$ as,

$$\hat{t}_{PC} = \hat{t}_{y,\pi} - \left(\hat{t}_{x,\pi} - t_x\right)' \hat{\boldsymbol{\beta}}_{PC,\pi}. \tag{3.12}$$

3.2.2 Properties of $\hat{t}_{PC}$ under the model and the sampling design

We study in this section, statistical properties of $\hat{t}_{PC}$ under the model ξ and the sampling design $p(.)$.

Result 10. *1. The bias under the model of the principal component estimator $\hat{t}_{PC}$ is given by,*

$$E_\xi(\hat{t}_{PC} - t_y) = (\mathbf{A}'_r - \mathbf{A}')\boldsymbol{\beta}$$

where $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$.

2. The principal component estimator $\hat{t}_{PC}$ of t_y is unbiased under (ξ, p), namely

$$E_p E_\xi(\hat{t}_{PC} - t_y) = 0$$

Proof. 1. Since, $\hat{t}_{PC} = \hat{t}_{y,\pi} - \hat{t}'_{z,\pi}\hat{\boldsymbol{\eta}}_\pi$, we get,

$$\begin{aligned}
\hat{t}_{PC} - t_y &= \hat{t}_{y,\pi} - \hat{t}'_{z,\pi}\hat{\boldsymbol{\eta}}_\pi - t_y \\
&= \sum_s \frac{y_k}{\pi_k} - \sum_s \frac{\tilde{\boldsymbol{z}}'_k\hat{\boldsymbol{\eta}}_\pi}{\pi_k} - \sum_U y_k \\
&= \sum_s \frac{\tilde{\boldsymbol{z}}'_k\boldsymbol{\eta} + \varepsilon_k}{\pi_k} - \sum_s \frac{\tilde{\boldsymbol{z}}'_k\hat{\boldsymbol{\eta}}_\pi}{\pi_k} - \sum_U (\tilde{\boldsymbol{z}}'_k\boldsymbol{\eta} + \varepsilon_k) \\
&= -\sum_s \frac{\tilde{\boldsymbol{z}}'_k}{\pi_k}(\hat{\boldsymbol{\eta}}_\pi - \boldsymbol{\eta}) - \underbrace{\sum_U \tilde{\boldsymbol{z}}'_k\boldsymbol{\eta}}_{0} - \sum_U \varepsilon_k + \sum_s \frac{\varepsilon_k}{\pi_k} \\
\hat{t}_{PC} - t_y &= -\sum_s \frac{\tilde{\boldsymbol{z}}'_k}{\pi_k}(\hat{\boldsymbol{\eta}}_\pi - \boldsymbol{\eta}) - \sum_U \varepsilon_k + \sum_s \frac{\varepsilon_k}{\pi_k}, \qquad (3.13)
\end{aligned}$$

and applying ξ-expectation, we get,

$$E_\xi(\hat{t}_{PC} - t_y) = -\left(\sum_s \frac{\tilde{\boldsymbol{z}}'_k}{\pi_k}\right) E_\xi(\hat{\boldsymbol{\eta}}_\pi - \boldsymbol{\eta}). \qquad (3.14)$$

Let us compute now the bias of $\hat{\boldsymbol{\eta}}_\pi$ under ξ. We know that $\tilde{\boldsymbol{\eta}} = \mathbf{A}'\boldsymbol{\beta}$ and $\boldsymbol{\eta} = \mathbf{A}'_r\boldsymbol{\beta}$. Consider,

$$\begin{aligned}
E_\xi(\hat{\boldsymbol{\eta}}_\pi) &= (\mathbf{Z}'_{r,s}\boldsymbol{\Pi}_s^{-1}\mathbf{Z}_{r,s})^{-1}\mathbf{Z}'_{r,s}\boldsymbol{\Pi}_s^{-1}E_\xi(\mathbf{y}_s) \\
&= (\mathbf{Z}'_{r,s}\boldsymbol{\Pi}_s^{-1}\mathbf{Z}_{r,s})^{-1}\mathbf{Z}'_{r,s}\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s\boldsymbol{\beta} \\
&= \underbrace{(\mathbf{Z}'_{r,s}\boldsymbol{\Pi}_s^{-1}\mathbf{Z}_{r,s})^{-1}\mathbf{Z}'_{r,s}\boldsymbol{\Pi}_s^{-1}\mathbf{Z}_{r,s}}_{I}\mathbf{A}'_r\boldsymbol{\beta} \\
&= \mathbf{A}'_r\boldsymbol{\beta}
\end{aligned}$$

Since $\mathbf{X}_r = \mathbf{Z}'_{r,s}\mathbf{A}'_r$. So finally,

$$\begin{aligned}
E_\xi(\hat{\boldsymbol{\eta}}_\pi) - \boldsymbol{\eta} &= \mathbf{A}'_r\boldsymbol{\beta} - \boldsymbol{\eta} \\
&= \mathbf{A}'_r\boldsymbol{\beta} - \mathbf{A}'\boldsymbol{\beta} \\
&= (\mathbf{A}'_r - \mathbf{A}')\boldsymbol{\beta}
\end{aligned}$$

so the bias of $\hat{\boldsymbol{\eta}}_\pi$ under the model ξ does not depend on the sample and it is the same $\xi-$bias as in a non-sampling framework (Gunst and Mason, 1977). Hence from (3.14) the bias of $\hat{t}_{PC}$ becomes,

$$E_\xi(\hat{t}_{PC} - t_y) = -\left(\sum_s \frac{\tilde{\boldsymbol{z}}'_k}{\pi_k}\right)(\mathbf{A}'_r - \mathbf{A}')\boldsymbol{\beta}. \tag{3.15}$$

2. Applying the design expectation on 3.15, we get,

$$E_p E_\xi(\hat{t}_{PC} - t_y) = -E_p\left(\sum_s \frac{\tilde{\boldsymbol{z}}'_k}{\pi_k}\right)(\mathbf{A}'_r - \mathbf{A}')\boldsymbol{\beta}.$$

We know that the Horvitz Thompson estimator is design unbiased, i.e., $E_p\left(\sum_s \frac{\tilde{\boldsymbol{z}}'_k}{\pi_k}\right) = \sum_U \tilde{\boldsymbol{z}}'_k = 0$. So we get,

$$E_p E_\xi(\hat{t}_{PC} - t_y) = 0 = Bias_{\xi,p}(t_{PC}).$$

□

3.2.3 Design-based properties

The estimator $\hat{t}_{PC}$ is no longer unbiased with respect to $p(.)$, we prove that it is asymptotically design-unbiased. In order to prove it, we consider the asymptotic framework as introduced by Isaki and Fuller (1982) and the following assumptions.

Hypothesis

(A1). $\pi_k > \lambda > 0\ \forall k \in U$

(A2). $\overline{\lim}_{N\to\infty} n \max_{k\neq l} |\pi_{kl} - \pi_k\pi_l| < c < \infty$

(A3). $\lim_{N\to\infty} \frac{1}{N}\sum_U y_k^2 < \infty$ with ξ-probability 1.

(A4). $\lim_{N\to\infty} \frac{n}{N} = \pi \in (0,1)$

(A5). $||\mathbf{x}_k|| < c < \infty$ for all $k \in U$.

Lemma 1. *Under assumptions (A1)-(A4), we have that*

$$\frac{1}{N}\left(\hat{t}_{y,\pi} - t_y\right) = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof. We calculate the variance under the sampling design of $\frac{1}{N}\left(\hat{t}_{y,\pi} - t_y\right)$. If

$$E\left[\frac{1}{N}(\hat{t}_{y,\pi} - t_y)\right]^2 = O\left(\frac{1}{n}\right)$$

implies that $\frac{1}{N}(\hat{t}_{y,\pi} - t_y) = O_p\left(\frac{1}{\sqrt{n}}\right)$. We have,

$$\begin{aligned} E\left[\frac{1}{N}(\hat{t}_{y,\pi} - t_y)\right]^2 &= Var\left(\frac{1}{N}\hat{t}_{y,\pi}\right) \\ &= \frac{1}{N^2}\sum_U\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \end{aligned} \tag{3.16}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ and $k, l \in U$. We partition 3.16 into two cases, $k = l$ and $k \neq l$ respectively.

$$E\left[\frac{1}{N}(\hat{t}_{y,\pi} - t_y)\right]^2 = \underbrace{\frac{1}{N^2}\sum_U \pi_k(1-\pi_k)\frac{y_k^2}{\pi_k^2}}_{(i)} + \underbrace{\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U_{l\neq k}}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}}_{(ii)}. \tag{3.17}$$

Consider 3.17(i) which is bounded by

$$\frac{1-\lambda}{\lambda}\frac{1}{N^2}\sum_U y_k^2 = \frac{1-\lambda}{\lambda}\frac{n}{N}\frac{1}{n}\frac{1}{N}\sum_U y_k^2 = O\left(\frac{1}{n}\right),$$

by hypothesis (A3) and (A4). Now, consider 3.17(ii). This term is bounded by

$$\frac{\max_{k\neq l}|\Delta_{kl}|}{N^2\lambda^2}\sum_{k\in U}\sum_{l\in U_{l\neq k}} y_k y_l \leq \frac{n}{n}\frac{\max_{k\neq l}|\Delta_{kl}|}{\lambda^2}\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U_{l\neq k}} y_k y_l.$$

We have

$$\frac{1}{N^2}\sum_{k\in U}\sum_{\substack{l\in U\\ l\neq k}} y_k y_l \leq \frac{1}{N^2}\left(\sum_{k\in U} y_k\right)^2 \leq \frac{1}{N}\sum_{k\in U} {y_k}^2. \tag{3.18}$$

We have in fact added the terms for $k = l$, so the left hand side of expression (3.18) is inferior than the right hand side. The last inequality is obtained by applying the Cauchy Schwarz inequality. Thus, (ii) is bounded by,

$$\frac{n}{n}\frac{\max_{k\neq l}|\Delta_{kl}|}{\lambda^2}\frac{1}{N}\sum_{k\in U} {y_k}^2 = O\left(\frac{1}{n}\right). \tag{3.19}$$

□

Proposition 3. *Under hypothesis (A1)-(A5), we have that:* $\hat{\boldsymbol{\eta}}_\pi - \hat{\boldsymbol{\eta}} = O_p\left(\frac{1}{\sqrt{n}}\right)$. *As consequence,* $\frac{1}{N}\left(\hat{t}_{PC} - t_y\right) = \frac{1}{N}\left(\hat{t}_{DIFF} - t_y\right) + o_p\left(\frac{1}{\sqrt{n}}\right)$, *where* $\hat{t}_{DIFF} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\tilde{\mathbf{z}}'_k}{\pi_k} - \sum_U \tilde{\mathbf{z}}'_k\right)\hat{\boldsymbol{\eta}}$ *where* $\hat{\boldsymbol{\eta}} = \left(\mathbf{Z}'_r\mathbf{Z}_r\right)^{-1}\mathbf{Z}'_r\mathbf{y}$. *The asymptotic variance of* $\hat{t}_{PC}$ *is the variance of* $\hat{t}_{DIFF}$,

$$AV(\hat{t}_{PC}) = \sum_U \sum_U (\pi_{kl} - \pi_k\pi_l)\frac{y_k - \tilde{\boldsymbol{z}}'_k\hat{\boldsymbol{\eta}}}{\pi_k}\frac{y_l - \tilde{\boldsymbol{z}}'_l\hat{\boldsymbol{\eta}}}{\pi_l}. \tag{3.20}$$

The asymptotic variance $AV(\hat{t}_{PC})$ *is not known and we suggest estimating it by:*

$$\hat{V}(\hat{t}_{PC}) = \sum_s \sum_s \frac{(\pi_{kl} - \pi_k\pi_l)}{\pi_{kl}}\frac{y_k - \tilde{\boldsymbol{z}}'_k\hat{\boldsymbol{\eta}}_\pi}{\pi_k}\frac{y_l - \tilde{\boldsymbol{z}}'_l\hat{\boldsymbol{\eta}}_\pi}{\pi_l}. \tag{3.21}$$

Proof. Now for the estimator of the principal component regression coefficient

$$\hat{\boldsymbol{\eta}}_\pi = \left(\sum_s \frac{\tilde{\mathbf{z}}_k\tilde{\mathbf{z}}'_k}{\pi_k}\right)^{-1}\left(\sum_s \frac{\tilde{\mathbf{z}}_k\mathbf{y}_k}{\pi_k}\right),$$

where $\sum_s \frac{\tilde{\mathbf{z}}'_k\mathbf{y}_k}{\pi_k}$ is the Horvitz-Thompson estimator of $\sum_U \tilde{\mathbf{z}}_k\mathbf{y}_k$. We can consider $\hat{\boldsymbol{\eta}}_\pi$ again as

$$\hat{\boldsymbol{\eta}}_\pi = \underbrace{(\mathbf{Z}'_{r,s}\boldsymbol{\Pi}_s^{-1}\mathbf{Z}_{r,s})^{-1}}_{\hat{\mathbf{Q}}_s}\underbrace{\mathbf{Z}'_{r,s}\boldsymbol{\Pi}_s^{-1}\mathbf{y}_s}_{\hat{\mathbf{q}}_s}$$

where $\mathbf{Q} = \sum_U \tilde{\mathbf{z}}_k\tilde{\mathbf{z}}'_k = \mathbf{Z}'_r\mathbf{Z}_r$ with $\hat{\mathbf{Q}}_s = \mathbf{Z}'_{r,s}\boldsymbol{\Pi}_s^{-1}\mathbf{Z}_{r,s}$ and $\mathbf{q} = \mathbf{Z}'_r\mathbf{y}_U$ with $\hat{\mathbf{q}}_s = \mathbf{Z}'_{r,s}\boldsymbol{\Pi}_s^{-1}\mathbf{y}_s$.

Here, if we show that,

a). $\frac{1}{N}||\hat{\mathbf{Q}}_s - \mathbf{Q}||_2 = O_p\left(\frac{1}{\sqrt{n}}\right)$ where $||.||_2$ is the trace norm defined for any matrix S by $||S||_2^2 = trace(S'S)$.

b). $\frac{1}{N}||\hat{\mathbf{q}}_s - \mathbf{q}|| = O_p\left(\frac{1}{\sqrt{n}}\right)$ where $||.||$ is the Euclidean norm,

then we can show that

$$||\hat{\boldsymbol{\eta}}_\pi - \hat{\boldsymbol{\eta}}||^2 = O_p\left(\frac{1}{n}\right), \tag{3.22}$$

since,

$$\begin{aligned}\hat{\boldsymbol{\eta}}_\pi - \hat{\boldsymbol{\eta}} &= \left(\frac{1}{N}\hat{\mathbf{Q}}_s\right)^{-1}\frac{1}{N}\hat{\mathbf{q}}_s - \left(\frac{1}{N}\mathbf{Q}\right)^{-1}\frac{1}{N}\mathbf{q} \\ &= N\underbrace{\left(\hat{\mathbf{Q}}_s^{-1} - \mathbf{Q}^{-1}\right)}\frac{1}{N}\left(\hat{\mathbf{q}}_s\right) + N\mathbf{Q}^{-1}\frac{1}{N}\left(\hat{\mathbf{q}}_s - \mathbf{q}\right).\end{aligned}$$

Using $\hat{\mathbf{Q}}_s^{-1} - \mathbf{Q}^{-1} = \hat{\mathbf{Q}}_s^{-1}\left(\mathbf{Q} - \hat{\mathbf{Q}}_s\right)\mathbf{Q}^{-1}$, we get,

$$\begin{aligned} & N||\hat{\mathbf{Q}}_s^{-1} - \mathbf{Q}^{-1}||_2 = N||\hat{\mathbf{Q}}_s^{-1}\left(\mathbf{Q} - \hat{\mathbf{Q}}_s\right)\mathbf{Q}^{-1}||_2 \\ \leq & N||\hat{\mathbf{Q}}_s^{-1}||_2 \cdot \underbrace{||\left(\mathbf{Q} - \hat{\mathbf{Q}}_s\right)\frac{1}{N}||_2}_{O_p} \cdot N||\mathbf{Q}^{-1}||_2 \left(\frac{1}{\sqrt{n}}\right) \leq N||\hat{\mathbf{Q}}_s^{-1}||_2 \cdot O_p\left(\frac{1}{\sqrt{n}}\right) \cdot N||\mathbf{Q}^{-1}||_2 \\ = & O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{3.23}$$

Since the eigenvalues of $\frac{1}{N}\mathbf{Q}$ are far from zero so $N||\mathbf{Q}^{-1}||_2$ is bounded. The same is true for $\frac{1}{N}\hat{\mathbf{Q}}_s$, so $||\frac{1}{N}\hat{\mathbf{Q}}_s|| = O_p(1)$.

Let prove now that $\frac{1}{N}||\hat{\mathbf{Q}}_s - \mathbf{Q}||_2 = O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\frac{1}{N}||\hat{\mathbf{q}}_s - \mathbf{q}|| = O_p\left(\frac{1}{\sqrt{n}}\right)$. We have $\tilde{\mathbf{z}}_k = (z_{ki})_{i=1}^r$ and

$$\frac{1}{N^2}||\hat{\mathbf{q}}_s - \mathbf{q}||^2 = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\left(\sum_{i=1}^r (z_{ki}y_k)(z_{li}y_l)\right)\alpha_k\alpha_l$$

where $\alpha_k = \frac{I_k}{\pi_k} - 1$, $k \in U$. Using the Lemma (3.6.2) of Goga (2003, page 182), the results follows if we prove that,

$$\underbrace{\frac{1}{N}\sum_{k\in U}\left(\sum_{i=1}^r (z_{ki}y_k)^2\right) < \infty}_{(i)}$$

and

$$\underbrace{\frac{1}{N^2}\sum\sum_{k\neq l}\left|\left(\sum_{i=1}^r (z_{ki}y_k)(z_{li}y_l)\right)\right| < \infty}_{(ii)}.$$

Now, the above relation (i) can be written as,

$$\begin{aligned} \frac{1}{N}\sum_{k\in U}\left(\sum_{i=1}^r (z_{ki}y_k)^2\right) &= \frac{1}{N}\sum_{k\in U}\left(\sum_{i=1}^r z_{ki}^2 y_k^2\right) \\ &= \frac{1}{N}\sum_{k\in U}\underbrace{\left(\sum_{i=1}^r z_{ki}^2\right)}_{||\tilde{\mathbf{z}}_k||^2} y_k^2 \\ &= \frac{1}{N}\sum_{k\in U}||\tilde{\mathbf{z}}_k||^2 \cdot y_k^2 \quad \text{, and using (A5),} \\ &\leq \frac{c}{N}\sum_{k\in U} y_k^2 < \infty \end{aligned}$$

Relation (ii) can be written as,

$$\begin{aligned}\frac{1}{N^2}\sum\sum_{k\neq l}\left|\left(\sum_{i=1}^{r}(z_{ki}y_k)(z_{li}y_l)\right)\right| &\leq \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\sum_{i=1}^{r}|z_{ki}y_k z_{li}y_l| \\ &= \frac{1}{N^2}\sum_{i=1}^{r}\left(\sum_{k\in U}|z_{ki}y_k|\right)^2 \leq \frac{1}{N}\sum_{k\in U}\left(\sum_{i=1}^{r}z_{ki}^2 y_k^2\right) < \infty.\end{aligned}$$

Now we shall prove (a). We have

$$\mathbf{Q} = \sum_{U}\tilde{\mathbf{z}}_k\tilde{\mathbf{z}}_k'$$

with

$$\hat{\mathbf{Q}}_s = \sum_{s}\frac{\tilde{\mathbf{z}}_k\tilde{\mathbf{z}}_k'}{\pi_k}.$$

So,

$$\hat{\mathbf{Q}}_s - \mathbf{Q} = \sum_{U}\tilde{\mathbf{z}}_k\tilde{\mathbf{z}}_k'\alpha_k.$$

Also,

$$||\hat{\mathbf{Q}}_s - \mathbf{Q}||_2^2 = \sum_{k\in U}\sum_{l\in U}tr(\tilde{\mathbf{z}}_k\tilde{\mathbf{z}}_k'\tilde{\mathbf{z}}_l\tilde{\mathbf{z}}_l')\alpha_k\alpha_l$$

and using the same lemma from Goga (2003), the results follows since

$$\frac{1}{N}\sum_{U}tr\left(\tilde{\mathbf{z}}_k\tilde{\mathbf{z}}_k'\cdot\tilde{\mathbf{z}}_k\tilde{\mathbf{z}}_k'\right) \leq \frac{1}{N}\sum_{U}||\tilde{\mathbf{z}}_k\tilde{\mathbf{z}}_k'||_2^2 \leq \frac{1}{N}\sum_{U}||\tilde{\mathbf{z}}_k||^4 < \infty.$$

□

3.2.4 Calibration with Principal Components

The calibration technique (Deville and Särndal (1992)) described briefly in chapter 1, deals with deriving a weighted estimator $\hat{t}_{wy}$ of population total using the sample calibrated weights w_k. For the chi-square distance $\sum_s (w_k - d_k)^2/d_k q_k$, the weights are the solution of the following optimization problem,

$$\begin{aligned}\mathbf{w}^c &= \text{argmin}_{\mathbf{w}_s}(\mathbf{w}_s - \mathbf{d}_s)'\widetilde{\mathbf{\Pi}}_s(\mathbf{w}_s - \mathbf{d}_s) \\ &\quad \text{subject to} \quad \mathbf{w}_s'\mathbf{X}_s = \mathbf{1}_U'\mathbf{X},\end{aligned}$$

where $\widetilde{\mathbf{\Pi}}_s = \text{diag}(q_k^{-1}d_k^{-1})_{k\in s}$ and q_k are positive constants most often equal to 1. The resulting calibration weights can be written as,

$$\mathbf{w}^c = \mathbf{d}_s - \widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s)^{-1}(\mathbf{d}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})'.$$

The estimator $\hat{t}_{PC}$ given by (3.10) may be obtained using the calibration approach (Deville and Särndal (1992)). The vector of auxiliary information is now composed of the first r principal components such that $\mathbf{Z}_r = (\mathbf{z}_1, \ldots, \mathbf{z}_r)$. More exactly, we construct the estimator $\hat{t}_{yw} = \sum_s w_k^c y_k$ calibrated on the finite totals of the principal components $\mathbf{z}_i$, $i = 1 \ldots, r$ instead of $\mathbf{X}_i$, $i = 1, \ldots, p$ variables. So, the weights $\mathbf{w}^c = (w_k^c)_{k\in s}$ satisfy

$$\begin{aligned} \mathbf{w}^c &= \text{argmin}_{\mathbf{w}} \sum_s \frac{(w_k - d_k)^2}{d_k q_k} \\ \text{subject to} \quad \mathbf{w}'^c\mathbf{Z}_{r,s} &= \mathbf{1}_U'\mathbf{Z}_r. \end{aligned} \tag{3.24}$$

The resulting PC calibrated weights are given as,

$$\mathbf{w}^c = \mathbf{d}_s - \widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{Z}_{r,s}(\mathbf{Z}_{r,s}'\widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{Z}_{r,s})^{-1}(\mathbf{d}_s'\mathbf{Z}_{r,s} - \mathbf{1}_U'\mathbf{Z})'. \tag{3.25}$$

The calibration estimator for the total t_y is in fact a GREG type estimator given by,

$$\begin{aligned} \hat{t}_{PC}^c = \mathbf{w}^{c'}\mathbf{y}_s &= \mathbf{d}_s'\mathbf{y}_s - \left(\mathbf{d}_s'\mathbf{Z}_{r,s} - \mathbf{1}_U'\mathbf{Z}_r\right)\left(\mathbf{Z}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{Z}_{r,s}\right)^{-1}\mathbf{Z}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s \\ &= \hat{t}_{y,\pi} - \left(\sum_s \frac{\mathbf{z}_k'}{\pi_k} - \sum_U \mathbf{z}_k'\right)\hat{\eta}_\pi \end{aligned}$$

The calibration weights obtained in this way will not allow to find exact totals of the initial auxiliary variables $\mathbf{X}_i$ for $i = 1, \ldots, p$. This property is verified in the projection space on $\mathbf{Z}_r$.

3.2.5 Calibration on second moment of the principal component variables

An interesting extension of the classical calibration approach can be obtained noting that the variance of the principal components variable $\mathbf{z}_i$ is the eigen value

λ_i,

$$\frac{1}{N}\mathbf{z}_i'\mathbf{z}_i = \frac{1}{N}\sum_{k\in U} z_{ki}^2 = \lambda_i, \quad \text{for all} \quad i = 1, \ldots, p$$

This means that we can add supplementary calibration equations on the second moment of the principal components. Consider $\mathbf{Z}_r^2 = (\mathbf{z}_1^2, \ldots, \mathbf{z}_r^2)$ with $\mathbf{z}_i^2 = (z_{ki}^2)_{k\in U}$. We want to find the calibration weights $\mathbf{w}^c$ that satisfy the following optimization problem

$$\begin{aligned} \mathbf{w}^c &= \text{argmin}_{\mathbf{w}} \sum_s \frac{(w_k - d_k)^2}{d_k q_k} \\ \text{subject to} \quad \mathbf{w}'^c\mathbf{Z}_{r,s} &= \mathbf{1}_U'\mathbf{Z}_r, \quad \mathbf{w}'^c\mathbf{Z}_{r,s}^2 = \mathbf{1}_U'\mathbf{Z}_r^2 \end{aligned}$$

where $\mathbf{Z}_{r,s}^2$ is the sample restriction of $\mathbf{Z}_r^2$. In order to compute the calibration weights, the objective function is written in a matrix form as,

$$\mathbf{w}^c = \text{argmin}_{\mathbf{w}}(\mathbf{w} - \mathbf{d}_s)'\tilde{\mathbf{\Pi}}_s(\mathbf{w} - \mathbf{d}_s)$$

where $\tilde{\mathbf{\Pi}}_s = \text{diag}(q_k)_{k\in s}^{-1}\mathbf{\Pi}_s$. We can form a matrix $\mathbf{T}_r$ of dimension $N \times 2r$ such that, $\mathbf{T}_r = (\mathbf{Z}_r, \mathbf{Z}_r^2)$ and its sample restriction as, $\mathbf{T}_{r,s} = (\mathbf{Z}_{r,s}, \mathbf{Z}_{r,s}^2)$. The set of calibration constraints can be re-structured as,

$$\mathbf{w}'\mathbf{T}_{r,s} = \mathbf{1}_U'\mathbf{T}_r.$$

We construct a Lagrangian function $\mathcal{L}(\mathbf{w}, \lambda)$,

$$\mathcal{L}(\mathbf{w}, \lambda) = (\mathbf{w} - \mathbf{d}_s)'\tilde{\mathbf{\Pi}}_s(\mathbf{w} - \mathbf{d}_s) - 2\left(\mathbf{w}'\mathbf{T}_{r,s} - \mathbf{1}_U'\mathbf{T}_r\right)\lambda$$

We take the first derivative of Lagrangian function with respect to $\mathbf{w}$ and λ

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = 2\tilde{\mathbf{\Pi}}_s(\mathbf{w} - \mathbf{d}_s) - 2\mathbf{T}_{r,s}\lambda$$

and put it equal to 0, we get,

$$\tilde{\mathbf{\Pi}}_s(\mathbf{w} - \mathbf{d}_s) - \mathbf{T}_{r,s}\lambda = 0$$

$$\mathbf{w} - \mathbf{d}_s = \tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s}\lambda.$$

Finally we get the following shape of the weights where λ is unknown

$$\mathbf{w} = \mathbf{d}_s + \tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s}\lambda.$$

We consider the calibration equations again and put the above acquired weights in it,

$$\mathbf{w}'\mathbf{T}_{r,s} = \mathbf{1}_U'\mathbf{T}_r$$

$$\mathbf{d}_s'\mathbf{T}_{r,s} + \lambda'\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s} = \mathbf{1}_U'\mathbf{T}_r$$

$$\mathbf{d}_s'\mathbf{T}_{r,s} - \mathbf{1}_U'\mathbf{T}_r + \lambda'\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s} = 0.$$

Hence we have λ as follows,

$$\lambda = -\left(\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s}\right)^{-1}\left(\mathbf{d}_s'\mathbf{T}_{r,s} - \mathbf{1}_U'\mathbf{T}_r\right)'.$$

The solution is given by

$$\mathbf{w}^c = \mathbf{d}_s - \tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s}\left(\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s}\right)^{-1}\left(\mathbf{d}_s'\mathbf{T}_{r,s} - \mathbf{1}_U'\mathbf{T}_r\right)' \tag{3.26}$$

The calibration estimator for the total t_y is in fact a generalized regression estimator for the $N \times (2r)$-dimensional auxiliary information $\mathbf{T}_r = (\mathbf{Z}_r, \mathbf{Z}_r^2)$ as follows

$$\begin{aligned}\hat{t}_{MPC}^c = \mathbf{w}^{c'}\mathbf{y}_s &= \mathbf{d}_s'\mathbf{y}_s - \left(\mathbf{d}_s'\mathbf{T}_{r,s} - \mathbf{1}_U'\mathbf{T}_r\right)\left(\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s}\right)^{-1}\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s \\ &= \hat{t}_{y,\pi} - \left(\sum_s \frac{\mathbf{t}_k'}{\pi_k} - \sum_U \mathbf{t}_k'\right)\hat{\mathbf{B}}_{z,z^2} \end{aligned} \tag{3.27}$$

where $\mathbf{t}_k = (\tilde{\boldsymbol{z}}_k', \tilde{\boldsymbol{z}}_{\boldsymbol{k}}^{\mathbf{2}\prime})$ is the k-th row of $\mathbf{T}_r$ and $\hat{\mathbf{B}}_{z,z^2} = \left(\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s}\right)^{-1}\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s$. The idea of finding estimates using the calibration on second order (or higher) moments of the auxiliary variables stems from the works done by Ren (2000) and Särndal (2007). Both articles show that an estimator constructed using the calibration on the second order (or higher) moments of the auxiliary variables is expected to perform better than the estimates constructed using only their first order moments. Nevertheless, calibration on the second moment adds r supplementary equations so a small number r should be used.

3.2.6 Partial Principal Component Calibration

Often there is the case when we want to find exact sample estimates of total for some auxiliary variables. This might be due to the importance associated with those auxiliary variables. Age, sex, socio-professional categories etc may be a few of those auxiliary variables. Following the idea of Bardsley and Chambers (1984) for partial ridge regression, we can modify the simple calibration into partial principal component calibration. Breidt and Chauvet (2011) have used the same technique but at the sampling stage. In their study, the sample was selected by the *cube* method.

For this purpose, we partition our data matrix into two parts such that it can be written as,

$$\mathbf{X} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$$

where $\tilde{\mathbf{X}}_1 = [\mathbf{X}_1, \ldots, \mathbf{X}_{p_1}]$ includes those $p_1 (<< p)$ variables which need to be exactly calibrated. To be more precise, the variables in $\tilde{\mathbf{X}}_1$ contain the maximum importance in the estimation procedure and are very few in numbers. On the other hand $\tilde{\mathbf{X}}_2 = [\mathbf{X}_{p_1+1}, \ldots, \mathbf{X}_p]$ contains $p - p_1$ variables such that $p - p_1 >> p_1$.

We shall now calculate the $p - p_1$ principal components $\tilde{\mathbf{Z}}_2$ corresponding to the variables in $\tilde{\mathbf{X}}_2$ and orthogonal on $\tilde{\mathbf{X}}_1$. First r_1 principal components are chosen from $\tilde{\mathbf{Z}}_2 = [\mathbf{z}_{p_1+1}, \ldots, \mathbf{z}_p]$ such that $r_1 << (p - p_1)$. We shall denote these r_1 principal components by $\tilde{\mathbf{Z}}_{2r_1} = [\mathbf{z}_{p_1+1}, \ldots, \mathbf{z}_{r_1}]$ and their selection can be made using any of the famous methods for selection of principal components (see Mason and Gunst (1985) and Jolliffe (2002)). Our auxiliary data matrix denoted by $\mathbf{M}$ becomes,

$$\mathbf{M} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{Z}}_{2r_1})$$

Thus our optimization problem becomes,

$$\mathbf{w}_{ppc}^{c} = \text{argmin}_{\mathbf{w}} \sum_{s} \frac{(w_k - d_k)^2}{d_k q_k}$$

subject to,

$$\mathbf{w}_{ppc}^{\prime c} \mathbf{M}_s = \mathbf{1}_U \mathbf{M}$$

where $\mathbf{M}_s = (\tilde{\mathbf{X}}_{1s}, \tilde{\mathbf{Z}}_{2sr_1})$ is the sample restriction of $\mathbf{M}$. The resulting weights will get the following shape obtained in a similar manner to the previous,

$$\mathbf{w}_{ppc}^{c} = \mathbf{d}_s - \tilde{\mathbf{\Pi}}_s^{-1}\mathbf{M}_s \left(\mathbf{M}_s'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{M}_s\right)^{-1} \left(\mathbf{d}_s'\mathbf{M}_s - \mathbf{1}_U'\mathbf{M}\right)' \tag{3.28}$$

These weights are in fact using maximum variation available in $\tilde{\mathbf{Z}}_{2sr_1}$ and on the same time minimizing the dimension of auxiliary data.(A certain aspect of the presence of the multicollinearity among the variables in $\tilde{\mathbf{Z}}_{2sr_1}$ has to be verified and its absence may ensure the improvement in the estimation procedure with reduction in the dimension). The estimator of the total for the above weights $\mathbf{w}_{ppc}^{\prime c}$ becomes,

$$\begin{aligned}
\hat{t}_{ppc}^{c} &= \mathbf{w}_{ppc}^{\prime c}\mathbf{y}_s \\
&= \mathbf{d}_s'\mathbf{y}_s - \left(\mathbf{d}_s'\mathbf{M}_s - \mathbf{1}_U'\mathbf{M}\right)\left(\mathbf{M}_s'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{M}_s\right)^{-1}\mathbf{M}_s'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s \\
&= \hat{t}_{y,\pi} - \left(\sum_s \frac{\mathbf{m}_k'}{\pi_k} - \sum_U \mathbf{m}_k'\right)\hat{\mathbf{B}}_m
\end{aligned} \tag{3.29}$$

where $\hat{\mathbf{B}}_m = \left(\mathbf{M}_s'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{M}_s\right)^{-1}\mathbf{M}_s'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s$.

3.2.7 Estimated Principal Component Calibration

We discussed deriving the principal components when the auxiliary data set $\mathbf{X}_1, \ldots, \mathbf{X}_p$ is available for all units $k \in U$. This however may not be possible practically. In this section we consider the case when we know $\mathbf{x}_k$ only for the sample units $k \in s \subset U$ but their population means and standard deviations $\sigma_1, \ldots, \sigma_p$ are known. As previously, we suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are standardized. We shall estimate the covariance matrix

$$\mathbf{\Sigma} = \frac{1}{N}(\mathbf{X}'\mathbf{X}) = \frac{1}{N}\sum_{k\in U}\mathbf{x}_k\mathbf{x}_k' = \frac{1}{N}(\mathbf{X} - \underbrace{\bar{\mathbf{X}}}_{0})'(\mathbf{X} - \underbrace{\bar{\mathbf{X}}}_{0}) \text{with} \quad \mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$$

by

$$\begin{aligned}
\hat{\mathbf{\Sigma}} &= \frac{1}{\hat{N}}\left(\mathbf{X}_s - \hat{\bar{\mathbf{X}}}\right)'\mathbf{\Pi}_s^{-1}\left(\mathbf{X}_s - \hat{\bar{\mathbf{X}}}\right), \\
&= \frac{1}{\hat{N}}\sum_s \frac{1}{\pi_k}(\mathbf{x}_k - \hat{\bar{\mathbf{X}}})(\mathbf{x}_k - \hat{\bar{\mathbf{X}}})'
\end{aligned}$$

where $\hat{\bar{\mathbf{X}}} = \frac{1}{\hat{N}} \sum_s \frac{\mathbf{x}_k}{\pi_k}$ and $\hat{N} = \sum_s \frac{1}{\pi_k}$. For example, in simple random sampling without replacement we have $\pi_k = \frac{n}{N}$.

This estimated covariance matrix $\hat{\mathbf{\Sigma}}$ has the eigenvalue-eigenvector pair as $(\hat{\lambda}_1, \hat{\mathbf{a}}_1), \ldots, (\hat{\lambda}_p, \hat{\mathbf{a}}_p)$ such that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$ and $\hat{\lambda}_i$, related to $\hat{\mathbf{a}}_i$ are design-based estimators of λ_i, related to $\mathbf{a}_i$. We have that

$$\hat{\mathbf{\Sigma}}\hat{\mathbf{a}}_i = \hat{\lambda}_i \hat{\mathbf{a}}_i \tag{3.30}$$

$$\hat{\mathbf{\Sigma}} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{\mathbf{a}}_i \hat{\mathbf{a}}_i', \tag{3.31}$$

and Cardot *et al* (2010) showed that under assumptions (A1), (A2), and (A4), (A5)

$$|\hat{\lambda}_i - \lambda_i| = O_p\left(\frac{1}{\sqrt{n}}\right)$$

and

$$||\hat{\mathbf{a}}_i - \mathbf{a}_i|| = O_p\left(\frac{1}{\sqrt{n}}\right).$$

We suggest to estimate the principal component $\mathbf{z}_i$ by,

$$\hat{\mathbf{z}}_i = \mathbf{X}\hat{\mathbf{a}}_i, \quad i = 1, \ldots, p.$$

Hence

$$\hat{\mathbf{Z}} = (\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_p) = \mathbf{X}\hat{\mathbf{A}},$$

where $\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_p)$. Remark that the $\hat{\mathbf{z}}_i$ is known for $i \in s$ but the total of $\hat{\mathbf{z}}_i$ over population U may be computed as

$$t_{\hat{\mathbf{z}}_i} = \left(t_x'\right) \hat{\mathbf{a}}_i = 0$$

$$tr(\hat{\mathbf{\Sigma}}) = p = \hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p.$$

The first r_2 PC's, $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots, \hat{\mathbf{z}}_{r_2}$ are selected on the basis of the first r_2 largest eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_{r_2}$. We use these estimated principal components $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots, \hat{\mathbf{z}}_{r_2}$ to construct the calibration estimator of the population total.

$$\hat{\mathbf{Z}}_r = (\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_{r_2}) = \left(\hat{\bar{\mathbf{z}}}_k'\right)_{\{k \in U\}} = \mathbf{X}\hat{\mathbf{A}}_r,$$

with $\hat{\mathbf{A}}_r = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_r)$. Our optimization problem in this case becomes,

$$\mathbf{w}^c_{pc,est} = \text{argmin}_{\mathbf{w}} \sum_s \frac{(w_k - d_k)^2}{d_k q_k}$$

subject to,

$$\mathbf{w}'^c_{pc,est} \hat{\mathbf{Z}}_{r,s} = \mathbf{1}_U \hat{\mathbf{Z}}_r$$

where $\hat{\mathbf{Z}}_{r,s} = \mathbf{X}_s \hat{\mathbf{A}}_r$ is the sample restriction of $\hat{\mathbf{Z}}$. The resulting weights are given by,

$$\mathbf{w}^c_{pc,est} = \mathbf{d}_s - \widetilde{\mathbf{\Pi}}_s^{-1} \hat{\mathbf{Z}}_{r,s} (\hat{\mathbf{Z}}'_{r,s} \widetilde{\mathbf{\Pi}}_s^{-1} \hat{\mathbf{Z}}_{r,s})^{-1} (\mathbf{d}'_s \hat{\mathbf{Z}}_{r,s} - \mathbf{1}'_U \hat{\mathbf{Z}})'. \tag{3.32}$$

The total estimator is given as,

$$\begin{aligned} \hat{t}_{pc,est} &= \mathbf{w}'^c_{pc,est} \mathbf{y}_s = \mathbf{d}'_s \mathbf{y}_s - \left(\mathbf{d}'_s \hat{\mathbf{Z}}_{r,s} - \mathbf{1}'_U \hat{\mathbf{Z}}_r \right) \left(\hat{\mathbf{Z}}'_{r,s} \tilde{\mathbf{\Pi}}_s^{-1} \hat{\mathbf{Z}}_{rs} \right)^{-1} \hat{\mathbf{Z}}'_{r,s} \tilde{\mathbf{\Pi}}_s^{-1} \mathbf{y}_s \\ &= \hat{t}_{y,\pi} - \left(\sum_s \frac{\hat{\tilde{z}}'_k}{\pi_k} - \sum_U \hat{\tilde{z}}'_k \right) \hat{\boldsymbol{\eta}}_{\pi,est} \end{aligned} \tag{3.33}$$

where

$$\hat{\boldsymbol{\eta}}_{\pi,est} = \left(\hat{\mathbf{Z}}'_{r,s} \tilde{\mathbf{\Pi}}_s^{-1} \hat{\mathbf{Z}}_{r,s} \right)^{-1} \hat{\mathbf{Z}}'_{r,s} \tilde{\mathbf{\Pi}}_s^{-1} \mathbf{y}_s. \tag{3.34}$$

The estimator $\hat{t}_{pc,est}$ can be written in function of $\mathbf{X}$ as,

$$\hat{t}_{pc,est} = \hat{t}_{y,\pi} - \left(\hat{t}_{x,\pi} - t_x \right)' \hat{\boldsymbol{\beta}}_{PC,\pi,est}, \tag{3.35}$$

where $\hat{\boldsymbol{\beta}}_{PC,\pi,est} = \hat{\mathbf{A}}_r \hat{\boldsymbol{\eta}}_{\pi,est}$.

Result 11. *Under the assumptions (A1)-(A5), we have* $\hat{\boldsymbol{\eta}}_{\pi,est} - \hat{\boldsymbol{\eta}} = o_p(1)$. *As a consequence,* $\hat{\boldsymbol{\beta}}_{PC,\pi,est} - \hat{\boldsymbol{\beta}}_{PC,\pi} = o_p(1)$.

Proof. We consider for simplicity that $q_k = 1$ for all $k \in U$. We show first that $N^{-1}(\hat{\mathbf{Z}}'_{r,s} \mathbf{\Pi}_s^{-1} \hat{\mathbf{Z}}_{r,s} - \mathbf{Z_r}' \mathbf{Z}_r) = O_p(n^{-1/2})$.

Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$ be the $N \times p$ matrix of eigenvectors estimated by $\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_p)$. Let also $\mathbf{A}_r = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$ be the $N \times r$ matrix of the first r eigenvectors estimated by the $n \times r$ matrix $\hat{\mathbf{A}}_r = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_p)$. We have from relation (3.2) that $\mathbf{Z}_r = (\mathbf{z}_1, \ldots, \mathbf{z}_r) = \mathbf{X}\mathbf{A}_r$ yielding

$$\begin{aligned} \frac{1}{N} \mathbf{Z}'_r \mathbf{Z}_r &= \mathbf{A}'_r \left(\frac{1}{N} \mathbf{X}'\mathbf{X} \right) \mathbf{A}_r \\ &= \text{diag}(\lambda_j)_{j=1}^r := \mathbf{\Lambda}_r. \end{aligned}$$

From (3.31) and under the given assumptions, we have that

$$\frac{1}{\hat{N}}\mathbf{X}_s'\mathbf{\Pi}_s^{-1}\mathbf{X}_s = \hat{\mathbf{\Sigma}} + \hat{\overline{\mathbf{X}}}\hat{\overline{\mathbf{X}}}' = \hat{\mathbf{A}}\hat{\mathbf{\Lambda}}\hat{\mathbf{A}}' + o_p(1),$$

which yields

$$\begin{aligned}\frac{1}{N}\hat{\mathbf{Z}}_{r,s}'\mathbf{\Pi}_s^{-1}\hat{\mathbf{Z}}_{r,s} &= \hat{\mathbf{A}}_r'\left(\frac{1}{N}\mathbf{X}_s'\mathbf{\Pi}_s^{-1}\mathbf{X}_s\right)\hat{\mathbf{A}}_r \\ &= \hat{\mathbf{\Lambda}}_r + o_p(1).\end{aligned}$$

Under assumptions (A1),(A2) and (A4)-(A5), we have that $\hat{\lambda}_j - \lambda_j = O_p(n^{-1/2})$ (Cardot *et al.*, 2010) and by consequence, $||\hat{\mathbf{\Lambda}}_r - \mathbf{\Lambda}_r||_2 = O_p(n^{-1/2})$ where the trace norm $||\cdot||_2$ defined for any matrix S by $||S||_2^2 = trace(S'S)$. So, we have proved that

$$N^{-1}(\hat{\mathbf{Z}}_{r,s}'\mathbf{\Pi}_s^{-1}\hat{\mathbf{Z}}_{r,s} - \mathbf{Z}_r'\mathbf{Z}_r) = \hat{\mathbf{\Lambda}}_r - \mathbf{\Lambda}_r = O_p(n^{-1/2}).$$

Since λ_j and $\hat{\lambda}_j$ are strictly positive for all $j = 1, \ldots, r$, we obtain that

$$N\left((\hat{\mathbf{Z}}_{r,s}'\mathbf{\Pi}_s^{-1}\hat{\mathbf{Z}}_{r,s})^{-1} - (\mathbf{Z}_r'\mathbf{Z}_r)^{-1}\right) = O_p(n^{-1/2}).$$

Hence,

$$\begin{aligned}\hat{\boldsymbol{\eta}}_{\pi,est} - \hat{\boldsymbol{\eta}} &= (\hat{\mathbf{Z}}_{r,s}'\mathbf{\Pi}_s^{-1}\hat{\mathbf{Z}}_{r,s})^{-1}\hat{\mathbf{Z}}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s - (\mathbf{Z}_r'\mathbf{Z}_r)^{-1}\mathbf{Z}_r'\mathbf{y} \\ &= \left(N(\mathbf{Z}_r'\mathbf{Z}_r)^{-1} + O_p(n^{-1/2})\right)\left(N^{-1}\mathbf{Z}_r'\mathbf{y} + O_p(n^{-1/2})\right) - (\mathbf{Z}_r'\mathbf{Z}_r)^{-1}\mathbf{Z}_r'\mathbf{y} \\ &= O_p(n^{-1/2})\end{aligned}$$

since $N(\mathbf{Z}_r'\mathbf{Z}_r)^{-1} = O(1)$ and $N^{-1}\mathbf{Z}_r'\mathbf{y} = O(1)$ by assumption (A5). □

Result 12. *Under the assumptions (A1)-(A5), we have* $N^{-1}(\hat{t}_{pc,est} - t_y) = N^{-1}(\hat{t}_{DIFF} - t_y) + o_p(n^{-1/2})$ *where* $\hat{t}_{DIFF} = \hat{t}_{y\pi} - \left(\hat{t}_{\mathbf{z}\pi} - t_{\mathbf{z}}\right)'\hat{\boldsymbol{\eta}}$.

Proof. We have

$$\begin{aligned}\frac{1}{N}(\hat{t}_{pc,est} - t_y) &= \frac{1}{N}(\hat{t}_{y\pi} - t_y) - \frac{1}{N}\left(\hat{t}_{\mathbf{z}\pi} - t_{\mathbf{z}}\right)'\hat{\boldsymbol{\eta}}_{\pi,est} \\ &= \frac{1}{N}(\hat{t}_{y\pi} - t_y) - \frac{1}{N}\left(\hat{t}_{\boldsymbol{z}\pi} - t_{\boldsymbol{z}}\right)'\hat{\boldsymbol{\eta}} - \frac{1}{N}\left(\hat{t}_{\boldsymbol{z}\pi} - t_{\boldsymbol{z}}\right)'(\hat{\boldsymbol{\eta}}_{\pi,est} - \hat{\boldsymbol{\eta}}) \\ &= \frac{1}{N}(\hat{t}_{y\pi} - t_y) - \frac{1}{N}\left(\hat{t}_{\boldsymbol{z}\pi} - t_{\boldsymbol{z}}\right)'\hat{\boldsymbol{\eta}} + o_p(n^{-1/2}).\end{aligned}$$

□

The asymptotic variance $AV(\hat{t}_{pc,est})$ is similar to the 3.20, given as,

$$AV(\hat{t}_{pc,est}) = \sum_U \sum_U (\pi_{kl} - \pi_k \pi_l) \frac{y_k - \tilde{\boldsymbol{z}}_k' \hat{\boldsymbol{\eta}}}{\pi_k} \frac{y_l - \tilde{\boldsymbol{z}}_l' \hat{\boldsymbol{\eta}}}{\pi_l}. \tag{3.36}$$

But its estimate is different than 3.21 and is written as,

$$\hat{V}(\hat{t}_{pc,est}) = \sum_s \sum_s \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{y_k - \tilde{\boldsymbol{z}}_k' \hat{\boldsymbol{\eta}}_{\pi,est}}{\pi_k} \frac{y_l - \tilde{\boldsymbol{z}}_l' \hat{\boldsymbol{\eta}}_{\pi,est}}{\pi_l}. \tag{3.37}$$

3.3 Simulation Study on the PC calibrated estimators

From a large population of Mediametrie data described in section (2.4), we took a large sample of 5930 individuals on the 49 columns for the first two weeks of September 2010 and considered this sample as our population. The $\mathbf{X}$ matrix is of dimension 5930×49. We used 21 variables which include 4 quantitative and 17 qualitative variables. Different number of class in 17 qualitative variables resulted in 45 columns and hence making the data matrix of dimension 5930×49. The number of columns in this simulation study is 49 compared to 19 in section 2.4. The objective of our simulation study is the estimate of total time watched on second Monday of September 2010 on a particular T.V. channel by 5930 individual. The true value for the variable of interest is $t_y = 228537.6$ minutes watched on a particular T.V. channel. As we saw in section (2.4), the GREG estimator did not work as it came out to be singular due to the seriously ill-conditioned data. We realize a simulation study considering the HT estimator and on three types of calibrated PC estimators is realized including,

(a). The Horvitz-Thompson estimator can be written as,

$$\hat{t}_{y\pi} = \sum_s \frac{y_k}{\pi_k}.$$

(b). The *Population PC calibrated estimator* for population total, this means that we have $\mathbf{x}_k$ for all $k \in U$ allowing computation of $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_r$. The expression for the PC calibrated weights is given by,

$$\mathbf{w}^c = \mathbf{d}_s - \widetilde{\mathbf{\Pi}}_s^{-1} \mathbf{Z}_{r,s} (\mathbf{Z}_{r,s}' \widetilde{\mathbf{\Pi}}_s^{-1} \mathbf{Z}_{r,s})^{-1} (\mathbf{d}_s' \mathbf{Z}_{r,s} - \mathbf{1}_U' \mathbf{Z})',$$

and the PC calibrated estimator of the population total is,

$$\begin{aligned}\hat{t}^c_{PC} = \mathbf{w}^{c'}\mathbf{y}_s &= \mathbf{d}'_s\mathbf{y}_s - \left(\mathbf{d}'_s\mathbf{Z}_{r,s} - \mathbf{1}'_U\mathbf{Z}_r\right)\left(\mathbf{Z}'_{r,s}\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{Z}_{r,s}\right)^{-1}\mathbf{Z}'_{r,s}\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s \\ &= \hat{t}_{y,\pi} - \left(\sum_s \frac{\mathbf{z}'_k}{\pi_k} - \sum_U \mathbf{z}'_k\right)\hat{\eta}_\pi\end{aligned}$$

where $\hat{\boldsymbol{\eta}}_\pi = \left(\mathbf{Z}'_{r,s}\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{Z}_{r,s}\right)^{-1}\mathbf{Z}'_{r,s}\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s$.

(c). The *Estimated PC calibrated estimator* for population total. The expression for estimated PC calibrated weights is

$$\mathbf{w}^c_{pc,est} = \mathbf{d}_s - \widetilde{\mathbf{\Pi}}_s^{-1}\hat{\mathbf{Z}}_{r,s}(\hat{\mathbf{Z}}'_{r,s}\widetilde{\mathbf{\Pi}}_s^{-1}\hat{\mathbf{Z}}_{r,s})^{-1}(\mathbf{d}'_s\hat{\mathbf{Z}}_{r,s} - \mathbf{1}'_U\hat{\mathbf{Z}})',$$

the total estimator is given as,

$$\begin{aligned}\hat{t}^c_{pc,est} &= \mathbf{w}'^c_{pc,est}\mathbf{y}_s = \mathbf{d}'_s\mathbf{y}_s - \left(\mathbf{d}'_s\hat{\mathbf{Z}}_{r,s} - \mathbf{1}'_U\hat{\mathbf{Z}}_r\right)\left(\hat{\mathbf{Z}}'_{r,s}\tilde{\mathbf{\Pi}}_s^{-1}\hat{\mathbf{Z}}_{r,s}\right)^{-1}\hat{\mathbf{Z}}'_{r,s}\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s \\ &= \hat{t}_{y,\pi} - \left(\sum_s \frac{\hat{\mathbf{z}}'_k}{\pi_k} - \sum_U \hat{\mathbf{z}}'_k\right)\hat{\boldsymbol{\eta}}_{\pi,est},\end{aligned}$$

where $\hat{\boldsymbol{\eta}}_{\pi,est} = \left(\hat{\mathbf{Z}}'_{r,s}\tilde{\mathbf{\Pi}}_s^{-1}\hat{\mathbf{Z}}_{r,s}\right)^{-1}\hat{\mathbf{Z}}'_{r,s}\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s$.

(d). The *PPC calibrated estimator* for population total. The PPC calibrated weights are given by

$$\mathbf{w}^c_{ppc} = \mathbf{d}_s - \tilde{\mathbf{\Pi}}_s^{-1}\mathbf{M}_s\left(\mathbf{M}'_s\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{M}_s\right)^{-1}\left(\mathbf{d}'_s\mathbf{M}_s - \mathbf{1}'_U\mathbf{M}\right)',$$

the estimator of the total for the above weights $\mathbf{w}'^c_{ppc}$ is,

$$\begin{aligned}\hat{t}^c_{ppc} &= \mathbf{w}'^c_{ppc}\mathbf{y}_s \\ &= \mathbf{d}'_s\mathbf{y}_s - \left(\mathbf{d}'_s\mathbf{M}_s - \mathbf{1}'_U\mathbf{M}\right)\left(\mathbf{M}'_s\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{M}_s\right)^{-1}\mathbf{M}'_s\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s \\ &= \hat{t}_{y,\pi} - \left(\sum_s \frac{\mathbf{m}'_k}{\pi_k} - \sum_U \mathbf{m}'_k\right)\hat{\mathbf{B}}_m\end{aligned}$$

where $\hat{\mathbf{B}}_m = \left(\mathbf{M}'_s\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{M}_s\right)^{-1}\mathbf{M}'_s\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s$.

The principal components matrix $\mathbf{Z}$ of dimension 5930×49 is calculated using the data matrix $\mathbf{X}$. We considered $r = 25$ as the first 25 PC's account almost 84% of the variability available in the covariance matrix $\frac{1}{N}\mathbf{X}'\mathbf{X}$. So the matrix of PC's ($\mathbf{Z}_r$) (in Equation 3.25) has the dimensions 5930×25. We shall divide our simulation study in two major parts,

(i). Performance of PC calibrated estimator.

(ii). Variance estimation.

For the first part of simulation study (i), the number of simulations $B = 1000$ and for the second part of simulation study (ii) $B = 3000$. Simple random sampling without replacement (SRSWOR) is used as a sampling design in our applied computation of the calibrated estimators. Several performance indicators are computed to evaluate each type of the 3 calibrated estimators given above. This includes,

(1). Coefficient of variation for the PC weights $\mathbf{w}_s$

$$cv(\mathbf{w}_s) = \frac{\sqrt{Var(\mathbf{w}_s)}}{mean(\mathbf{w}_s)}. \tag{3.38}$$

(2). Gain for PC calibrated estimator with respect to the Horvitz-Thompson estimator

$$Gain = \frac{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_{PC}^{(b)} - \theta\right)^2}{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_{HT}^{(b)} - \theta\right)^2}. \tag{3.39}$$

(3). Relative bias (RB) for PC calibrated estimators

$$RB = \frac{\sum_{b=1}^{B}\hat{\theta}^{(b)}/1000 - t_y}{t_y}. \tag{3.40}$$

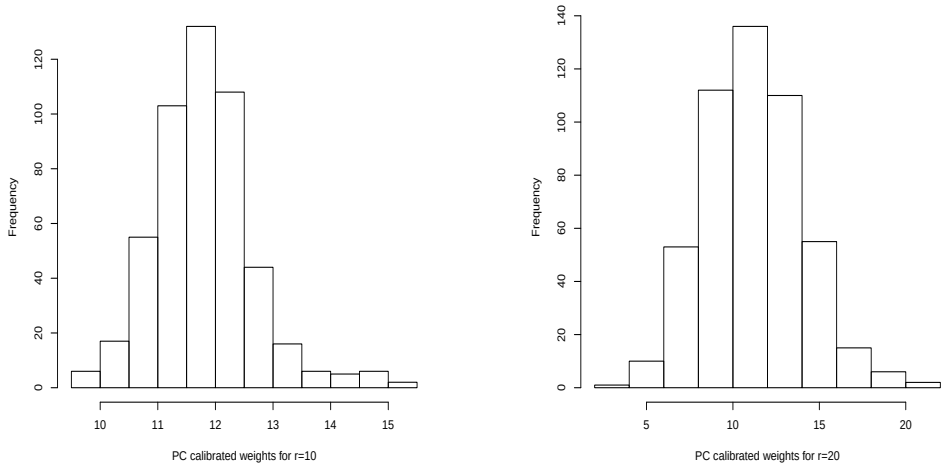
(4). Relative error (RE) for PC calibrated estimators

$$RE = \frac{\frac{1}{B}\sum_{b=1}^{B}\hat{t}_{PC}^{(b)} - t_y}{t_y}. \tag{3.41}$$

We considered two simulation cases for each estimator to check its performance,

(i). PC estimator for fixed sample size (n=500 and n=1000) and variable number of PC's (r=1,5,10,15,20,25,30,35,40,45,47)

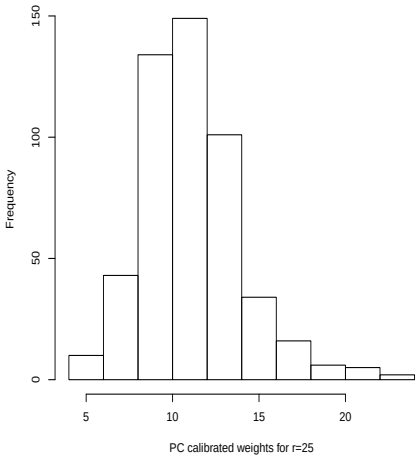
(ii). PC estimator for variable sample size (n=250,500,750,1000,1250,1500,1750,2000) and fixed number of PC's (r=25).



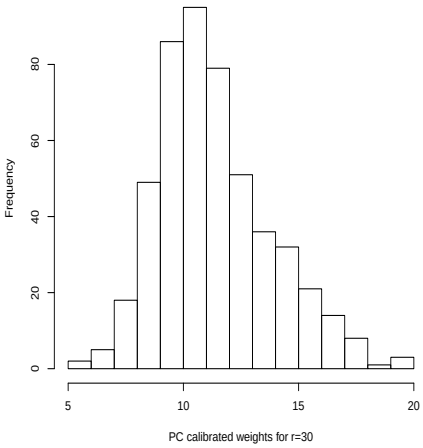
a. r=10 b. r=20

Figure 3.1: PC calibrated weights for n=500

We drew histograms for the PC calibrated weights for different number of PC's at $n = 500$. For $r = 20$ (figure 3.1(b)), the division of PC calibrated weights is more symmetrical compared to $r = 10$ (figure 3.1(a)) or $r = 25$ (figure 3.2(a)). We also sketched histograms for the ratio between the PC calibrated weights and Horvitz Thompson weights for different number of PC's. For $r = 10$ (figure 3.4(a)), the interval between the minimum and maximum limits of the ratio is 0.5 which increases to 1.0 for $r = 20$ (figure 3.4(b)), 1.5 for $r = 25$ (figure 3.5(a)), 2.0 for $r = 40$ (figure 3.6(a)) and 2.0 for $r = 45$ (figure 3.6(b)). The exception emerged for $r = 30$ (figure 3.5(b)), where the respective interval between the lower and upper limit remains 1.2.

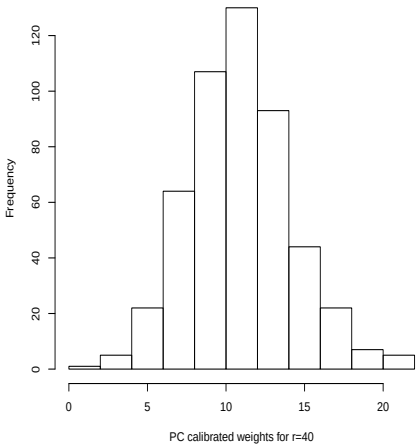


a. r=25

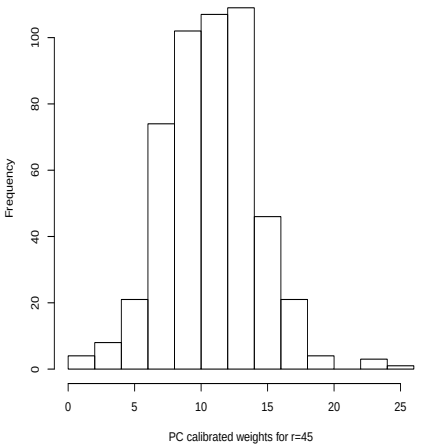


b. r=30

Figure 3.2: PC calibrated weights for n=500

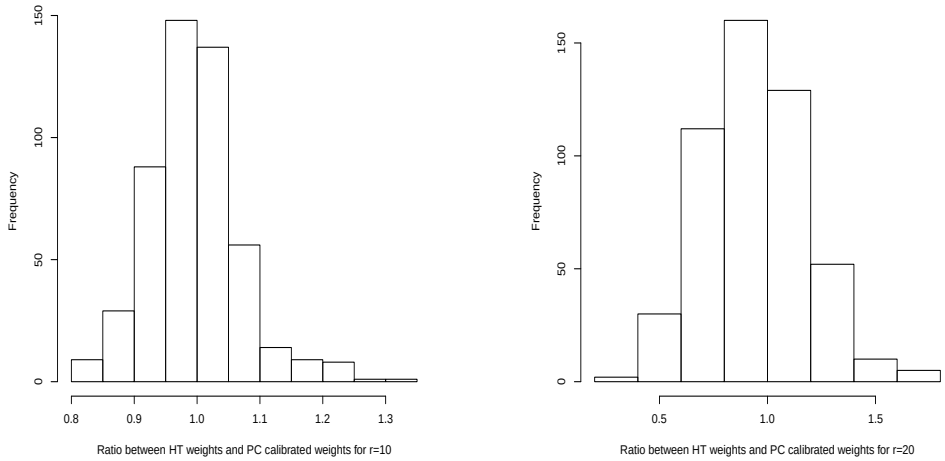


a. r=40



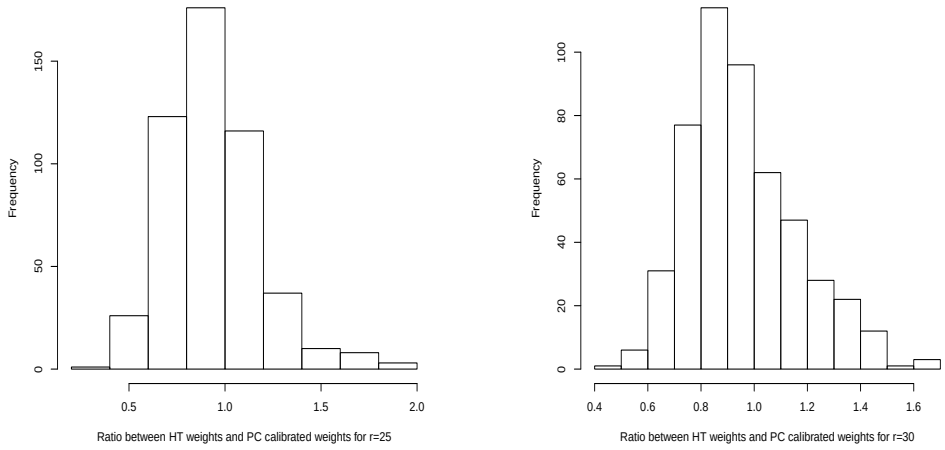
b. r=45

Figure 3.3: PC calibrated weights for n=500



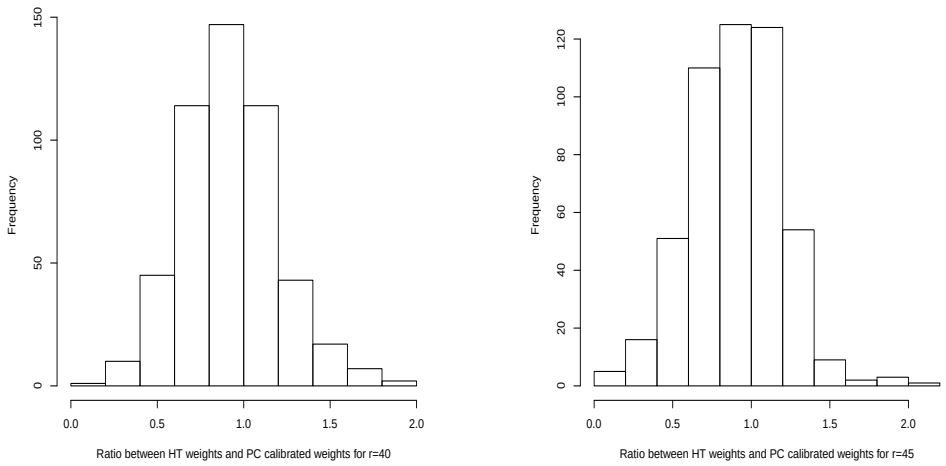
a. r=10 b. r=20

Figure 3.4: Ratio between PC calibrated weights and HT weights for n=500



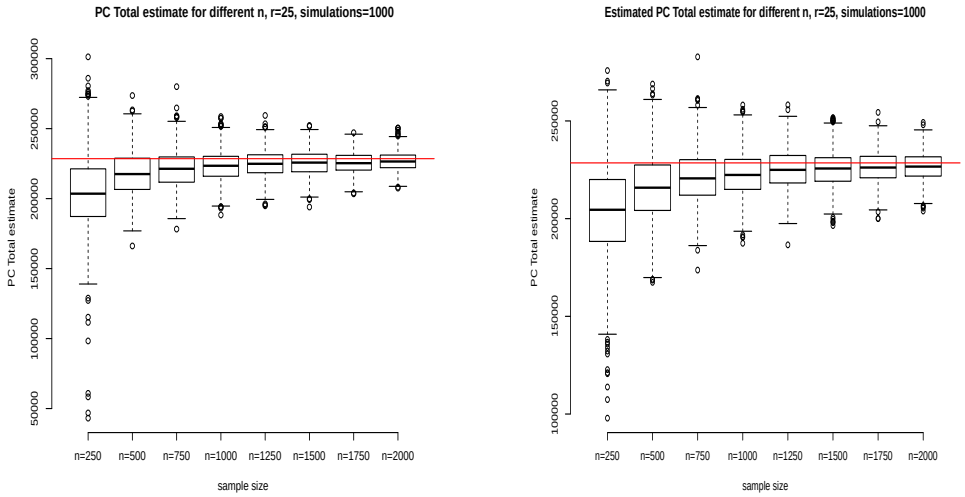
a. r=25 b. r=30

Figure 3.5: Ratio between PC calibrated weights and HT weights for n=500



a. r=40 b. r=45

Figure 3.6: Ratio between PC calibrated weights and HT weights for n=500

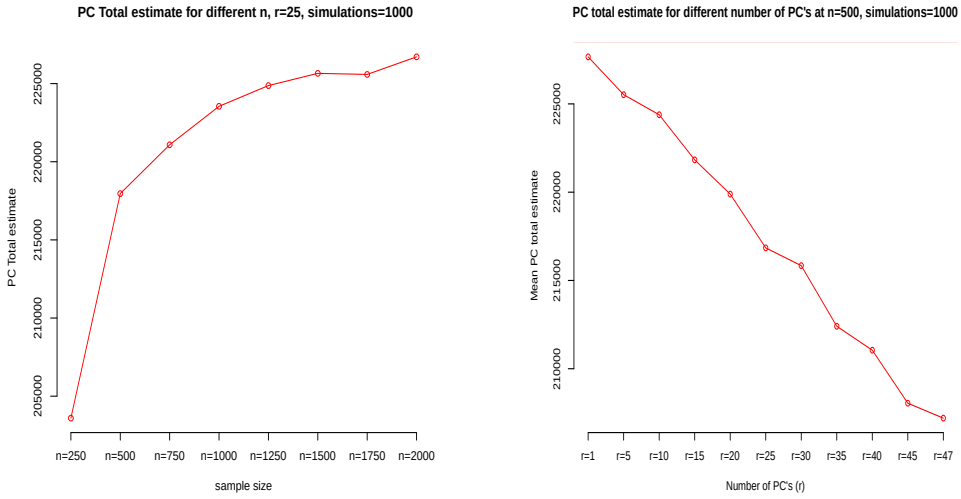


a. True b. Estimated

Figure 3.7: PC Total estimate for different sample size n, r=25, simulations=1000

The boxplot total estimates for different sample size (see figures 3.7, 3.8) tends towards the true value of the variable of interest with the increase in sample size. The red line passes through the true value of the population total $t_y = 228537.6$. In both cases, we can see that with the increase in sample size, the distance between the true and estimated value diminishes. However, the median value of estimates remains smaller than the true value despite of being very close. For example for $n = 500$, the mean total estimator $\hat{t}_{pc} = 217962.5$ and for $n = 1000$, it is, $\hat{t}_{pc} = 223547.04$. This hints us that our estimator performs an under-estimation which is even serious for the smaller sample size. On the other hand as we increase the number of PC's, the mean estimate of the $\hat{t}_{PC}$ lower than the true value $ty = 228537.6$. So, we can conclude that for smaller sample size and increasing the number of PC's after a certain number, our estimator $\hat{t}_{PC}$ tend to under-estimate the population total (see table 3.1).

The figures (3.9(a) and 3.10(a)) indicate that the increase in the sample size results in the fall of the coefficient of variation for the PC weights. A higher value for the mean coefficient of variation, 0.34 at $n = 250$ drops down to 0.22 at $n = 500$



a. Variable sample size n

b. Variable number of PC's r

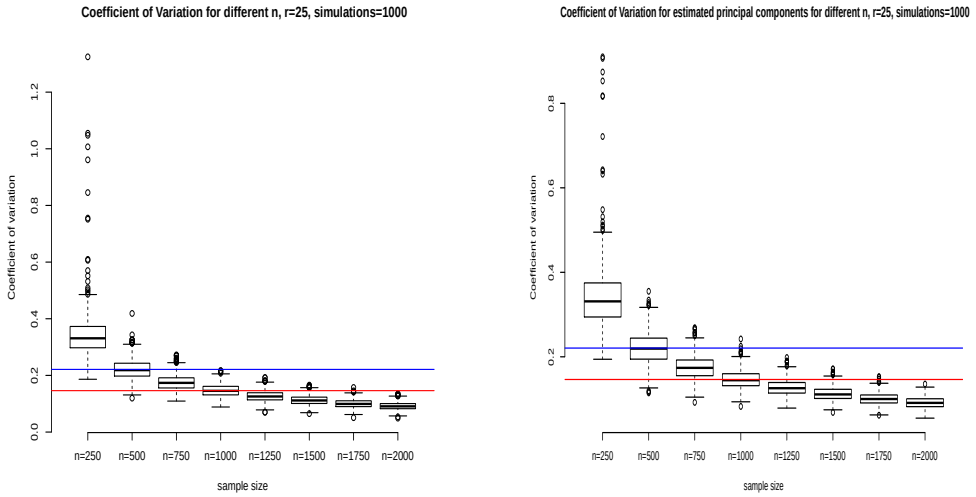
Figure 3.8: PC Total estimate

Table 3.1: PC Estimates, r=25

Estimator	n=500	n=1000
$\hat{t}_{PC}$	217962.5	223547.4
$\hat{t}_{pc,est}$	216020.1	222631.8

and further stables to 0.15 at $n = 1000$. The blue and red lines in figures 3.9(a) and 3.9(b) represent the value for the mean coefficient of variation at $n = 500$ and $n = 1000$ respectively and notably the respective values are somewhat similar in both cases. This pattern continues and for $n = 2000$, the mean coefficient of variation turns to 0.09. For the estimated PC's, the PC estimator performs slightly better than the estimator drawn from the original PC's (see figures 3.9(b),3.10(b)). The blue and green lines in figure 3.10(b) pass through the mean coefficient of variation at $n = 500$ and $n = 1000$ respectively.

On the other hand, in case of increasing the number of principal components (see figures 3.11(a), 3.12(a) and 3.13(a)), the trend is inverse. That is, increase in the number of principal components also increases the coefficient of variation



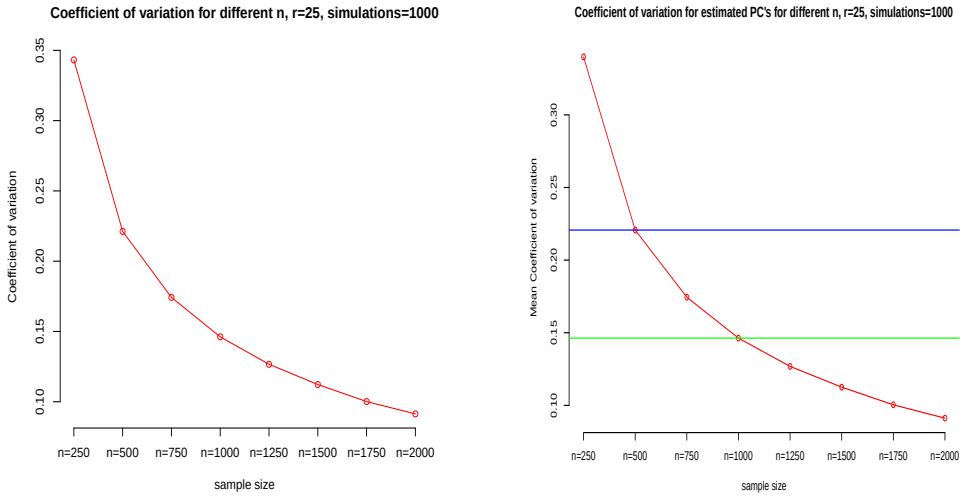
a. True b. Estimated

Figure 3.9: Coefficient of variation for different sample size

and the same is true for the estimated PC calibrated weights (see figures 3.11(b), 3.12(b) and 3.13(b)).

This reveals that our PC estimator of population total t_y performs better with the increase in the sample size but due to the cost issue we may be restricted to a rather smaller sample size (say $n = 500$ or $n = 1000$) which also give us reasonable reduction in the coefficient of variation. In the second case of variable number of principal components, we see that for $r = 25$ (which takes almost 84% of the variation into account) at $n = 500$ the mean value (0.22) of coefficient of variation is relatively higher as compared to the mean coefficient of variation (0.14) case when $n = 1000$ (see tables 3.2 and 3.3). Another important fact comes out is that when we estimate our principal components, the variability in terms of the coefficient of variation is a bit lesser than the PC estimator for population principal components.

This may be due to the fact that we estimated the PC's, so the use of smaller set (sample) of the standardized observations is for the their estimation also reduced the variability. The difference in performance in terms of the coefficient of variation



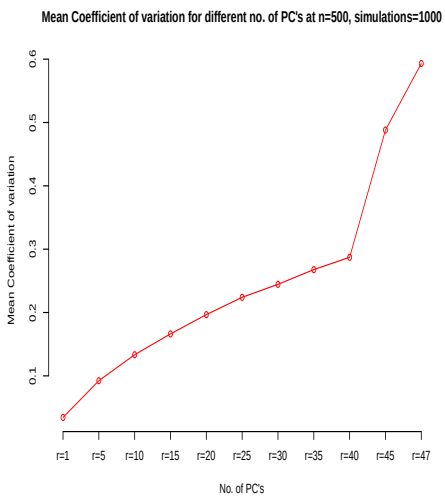
a. True b. Estimated

Figure 3.10: Coefficient of variation for different sample size

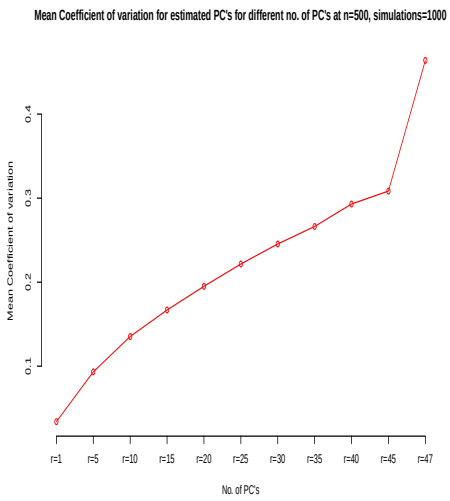
(variability) gets more clear for smaller sample size and large number of PC's.

The gain, which in fact is the ratio of the variance of $\hat{t}_{PC}$ and the variance of $\hat{t}_{y\pi}$, decreases with the increase in the sample size. It shows the relative benefit we gain with respect to the Horvitz-Thompson estimator $\hat{t}_{y\pi}$.

The figure 3.14 shows that the larger sample, the smaller value for the ratio of the variance (gain). This means that as we increase our sample size, the benefit increases with respect to the $\hat{t}_{y\pi}$ estimator in terms of the gain. This, however is different for the change in the number of the PC's. That is, as we increase the number of PC's, the value of gain also increases hence our estimator becomes less efficient for larger number of PC's (see figures 3.15(a) and 3.16(a)). For $n = 500$, the gain value goes more than 1 after $r = 25$, making our proposed estimator $\hat{t}_{PC}$ less efficient. For $n = 1000$, the gain value remains under 1 even for the maximum number of the PC's, hence advocates the efficiency for the larger sample.

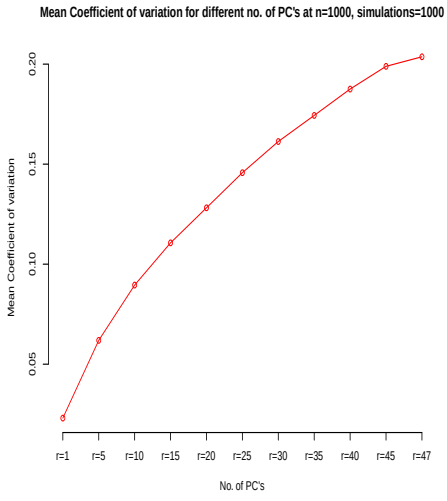


a. True

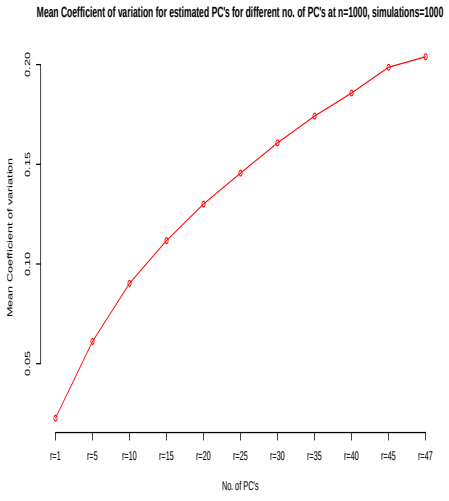


b. Estimated

Figure 3.11: Coefficient of variation for different number of PC's, n=500

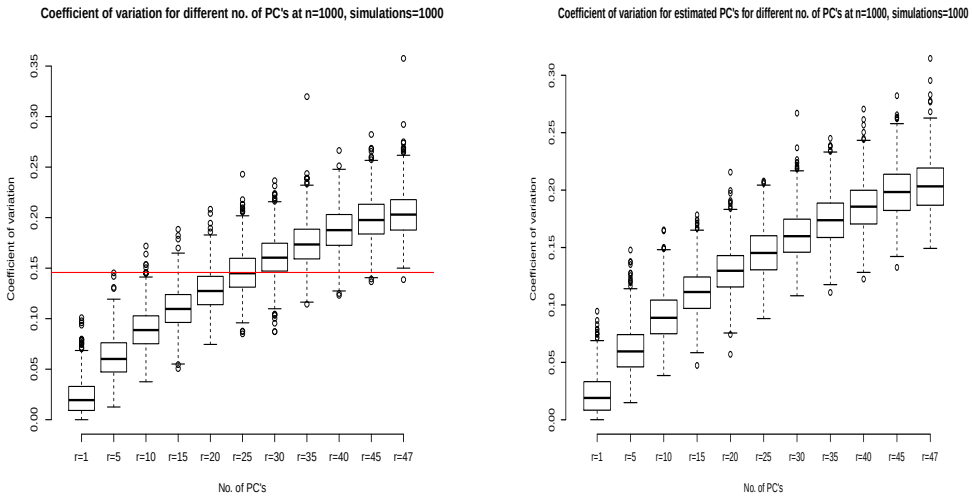


a. True



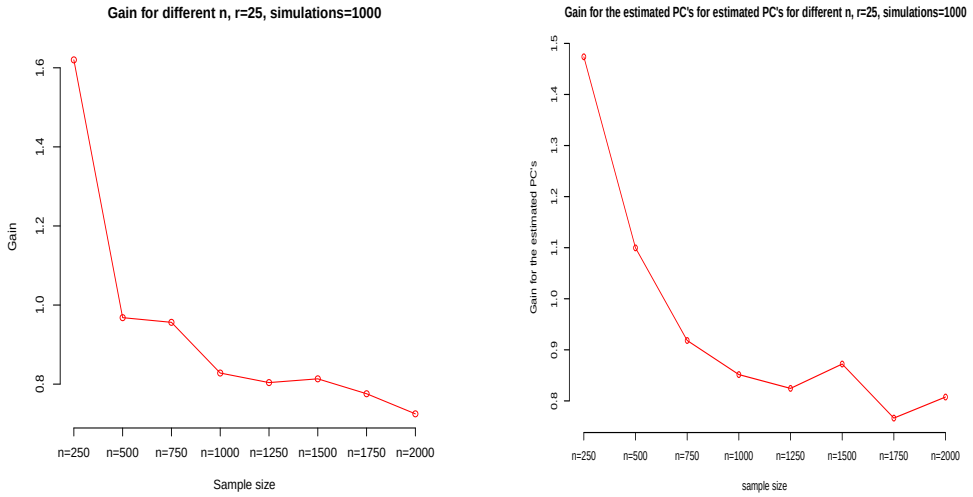
b. Estimated

Figure 3.12: Coefficient of variation for different number of PC's, n=1000



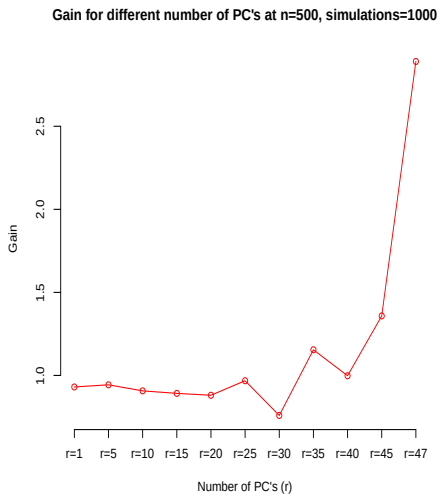
a. True b. Estimated

Figure 3.13: Coefficient of variation for different number of PC's, n=1000

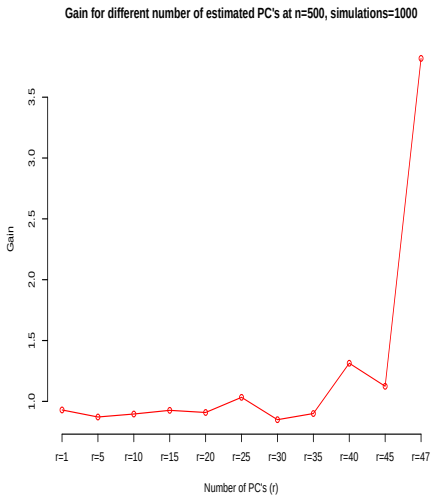


a. True b. Estimated

Figure 3.14: Gain for different sample size, r=25

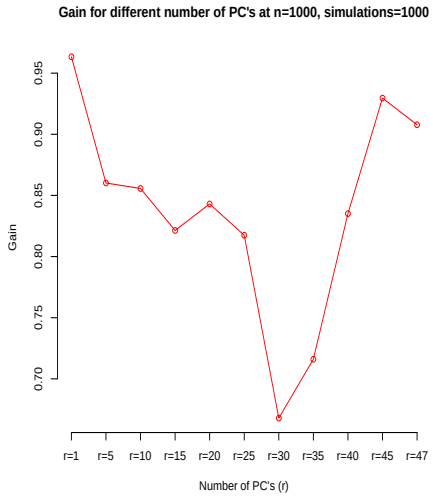


a. True

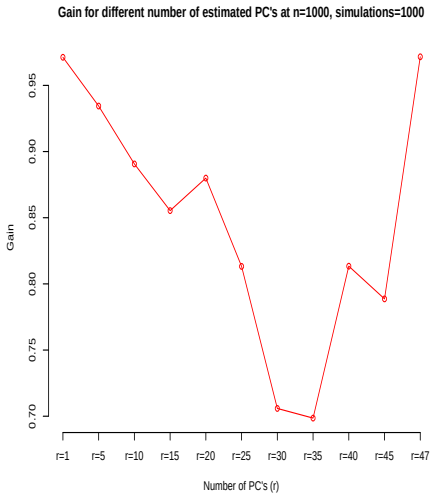


b. Estimated

Figure 3.15: Gain for different number of PC's, n=500

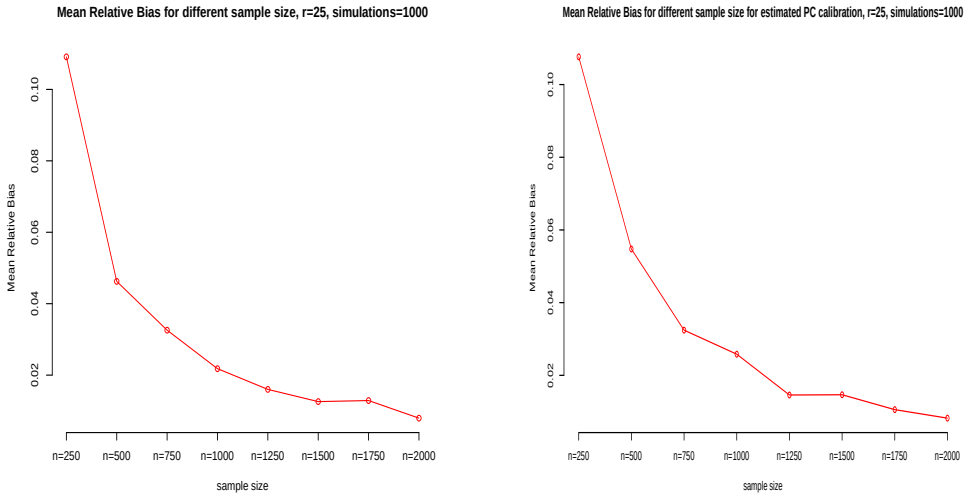


a. True



b. Estimated

Figure 3.16: Gain for different number of PC's, n=1000

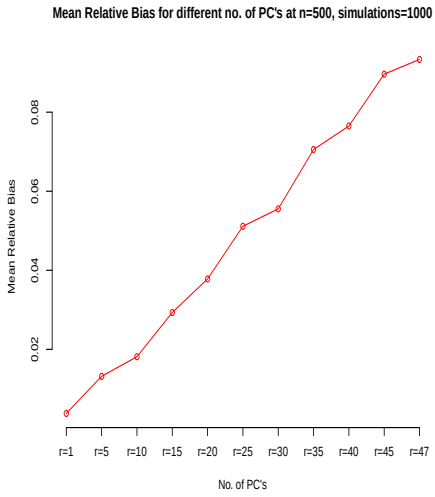


a. True b. Estimated

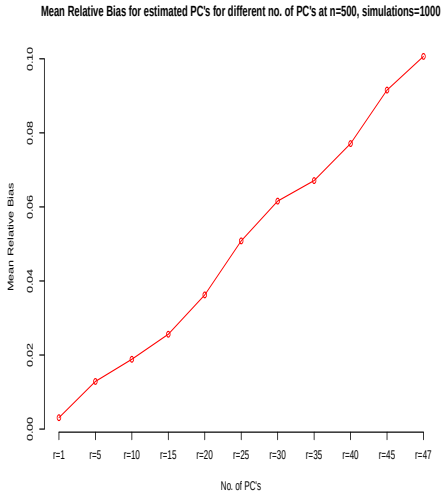
Figure 3.17: Relative bias for different sample size

However, in case of the estimated PC's, the gain value is slightly on the higher side both for $n = 500$ and $n = 1000$. Our estimator $\hat{t}_{PC}$ has the relative bias for $n = 500$ almost 5.1% which is slightly lower as 5% for $\hat{t}_{pc,est}$ (see table 3.2). For $n = 1000$, the relative bias goes from almost 2.1% to 1.9% (see table 3.3) for $\hat{t}_{PC}$ and $\hat{t}_{pc,est}$ respectively which is not huge.

Similarly, the relative error also follows the trend as of the relative bias, gain and coefficient of variation. That is, as n increases, the relative error decreases (see figure 3.22) and as r increases, the relative error also increases after a certain value of r (see figures 3.23, 3.24). However, for the estimated $\hat{t}_{PC}$, the relative error for $\hat{t}_{pc,est}$ is slightly higher than of the $\hat{t}_{PC}$ (see tables 3.2 and 3.3). For $n = 500$, $RE = 7\%$ which is a bit higher and for $n = 1000$, $RE = 4\%$.

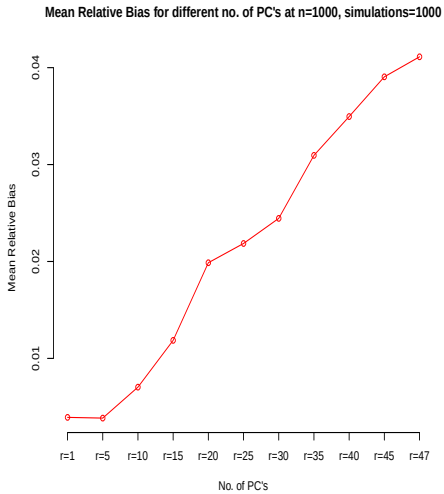


a. True

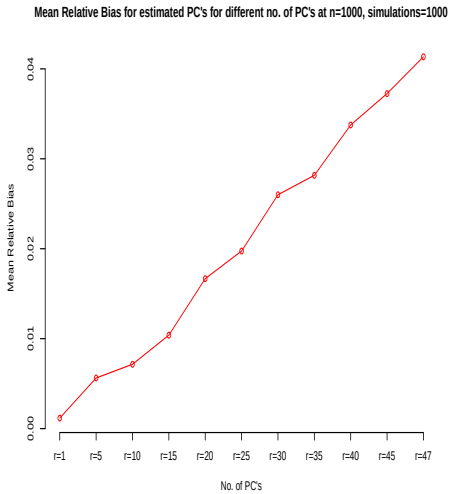


b. Estimated

Figure 3.18: Relative bias for different number of PC's, n=500

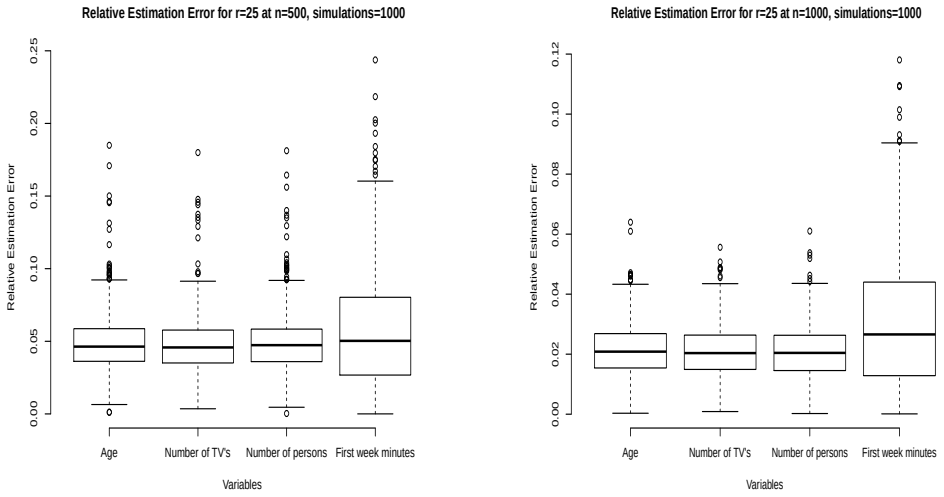


a. True



b. Estimated

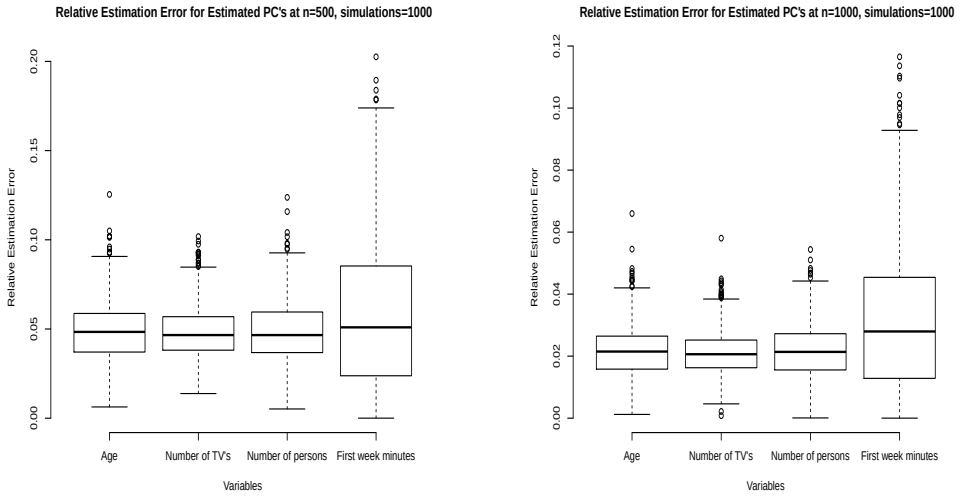
Figure 3.19: Relative bias for different number of PC's, n=1000



a. n=500 b. n=1000

Figure 3.20: Relative Estimation error at r=25

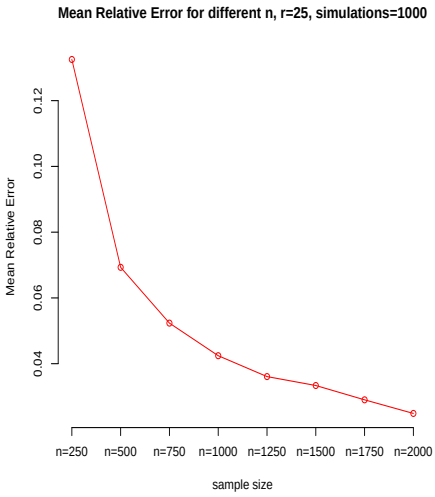
Although relative error is on higher side but it is understandable because $r = 25$ is almost half of the total number of PC's available. Now, how to decide what number of PC's should be taken and what should be the sample size? In our case, first 25 PC's ($r = 25$) take almost 85% of the variation into account and has coefficient of variation of 2% and 1.5% for $n = 500$ and $n = 1000$ respectively. Similarly, the data dimension is reduced from 49 variables to the 25 and yet conceded only 7% of the relative error and coefficient of variation is also on the lower side. The benefit in terms of the gain is also higher. So our calibration weights using the PC's is doing well even for a small sample size $n = 500$ which is almost 8% of the total population. We, then applied our calibration weights to estimate the known totals for some of the original auxiliary variables **Age ($\mathbf{X}_{40}$)**, **Number of T.V.'s in a house ($\mathbf{X}_{47}$)**, **Number of persons in a house ($\mathbf{X}_{48}$)** and **First week watched minutes ($\mathbf{X}_{49}$)**. We wanted to verify that how far the weights estimate the $\mathbf{X}$ totals. Relative estimation errors are found for these variables and compared between them for $n = 500$ and $n = 1000$ (see figures 3.20 and 3.21).



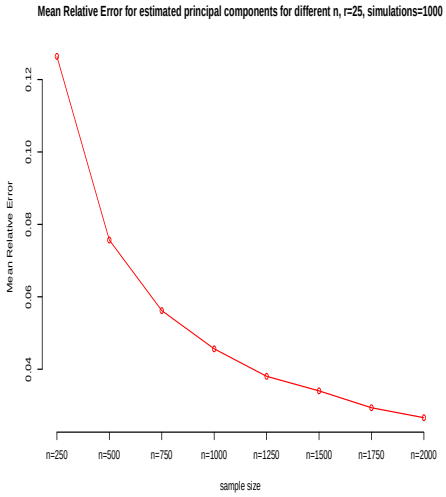
a. n=500 b. n=1000

Figure 3.21: Relative Estimation error for estimated PC's at r=25

Clearly we can see that for all the four variables, the relative estimation error is lower for the estimator using estimated PC's at $n = 500$. For example for Age, the scatter of the values of the estimation error goes up to almost 20% (figure 3.20(b)) for population PC estimator but it remains well under 15% for the estimated PC estimator. For the No. of T.V.'s in a house ($(\mathbf{X}_{47})$), the difference is even more clear. For population PC estimator is almost 19% and for estimated PC estimator it remains up to 10%.

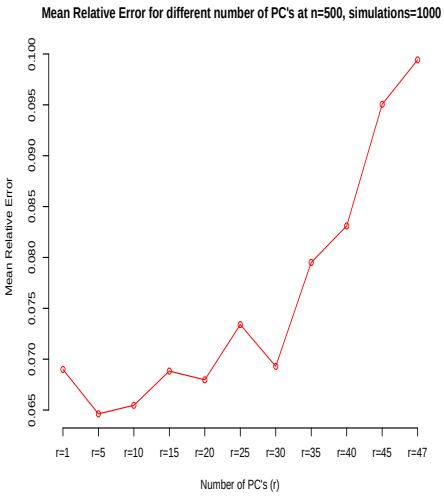


a. True

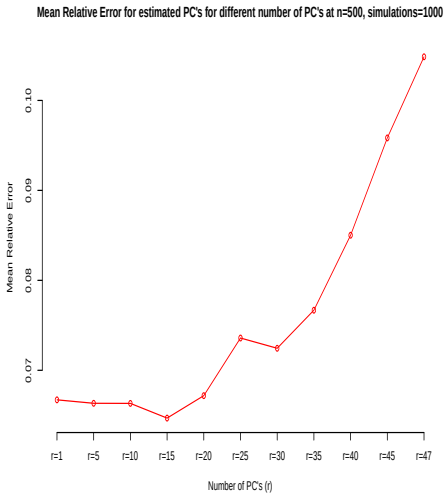


b. Estimated

Figure 3.22: Relative error for different sample size

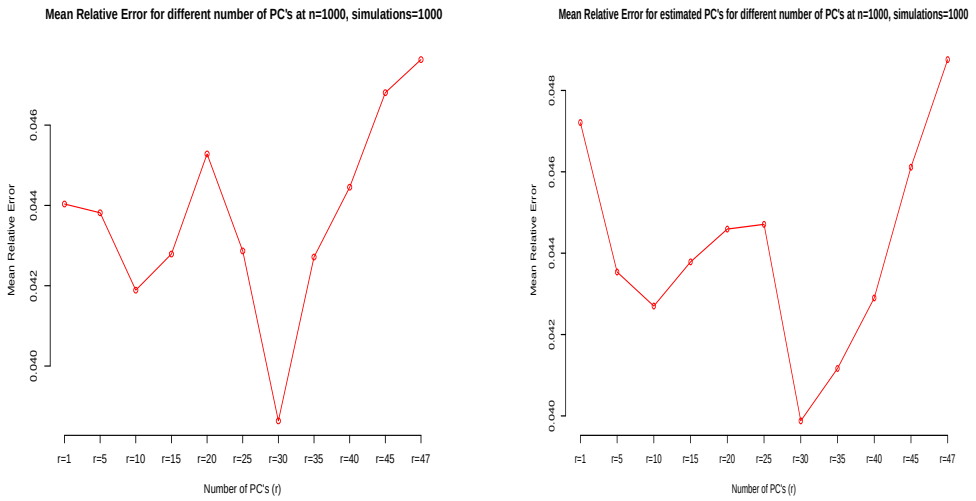


a. True



b. Estimated

Figure 3.23: Relative error for different number of PC's, n=500



a. True

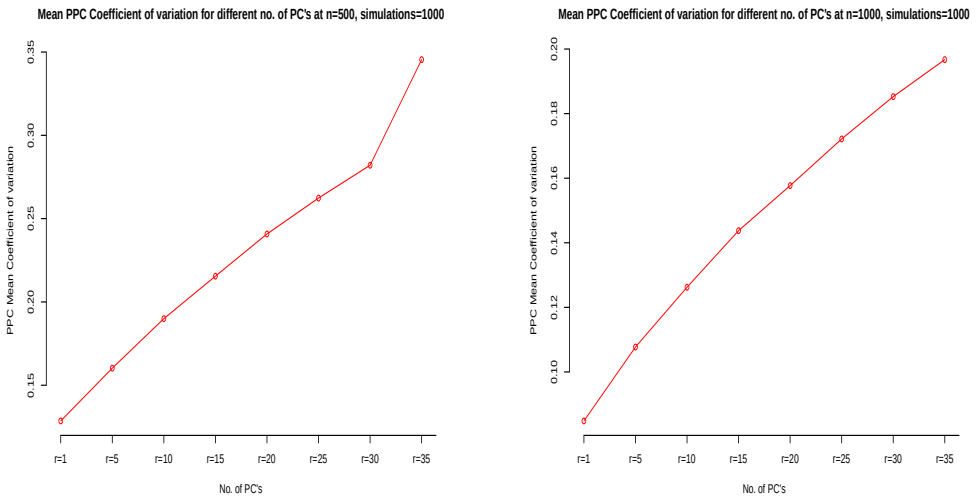
b. Estimated

Figure 3.24: Relative error for different number of PC's, n=1000

Similarly for Number of persons in a house ($\mathbf{X}_{48}$), for estimated PC estimator, the maximum relative estimation error remains almost 12% as compared to about 18% of the population PC estimator. For the variable *First week minutes*, the limits remain 25% and 20% for population PC estimator and estimated PC estimator respectively. Again for a large sample size $n = 1000$, the relative estimation error almost cuts off into half and are not so different between them. Thus we can say that as the sample size increases, the estimation error for the estimator using population PC and estimated PC become less distant.

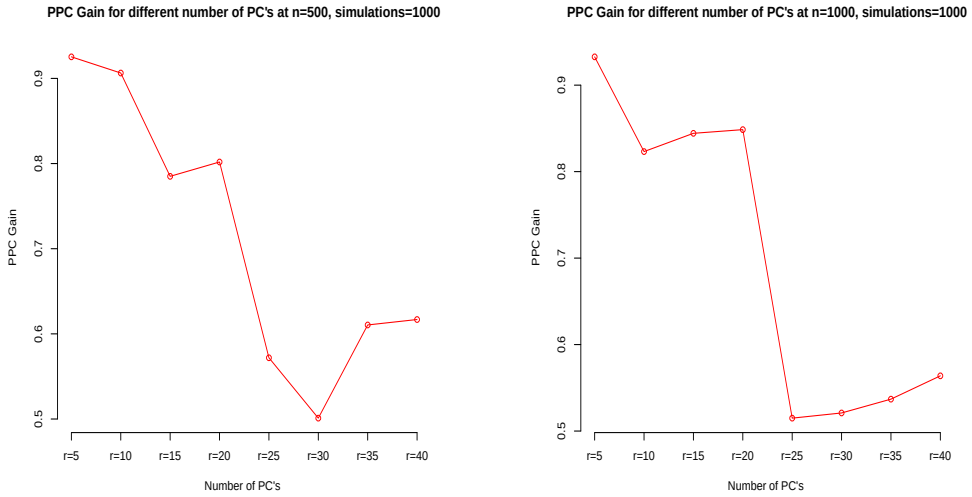
Table 3.2: Performance of PC estimator, n=500

r=25	n=500	Estimated PC n=500
Mean coefficient of variation	0.22	0.22
Mean gain	0.97	1.01
Mean relative bias	0.05	0.05
Mean relative error	0.07	0.08



a. n=500 b. n=1000

Figure 3.25: PPC Coefficient of Variation for different number of PC's



a. n=500 b. n=1000

Figure 3.26: PPC Gain for different number of PC's

Table 3.3: Performance of PC estimator, n=1000

r=25	n=1000	Estimated PC, n=1000
Mean coefficient of variation	0.15	0.15
Mean gain	0.83	0.85
Mean relative bias	0.02	0.02
Mean relative error	0.04	0.04

Table 3.4: Performance of PPC estimator

r_1=24	n=500	n=1000
Mean coefficient of variation	0.24	0.1563374
Mean gain	0.85	0.754646
Mean relative bias	0.008	0.003
Mean relative error	0.06	0.04

Applying the partial principal component calibration on our media data, we partitioned our matrix $\mathbf{X}$ such that $\tilde{\mathbf{X}}_1 = (\mathbf{X}_{sex}, \mathbf{X}_{age}, \mathbf{X}_{nbtv}, \mathbf{X}_{npf})$ and $\tilde{\mathbf{X}}_2$ con-

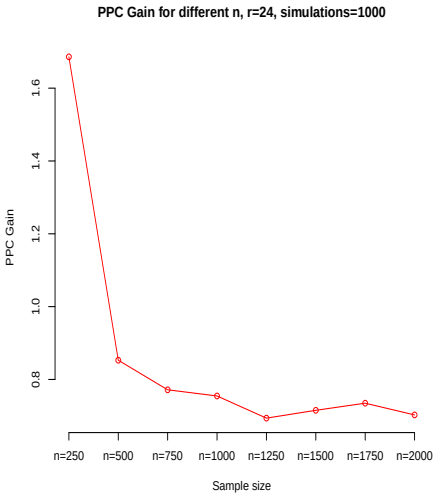


Figure 3.27: Gain for PPC

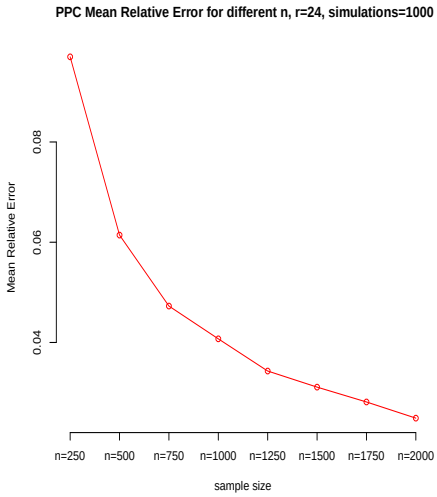


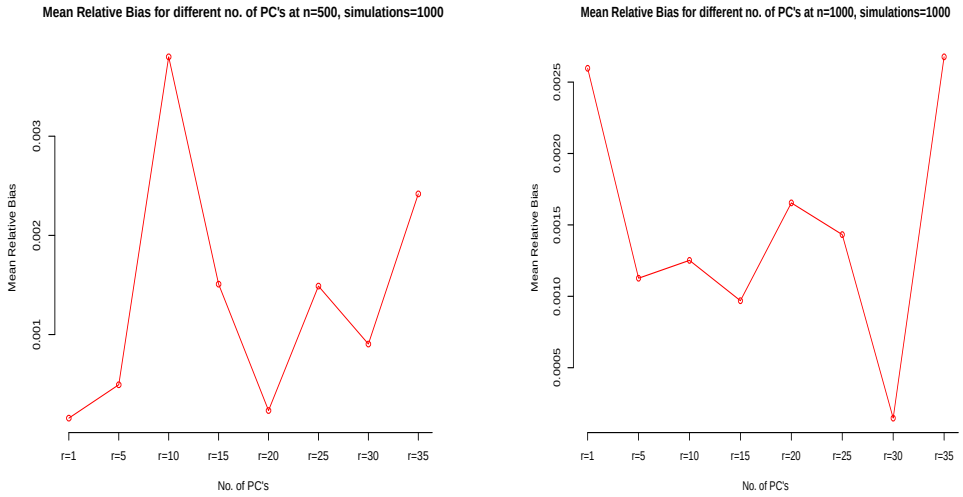
Figure 3.28: Relative Error for PPC

Table 3.5: Original Total vs PC estimates for different n

Variables	Original Total	n=500	n=1000
Type.menag	3749	3566.283	3666.612
CSP	5901	5619.041	5773.36
Internet	5928	5644.765	5800.479
Enfants	2819	2683.061	2758.391
First Week Minutes	1703739	1616919	1666277

Table 3.6: Original Total vs Estimated PC estimates for different n

Variables	Original Total	n=500	n=1000
Type.menag	3749	3570	3668.009
CSP	5901	5619	5774.971
Internet	5928	5644.825	5801.74
Enfants	2819	2687.525	2758.287
First Week Minutes	1703739	1614463	1666277



a. n=500 b. n=1000

Figure 3.29: PPC Relative Bias for different number of PC's

Table 3.7: Original Total vs PPC estimates for different n

Variables	Original Total	n=500	n=1000
Type.menag	3749	3740.857	3745.39
CSP	5901	5891.604	5897.596
Internet	5928	5919.314	5924.652
Enfants	2819	2817.28	2818.324
First Week Minutes	1703739	1695369	1700725

tained the rest of variables. The principal component matrix $\tilde{\mathbf{Z}}_2$ from $\tilde{\mathbf{X}}_2$ is computed and first 24 principal components associated to the largest 24 eigenvalues which account for almost 85% of the total variation. That is,

$$\tilde{\mathbf{Z}}_{2r_1} = (\mathbf{Z}_{2(1)}, \ldots, \mathbf{Z}_{2(24)})$$

and therefore the partial principal component matrix $\mathbf{M}$ is a (5930×29) matrix such that,

$$\mathbf{M} = (\mathbf{X}_6, \mathbf{X}_7, \mathbf{X}_{40}, \mathbf{X}_{47}, \mathbf{X}_{48}, \mathbf{Z}_{2(1)}, \ldots, \mathbf{Z}_{2(24)}).$$

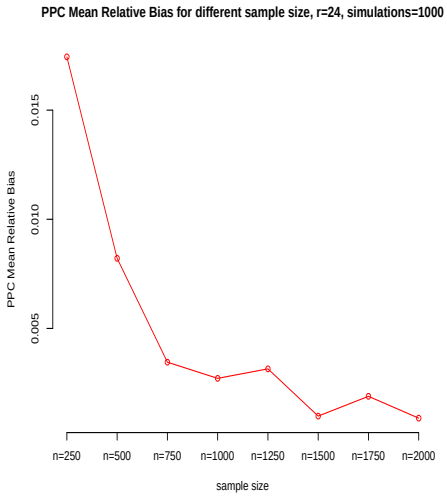


Figure 3.30: Relative Bias for PPC

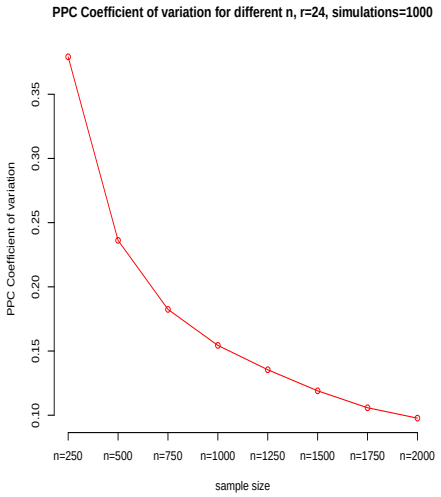
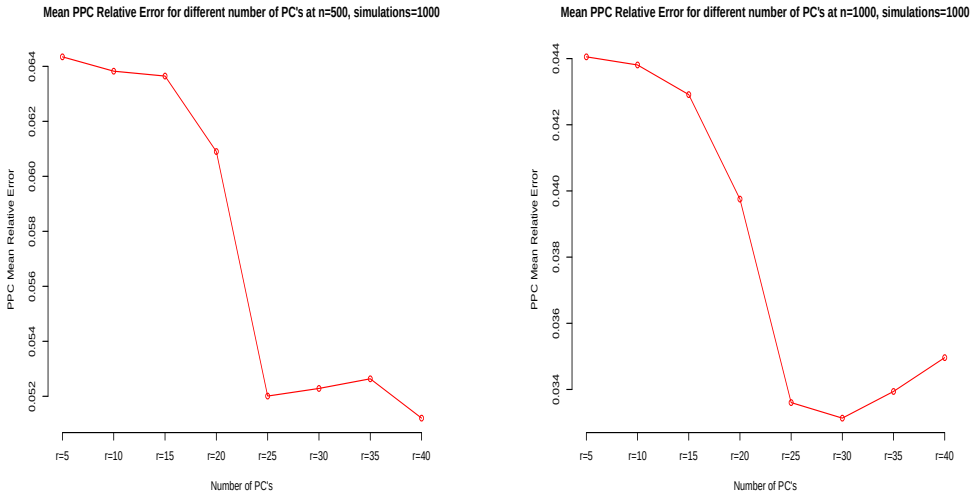


Figure 3.31: C.V. for PPC

Simulation study similar to the PC and estimated PC calibration is done. C.V., relative gain, relative bias and relative error are calculated to evaluate the performance of the PPC calibration estimator. Comparing the figure (3.31) with figure (3.10) and figure (3.25) with figures (3.11 and 3.12), the coefficient of variation for partial principal component (PPC) calibration, we can say that it follows the same trend as for the coefficient of variation for PC and estimated PC calibration estimators. Referring to the tables 3.2, 3.3 and 3.4, we can see that there is not much difference between the coefficient of variations between the different calibration methods. So the coefficient of variation is almost similar to the previous calibration methods.

The gain for PPC calibration for different sample size (figure 3.27) attains the trend identical to the PC and estimated PC calibration estimators (figure 3.14). However, for the variable number of r_1, the gain curve for PPC (figure 3.26) depicts smaller values than the PC and estimated PC calibration estimators (figures 3.15, 3.16). For $n = 500$, the PPC gain is 0.8528435 (table 3.4) which is almost similar to the gain for the PC and estimated PC at $n = 1000$ (table 3.3). The improvement in the estimation procedure due to the PPC calibration in terms



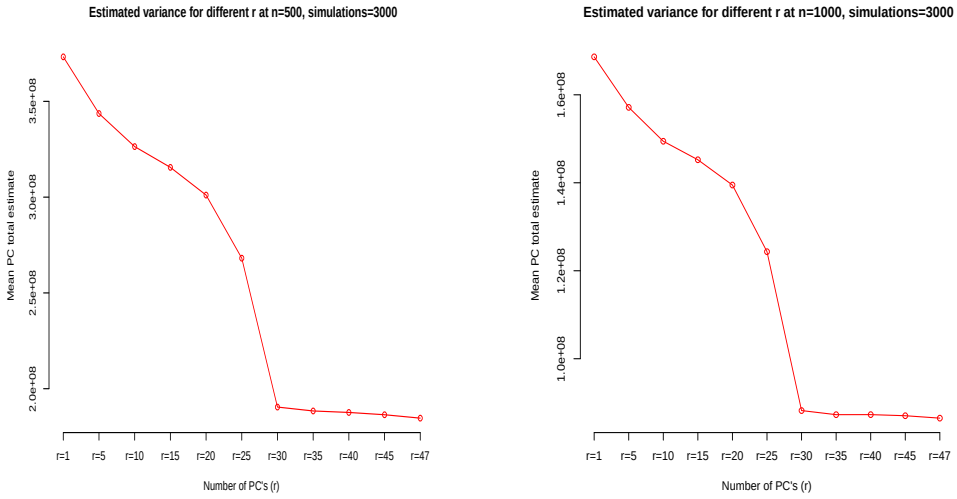
a. n=500 b. n=1000

Figure 3.32: PPC Relative Error for different number of PC's

of the relative gain is more prominent for $n = 1000$.

The relative bias for the PPC calibration method attains serious improvement (compare figures 3.17 and 3.30). That is, for PPC calibration at $n = 500$ the relative bias is less than 0.8% (table 3.4) compared to 4.5% for the PC and estimated PC calibration (table 3.2). This is due to the inclusion of the $\mathbf{M}$ which contains the original $\tilde{\mathbf{X}}_1$ variables and hence reduces their part of bias. It also shows the greater importance of the variables included in $\tilde{\mathbf{X}}_1$. We, however may not put any variable of $\mathbf{X}$ in $\tilde{\mathbf{X}}_1$. We tried some other variables which seemingly were important to be calibrated exactly but they resulted in the singularity problem. So, the choice of the $\tilde{\mathbf{X}}_1$ variables in the partial principal component calibration matrix $\mathbf{M}$ may need some work to do.

For $n = 500$, the relative error for the PPC calibration estimator (figure 3.28, 3.32 and table 3.4) is slightly lower than for the PC and estimated PC calibration estimators (figures 3.22, 3.23 and table 3.4). For the PPC calibration, the relative error decreases with the increase in n. For the variable r_1 (figure 3.32), the relative error increase with the increase in PPC up to $r_1 = 24$, then it starts increasing



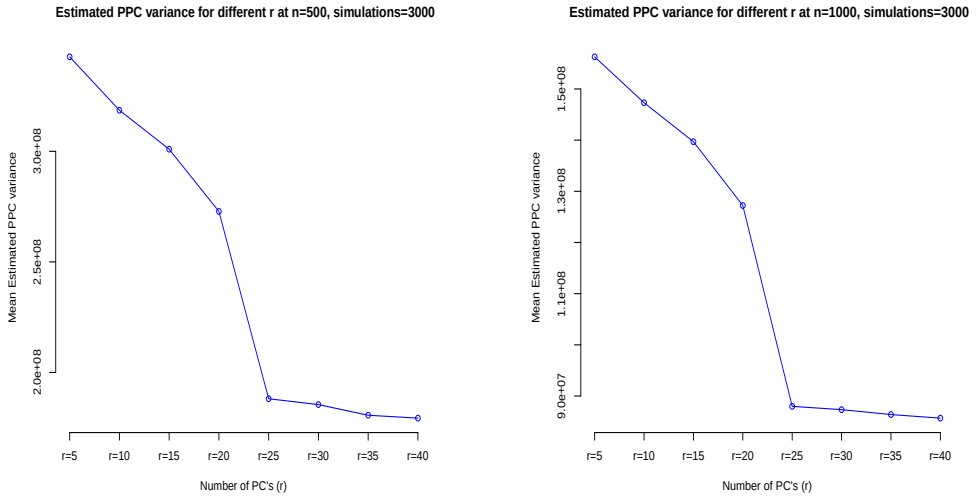
a. n=500 b. n=1000

Figure 3.33: PC Variance for different number of PC's

and then gets stable.

In the next step, we used the PPC calibration weights w_{ppc} to calibrate the totals of the original auxiliary variables $\mathbf{X}$ to see that how much estimation error is caused by the use of w_{ppc}. The four variables $\mathbf{X}_{sex}, \mathbf{X}_{age}, \mathbf{X}_{nbtv}$, and $\mathbf{X}_{npf}$ which were used in the matrix $\mathbf{M}$ are exactly calibrated by the partially principal component weights w_{ppc}. We calculated total estimators for $\mathbf{X}_{type_menag}$, $\mathbf{X}_{csp}$, $\mathbf{X}_{internet}$, $\mathbf{X}_{enfants}$ and $\mathbf{X}_{(first.week.minutes)}$ for $n = 500$ and $n = 10000$ using w_{pc} (table 3.5) , $w_{ppc,est}$ (table 3.6) and w_{ppc} (table 3.7) and compared them with their original totals.

Clearly, we can see that the results for w_{pc} (table 3.5) and $w_{pc,est}$ (table 3.6) are almost similar and a bit far from their original totals. But the totals estimated using w_{ppc} are significantly close to their respective original totals. The relatively lower relative estimation error (figures 3.38(a) and 3.38(b)) for w_{ppc} further clears the picture that the estimation of the population totals using PPC calibration gives better results than the PC calibration and the estimated PC calibration (figure 3.20).



a. n=500 b. n=1000

Figure 3.34: PPC Variance for different number of PC's

Variance estimation

We then calculated the estimated variance for the PC, estimated PC and PPC estimator for different n and different r with number of simulations equal to 3000.

The estimated variance for the PC (figure 3.35) and estimated PC estimator (figure 3.36)for different sample size decreases simultaneously in almost similar pattern. The trend remain identical for the estimated variance of PPC estimator (figure 3.37). For variable number of PC's, the estimated variance goes down smoothly until $r = 25$ and then suddenly falls immensely for $r = 30$ and then smooths up to $r = 47$ (figure 3.33). The estimated variance for the PPC estimator for different r (figure 3.34) is better than that of PC and estimated PC estimator. Relative error for the variance is also calculated for different estimator. Interestingly, the RE for variance for PC estimator (figure 3.42) and PPC estimator (figure 3.43) attain the similar pattern for $n = 500$. It increases rapidly with the increase in r up to a certain level ($r = 30$ for PC estimator, figure 3.42) and then smooths from onwards. For variable n, the RE of estimated variance are not very

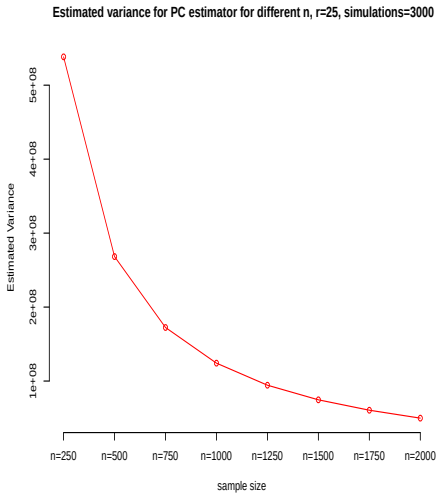


Figure 3.35: Variance for PC

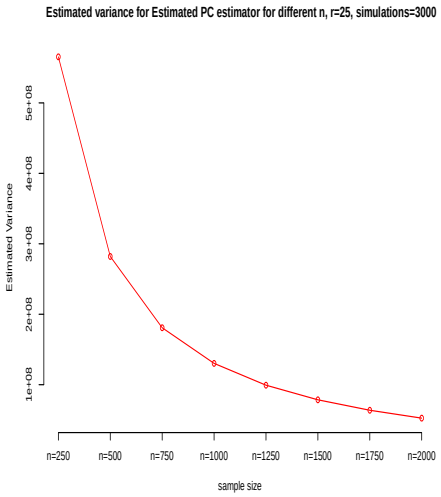


Figure 3.36: Variance for estimated PC

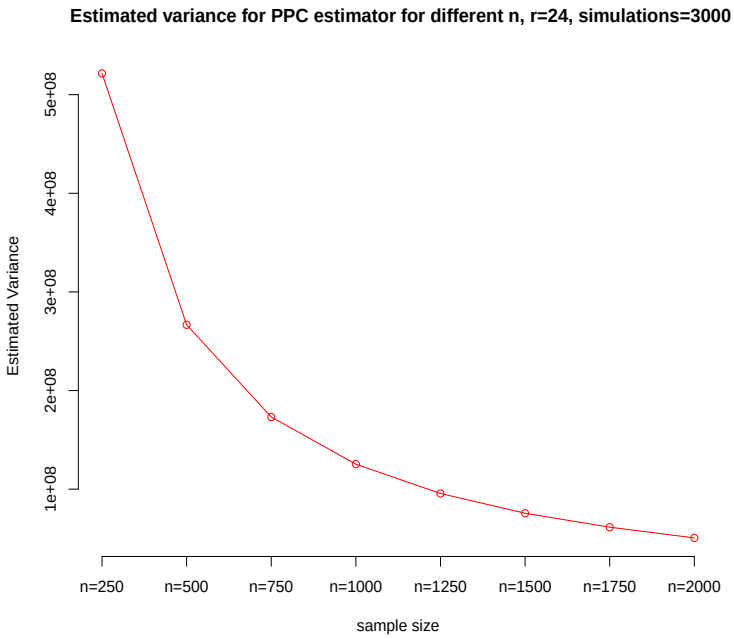
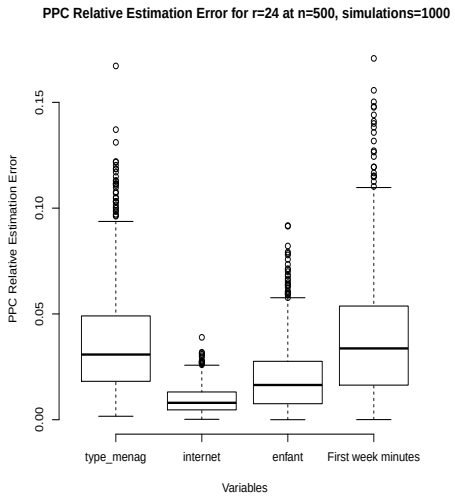
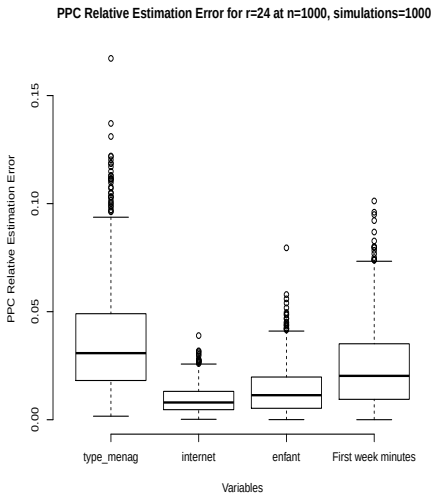


Figure 3.37: PPC Variance for different sample size

much different for PC and PPC estimator (see figures 3.40 and 3.41).

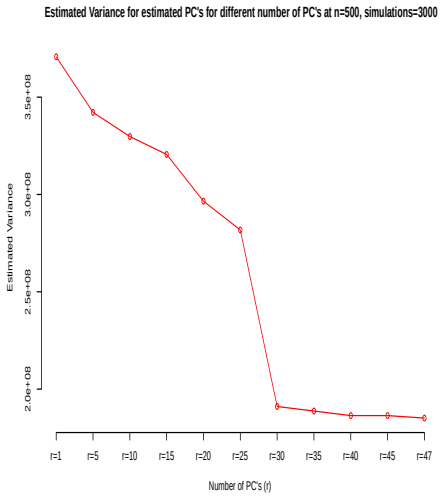


a. n=500

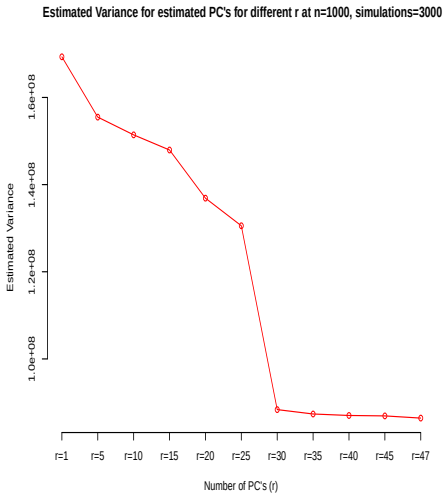


b. n=1000

Figure 3.38: PPC Estimation error for different number of PC's



a. True



b. Estimated

Figure 3.39: PC Variance for different number of PC's

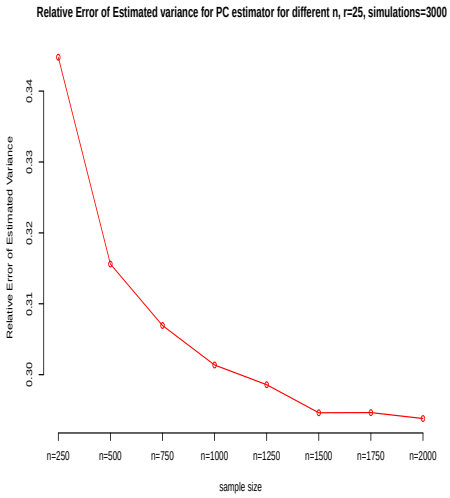


Figure 3.40: RE for PC variance for different n

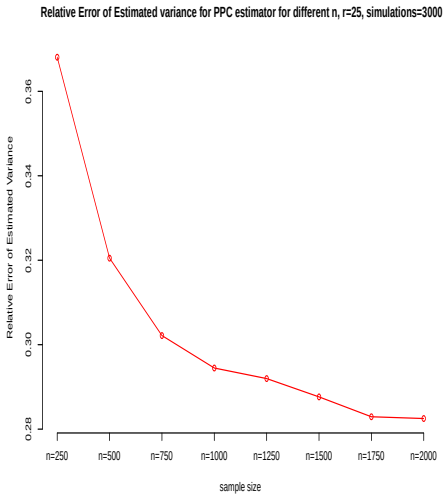


Figure 3.41: RE for PPC variance for different n

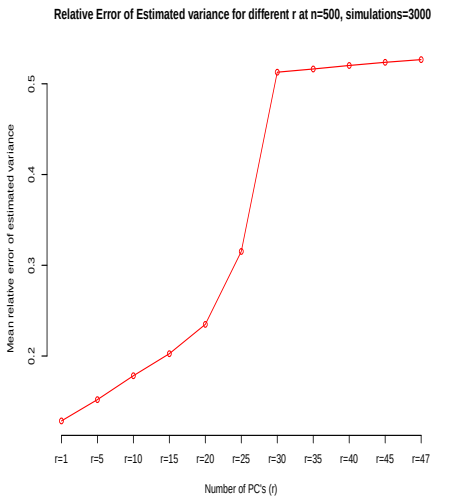


Figure 3.42: RE for PC variance for n=500

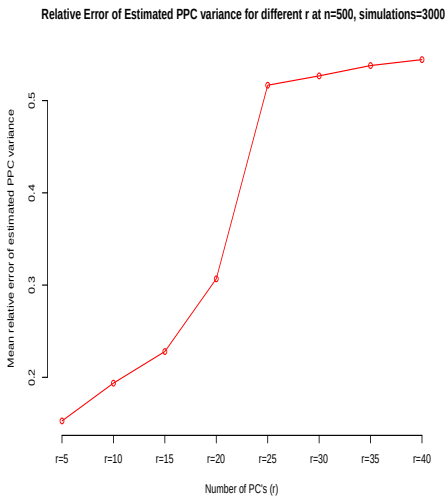


Figure 3.43: RE for PPC variance for n=500

Chapter 4

Discussion and Perspectives

This thesis report deals with the estimation of a population total when an in-hand large dimensional auxiliary data is severely ill-conditioned. Two types of methods are proposed to deal with the ill-conditioned auxiliary variables in the estimation of a population total and their variances are estimated. The first technique consists of penalizing the diagonal of the covariance matrix. A detailed overview of the different existing ridge regression solution viable in various statistical circumstances, is given in chapter 2. Estimation of regression coefficient using optimization problems in model-based and model-assisted cases are obtained and therefore used to construct the respective estimators for the population total. Similar ridge estimators are devised using a model-free approach called ridge calibration or penalized calibration and their equivalence is proved. A special case of ridge regression estimator (Bardsley and Chambers, 1984) is considered and its equivalence to the partially penalized calibration estimators (Guggemos and Tillé, 2010) is shown (proposition 2). Different interpretations of the ridge regression estimation are discussed and a link is established between them. Statistical properties are derived for ridge estimator and their improvement over least square estimator is shown. Clearly, the ridge estimator performs better than the least square estimator of the population total in terms of the MSE. We applied the ridge technique on the Mediametrie data set, which was seriously ill-conditioned

and multicollinear having a significant percentage of minimum eigenvalues zeros.

The second method studied was the principal component regression (PCR). GREG-type estimator is constructed using PC's and also different types PC calibration is introduced such as PC calibration on the second moment, partial principal component calibration and calibration using estimated PC's. Compared to the ridge regression estimator, which is a penalizing method, PC calibration is rather a dimension reduction technique (Jolliffe, 2002). Application of these methods on the Mediametrie data is done to estimate the population total of the variable of interest using the proposed PC calibration techniques and found that these techniques perform better than the Horvitz-Thompson estimator. Graphical and tabular comparisons between these PC calibration techniques are established.

The development of these methods was in fact inspired and motivated by a statistical data problem named ill-conditioning or some times multicollinearity present in Mediametrie (Paris) data available to us for different T.V. channels and we saw that our newly proposed methods gave improved results.

Although, we used only, simple random sampling without replacement (SRSWOR) as a sampling design, our approach is general and can be applied to other sampling techniques such as stratified random sampling. These methods can also be applied on other data sets such as in website data where the number of users are enormous and the estimation of total users visiting a certain website page may be of particular interest. Also in the telecommunications domain, the estimation of total number of calls made from a particular network to any other particular network may be of interest.

An interesting extension of this work may be to use a sampling design with unequal probabilities (Brewer, 1999). The variance and its convergence will be interesting to develop in this case under calibration. In this case, the total estimator and its asymptotic variance will get different shapes. Certain conditions will be necessary in assigning weights to each unit. The construction of the estimators using unequal probabilities is complex as each unit will be assigned different

weights according to its size.

PC Calibration methods can also be applied under the small area estimation scenario (see Rao (2003), Chambers (2005) and Wang *et al*, 2008). Small area estimation is getting more and more importance due to the need of reliable small area statistics when only a small sample is available for these areas (Pfefferman, 2002). We may also look for the development of more adequate methods to handle qualitative variables. Multiple correspondence analysis (MCA) is used for this purpose (see Kaciak and Louviere, (1990) and Greenacre and Blasius (2006)). Cross validation analysis may also be used as a tool to select the number of principal components to be included in the analysis (Jolliffe, (2002), Krzanowski, (1987) and Josse and Husson (2011)). Cross validation may equally be applied to find an optimal ridge parameter (Jung, (2009), Golub *et al*, 1979).

Bibliography

Bardsley, P. and Chambers, R.L. (1984), Multipurpose estimation from unbalanced samples, *Applied Statistics*, **33**, 290-299.

Beaumont, J.-F. and Bocci, C. (2008), Another look at ridge calibration, *Metron-International Journal of Statistics*, **LXVI**, 5-20.

Bellhouse, D. R. (1987), Model-based estimation in finite population sampling, *Journal of the American Statistical Association*, **41**, 260-262.

Breidt, F. J. and Chauvet G. (2011), Penalized balanced sampling, *Accepted in Biometrika*,

Brewer, K. R. W. (1999), Cosmetic calibration with unequal probability sampling, *Survey Methodology*, **25(2)**, 205-212.

Cardot, H., Chaouch, M., Goga, C., and Labruére, C. (2010), Properties of design-based functional principal components analysis, Journal of Statistical Planning and Inference, **140**, 75-91.

Cassel, C. M., Särndal, C.-E., and Wretman, J. H. (1976), Some Results on Generalized Difference Estimation and Generalized Regression Estimation forFinite Populations, *Biometrika*, **63(3)**, 615-620.

Cassel, C. M., Särndal, C.-E., and Wretman, J. H. (1977), Foundations of inference in survey sampling, *John Wiley and sons*

Chambers, R.L. (1996), Robust case-weighting for multipurpose establishment surveys, *Journal of Official statistics*, **12**, 3-32.

Chambers, R.L. (2005), Calibrated weighting for small area estimation, *Southampton statistical sciences research institute methodology working paper* , **M05/04**.

Chen, J., Sitter, R.R. and Wu, C. (2002), Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys, *Biometrika*, **89**, 230-237.

Cochran, W. G. (1939), The use of analysis of variance in enumeration by sampling, *JASA*, **34**, 492-510.

Cochran, W. G. (1942), Sampling theory when the sampling units are of unequal sizes, *JASA*, **37**, 199-212.

Cochran, W. G. (1946), Relative accuracy of systematic and stratified random samples for a certain class of populations, *Ann. Math. Statist.*, **17**, 1-24.

Conniffe, D. and Stone, J. (1973), A critical view of ridge regression, *The Statistician*, **22**, 181-187.

Deming, W. E., and Stephan, F. (1947), On interpretation of census as samples, *Journal of the American Statistical Association*, **36**, 46-49.

Dunteman, G. H. (1989), Principal Components Analysis, *SAGE Publications.*

Deville, J.C., (1999), Simultaneous calibration of several surveys *Proceedings of Statistics Canada Symposium 99 of Statistics Canada*, May, 1999

Deville, J.-C., and Särndal, C.-E. (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376-382.

Dunstan, R. and Chambers, R.L. (1986), Model-based confidence intervals in multipurpose surveys, *Applied Statistics*, **35**, 276-280.

Fuller, W. A. (1973), Regression for sample surveys, *Paper presented at meeting of International Statistical Institute.*

Fuller, W. A. (1975), Regression analysis for sample survey, *Sankhya*, **37(C)**, 117-1332.

Fuller, W. A. (2002), Estimation par régression appliquée à l'échantillonnage, *Techniques d'enquête*, **28**, 5-25.

Goga, C. (2003), Estimation de la variance dans les sondages à plusiers échantillons et prise en compte de l'information auxiliaire par des modèles nonparametriques, *PhD Dessertation of Rennes 2 University.*

Goga, C., Muhammad-Shehzad, A. and Vanheuverzwyn, A. (2011), Principal component regression with survey data. application on the french media audience, *Proceedings of the 58th World Statistics Congress of the International Statistical Institute, Dublin, Ireland.*

Golub, G. H., Heath, M., and Wahba, G. (2006), Generalized cross-validation as a methods for choosing good ridge parametr, *Technometrics*, **21**, 215-223.

Greenacre, M., and Blasius, N., (2006), Multiple Correspondence Analysis and Related Methods, *4th Edition*, *Chapman and Hall, London.*

Guggemos, F. and Tillé, Y. (2010), Penalized calibration in survey sampling: Design-based estimation assisted by mixed models, *Journal of Statistical Planning and Inference*, **140**, 3199-3212.

Gujarati, D. N., (2003), Econometrics, *4th Edition*, *New York: McGraw Hill.*

Gunst, R.F. and Mason, R.L. (1977), Biased estimation in regression: an evaluation using mean squared error, *Journal of the American Statistical Association*, **72**, 616-628.

Gunst, R.F. and Mason, R.L. (1979), Some considerations in the evaluation of alternate prediction equations, *Technometrics*, **21**, 55-63.

Hadi, A.S. and Ling, R.F. (1998), Some cautionary notes on the use of principal components regression, *The American Statistician*, **52**, 15-19.

Hájek, J. (1971) Comment on a paper by Basu, D., *In Foundations of Statistical Inference (eds. V. P. Godambe and D. A. Sprott), 236.* **Toronto**: Holt, Rinehart and Winston.

Hawkins, D. M. (1973), On the investigation of alternative regressions by principal component analysis, *Appl. Statist.*, **22**, 275-286.

Henderson, H. V., and Searle, S. R. (1981), On deriving the inverse of a sum of matrices, *SIAM Review*, **23(1)**, 53-60.

Hoerl, A. E. (1962), Application of ridge analysis to regression problems, *Chemical Engineering Progress*, **58**, 54-59.

Hoerl, A. E., and Kennard, R.W. (1968), On regression analysis and biased estimation, *Technometrics*, **10**, 422-423.

Hoerl, A. E., and Kennard, R.W. (1970), Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.

Hoerl, A. E., and Kennard, R.W. (1976), Ridge regression: iterative estimation of the biasing parameter, *Communications in Statistics*, **5**, 77-88.

Hoerl, A. E., Kennard, D.J., and Baldwin, K. F. (1975), Ridge regression: some simulations, *Communications in Statistics*, **4**, 105-123.

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

Hotelling, H. (1933), Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, **24**, 417-441.

Hotelling, H. (1955), The relations of the newer multivariate statistical methods to factor analysis, *Brit. J. Statist. Psychol.*, **10**, 69-79.

Izenman, A. J. (2008), Modern Multivariate Statistical Techniques, *Springer*.

Jeffers, J. N. R. (1967), Two case studies in the application of principal component anlysis, *Appl. Statist.*, **16**, 225-236.

Jessen, R.J. (1942), Statistical investigation of a sample survey for obtaining farm facts, Iowa Agriculture Experiment Station Research Bulletin, **304**.

Johnson, R. A., and Wichern, D. W. (2002), Applied Multivariate Statistical Analysis, *Prentice Hall International, Inc.*

Jolliffe, I. T. (1982), A note on the use of principal components in regression, *Journal of the Royal Statistical Society, Series C*, **31(3)**, 300-303.

Jolliffe, I. T. (2002), Principal Component Analysis. *Springer* 2nd ed.

Josse, J. and Husson, F. (2011), Selecting the number of components in principal component analysis using cross-validation approximations, *Computational Statistics and Data Analysis*, **56(6)**, 1869ñ1879.

Jung, K.-M. (2009), Robust cross validations in ridge regression, *Journal of Appl. Math. and Informatics, Series C*, **27(3-4)**, 903-908.

Kaciak, E. and Louviere, J. (1990), Multiple Correspondence Analysis of Multiple Choice Experiment Data, *Journal of Marketing Research*, **27(4)**, 455-465.

Kapat, P. and Goel, P.K. (2010), Anomalies in the foundations of ridge regression: some clarifications, *International Statistical Review*, **78**, 209-215.

Kendall, M. G. (1957), A course in multivariate analysis, *London: Griffin.*

Kim, J. K., and Park, M. (2010), Calibration estimation in Survey Sampling, *International Statistical Review*, **78(1)**, 21-39.

Krzanowski, W. J. (1987), Cross validation in principal component analysis, *Biometrics*, **43(3)**, 575-584.

Kung, E. C., and Sharif, T. A. (1980), Multi-regression forecasting of Indian summer monsoon with antecedent pattern of the large scale circulation, *In WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*, 295-302.

Lohr, S. L. (1999), Sampling: design and analysis. Duxbury Press.

Lohr, S. L. (2007), Comment: Struggles with survey weighting and regression modeling , *Statistical Science*, **22**, 175-178.

Madow, W. G., and Madow, L. H. (1944), On the theory of systematic sampling, *Ann. Math. Statist.*, **15**, 1-24.

Marquardt, D. W., and Snee, R. D. (1975), Ridge regression in practice, *Amer. Statist.*, **29(1)**, 3-20.

Martens, H. and Naes, T. (1989), Multivariate calibration. New York: Wiley.

Mason, R. L., and Gunst, R. L. (1985), Selecting principal components in regression, *Statistics and Probability Letters*, **3**, 299-301.

Matthews, D.J. and Newman, J. A. (2012), Improving Correlation Function Fitting with Ridge Regression: Application to Cross-Correlation Reconstruction. *The Astrophysical Journal, to appear*, http://arxiv.org/abs/1109.2121.

Montanari, G.E. (1987), Post-sampling efficient QR-prediction in large-scale surveys, *International Statistical Review*, **55**, 191-202.

Narain, R.D. (1951) On sampling without replacement with varying probabilities, *J. Ind. Soc. Agril. Statist.*, **3**, 169-174.

Park, M. and Yang, M. (2008), Ridge regression estimation for survey samples, *Communications in Statistics. Theory and Methods*, **37**, 532-543.

Pearson, K. (1901), On lines and planes of closest fit to systems of points in space, *Philosophical Magazine*, **2**, 559-572.

Qian, G., Gaber, G., and Gupta, R. P. (1994), Principal components selection by the criterion of the minimum mean difference of complexity, *Journal of Multivariate Analysis*, **49**, 49-75.

Rao, J.N.K. and Singh, A. C. (1997), A ridge-shrinkage method for range-restricted weight calibration in survey sampling, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 57-65.

Rao, J.N.K. (2003), Small area estimation, *Wiley series in survey methodology.*

Ren, R. (2000), Utilisation d'information auxiliaire par calage sur fonction de repartition, *Unpublished PhD Thesis of Rennes 2 University.*

Royall, R. M. (1970), On finite population sampling theory under linear regression models, *Biometrika*, **57**, 377-387.

Royall, R.M. (1976), The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, **71**, 657-664.

Royall, R. M. and Herson, J. (1973), Robust estimation in finite populations I. *Journal of the American Statistical Association*, **68**, 880-889.

Särndal, C.E. (1980), On π-inverse weighting versus best linear unbiased weighting in probability sampling, *Biometrika*, **67**, 639-650.

Särndal, C.E. (2007), The calibration approach in survey theory and practice *Survey Methodology*, **33(02)**, 99-119.

Sen, A. R. (1953), On the estimate of the variance in sampling with varying probabilities, J. Indian Soc. Agri. Statisti, **5**, 119-127.

Särndal, C.E., Swenson, B., and Wretman, J.(1992), Model Assisted Survey Sampling, Springer.

Silva, P.L.N. and Skinner, C. (1997), Variable selection for regression estimation in finite populations, *Survey Methodology*, **23**, 23-32.

Singh, A.C. and Mohl, C.A. (1996), Understanding calibration estimation in survey sampling, *Survey Methodology*, **22**, 107-115.

Smith, G., and Campbell, F. (1980), A critique of some ridge regression methods, *JASA* , **75**, 74-103.

Swindel, B.F. and Chapman, D.D. (1973), Good ridge estimators, *Abstracts Booklet, Joint Statistical Meetings in New York City*, 126.

Théberge, A. (2000), Calibration and restricted weights, *Survey Methodology*, **26**, 99-107.

Theobald, C. M. (1974), Generalizations of mean square error applied to ridge regression, *Journal of the Royal Statistical Society B*, **36**, 103-106.

Trenkler, G. (1984), On the performance of biased estimators in the linear regression model with correlated or heteroscedastic errors, *Journal of Econometrics*, **25**, 179-190.

Vinod, H. D., and Ullah, A. (1981), Recent advances in regression methods, *Statistics: Textbooks and Monographs*, New York: Marcel Dekker Inc. **41**.

Wang, J., Fuller, W., and Qu, Y. (2008), Small area estimation under a restriction, *Survey methodology*, **34(1)**, 29-36.

Yates, F., and Grundy, P. M. (1953), Selection without replacement from within strata with probability proportional to size, *J. R. Statist. Soc. B*, **15**, 253-261.