International Journal of Mathematics and
Computer Science, **15**(2020), no. 1, 275–289

Some statistical properties of the PLS-calibration estimator

Sara Nahchel[1], Jelloul Allal[1], Zoubir Zarrouk[2]

[1]Faculté des Sciences
Université Mohamed Premier
Oujda, Morocco

[2]Faculté des Sciences Juridiques, Economiques et Sociales
Université Mohamed Premier
Oujda, Morocco

email: s.nahchel@ump.ac.ma

(Received September 4, 2019, Revised November 1, 2019,
Accepted November 10, 2019)

Abstract

The purpose of this paper is to provide statistical properties of the estimator given by a new calibration method, named the "PLS-calibration", that is used particularly in the context of multicollinear auxiliary variables.
An application on real data and some simulations will show the efficiency of the PLS-calibration estimator compared to the Horvitz-Thompson estimator when the ordinary calibration couldn't work because of the multicollinearity.

Introduction

The aim of survey sampling is to get a consistent estimator by using a sample that is supposed to be a photo-reduction of the population. For this reason, different estimating methods have been proposed in the literature. At first, Horvitz and Thompson (1952) [16] have proposed an unbiased one

Key words and phrases: Calibration, Horvitz-Thompson estimator, multicollinearity, PLS-calibration.
AMS (MOS) Subject Classifications: 62A86, 62B86, 62H86, 62J86.
ISSN 1814-0432, 2020, http://ijmcs.future-in-tech.net

in the class of linear estimates (Godambe and Joshi, 1965 [11]). After that, the regression estimator has been suggested to increase the precision of the Horvitz-Thompson estimator. However, the unbiasedness property to be replaced by the asymptotic unbiasedness (Sárnadl, Swensson and Wretman, 1992 [19]) was lost. Later, the calibration estimator (Deville and Särndal, 1992 [7]) has appeared saving the same properties of the regression estimator without refering to the regression model. Unfortunately, the nature of real data imposes the problem of multicollinearity between the auxiliary variables, which disrupts the calibration method. Therefore, the calibration estimator will lose its precision or won't be able to be calculated. So some remedies have been suggested in the literature; namely, the Ridge calibration (Bardsley and Chambers, 1984[3]), the PC calibration (Goga, Shehzad and Vanheuverzwyn, 2011[12]), the lasso calibration (Chen, 2016[5]) and more recently the PLS-calibration (Nahchel, Allal and Zarrouk, 2018 [18]).
The purpose of this article is to study the statistical properties of the PLS-calibration estimator. Thereby, it will be structured as follows: after reminding the reader about the calibration method and the PLS-calibration, we will study for the first time the bias and the variance of the estimator given by the new method. Then, the results found will be tested through an application on real data in addition to some simulations.

1 Calibration

Calibration, as defined by Deville and Särndal in 1992 [7], is a technique that takes into account additional auxiliary information other than those used at the sampling phase in order to adjust the sample structure. So, the sample will match the population structure. In other terms, calibration uses auxiliary information to assign some weights to each individual in the sample in a way that makes the sample a photo-reduction of the population. In consequence, the calibration estimator will ensure better precision than the Horvitz-Thompson estimator (1952).

Mathematically, if s is a sample of size n drawn from a finite population U that contains N individuals, $y = (y_1, ..., y_N)^{'}$ is the variable of interest and $x_k = (x_{k1}, ..., x_{km})^{'}$ is the vector of m auxiliary variables observed on the kth element and d_k is the sample weight that is equal to the inverse of the inclusion probability π_k, the calibration procedure (Deville and Särndal,

1992 [7]) is defined by

$$\begin{cases} \min \sum_{k\in s} H(d_k, w_k) \\ \sum_{k\in s} w_k x_k = \sum_{k\in U} x_k \end{cases} \tag{1}$$

where w_k are the calibration weights that we are looking for and $H(.,.)$ is a pseudo-distance on $\mathbb{R}$ defined by $H(d_k, w_k) = d_k G\left(\frac{w_k}{d_k}\right)$, where $G()$ is one of the convex distance functions proposed in the literature (for further details see Nahchel, Allal and Zarrouk (2018)[18], Deville and Särndal (1992)[7], Husain (1969)[15] and Deville, Ireland and Kullback (1968)[17]).
So, the calibration estimator will be written as follows

$$\hat{Y}_{cal} = \sum_{k\in s} w_k y_k. \tag{2}$$

By using the asymptotic equivalence between the calibration estimator and the regression estimator (Deville and Särndal, 1992 [7]), the calibration weights w_k can be written as

$$w_k = d_k + \left(\sum_{k\in U} x_k^{'} - \sum_{k\in s} d_k x_k^{'}\right)\left(\sum_{k\in s} d_k q_k x_k x_k^{'}\right)^{-1}\left(\sum_{k\in s} d_k q_k x_k\right) \tag{3}$$

where $q_k = \frac{1}{\sigma_k^2}$ and σ_k^2 is the variance of y_k with respect to the model $\xi: \ y = X\beta + e$ such that $X = (x_k^{'})_{k\in U}$, $\beta = (\beta_1, ..., \beta_m)^{'} = \left(\sum_{k\in U} \frac{x_k x_k^{'}}{\sigma_k^2}\right)^2 \sum_{k\in U} \frac{x_k y_k}{\sigma_k^2}$ and $e = (e_1, ...e_N)^{'}$ the residual vector.
To evaluate the precision of the calibration estimator, the following expression of the asymptotic variance (Deville and Särndal, 1992 [7]) can be used

$$AV\left(\hat{Y}_{cal}\right) = \sum_{l\in U}\sum_{k\in U} \Delta_{kl}(d_l e_l)(d_k e_k) \tag{4}$$

with $e_k = y_k - x_k^{'}\beta$ and $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l = Prob(k, l \in s) - Prob(k \in s)Prob(l \in s)$. But the e_k for $k \in U$ are unknown. For that reason the asymptotic variance will be estimated by

$$\hat{AV}\left(\hat{Y}_{cal}\right) = \sum_{l\in s}\sum_{k\in s} \tilde{\Delta}_{kl}(d_l \tilde{e}_l)(d_k \tilde{e}_k) \tag{5}$$

such that $\tilde{e}_k = y_k - x_k^{'}\hat{\beta}$, $\hat{\beta}$ is the estimate of the unknown parameter β and $\tilde{\Delta}_{kl} = \dfrac{\Delta_{kl}}{\pi_{kl}}$.

2 PLS-calibration

2.1 Definition

The PLS-calibration (Nahchel, Allal and Zarrouk, 2018 [18]) refers simply to the combination of the PLS regression (Wold, 1966) and the calibration procedure. More precisely, the PLS-calibration has two phases: The first one consists of applying the PLS regression between the variable of interest and the auxiliary variables all observed on the sample s while the second one is preserved to calibrate s on the PLS components saved before. Therefore, the PLS-calibration reduces the dimension of the auxiliary variables based on the Akaike information criterion (Akaike, 1973 [1]), eliminates the multicollinearity and then calibrates the sample s without any problems.
Mathematically, the PLS calibration is defined by

$$\begin{cases} \min \sum\limits_{k\in s} H(d_k, w_k) \\ \sum\limits_{k\in s} w_k l_k = \sum\limits_{k\in U} l_k, \end{cases} \tag{6}$$

where $l_k = (l_1, ..., l_v)^{'}$ is the vector of the first v PLS components for the kth element.
Due to the asymptotic equivalence between the calibration estimator and the regression estimator the new calibration, weights can be written as

$$w_{PLS,k} = d_k + \left(\sum_{k\in U} l_k^{'} - \sum_{k\in s} d_k l_k^{'}\right)\left(\sum_{k\in s} d_k q_k l_k l_k^{'}\right)^{-1}\left(\sum_{k\in s} d_k q_k l_k\right). \tag{7}$$

In consequence, the PLS-calibration estimator can be expressed by

$$\begin{aligned} \hat{Y}_{PLS} &= \sum_{k\in s} w_{PLS,k} y_k \\ &= \sum_{k\in s} d_k y_k + \left(\sum_{k\in U} l_k^{'} - \sum_{k\in s} d_k l_k^{'}\right)\left(\sum_{k\in s} d_k q_k l_k l_k^{'}\right)^{-1}\left(\sum_{k\in s} d_k q_k l_k y_k\right) \end{aligned}$$

$$= \hat{Y}_{HT} + \left(\sum_{k \in U} l_k^{'} - \sum_{k \in s} d_k l_k^{'} \right) \hat{\beta}_{PLS} \tag{8}$$

where $\hat{Y}_{HT} = \sum_{k \in s} d_k y_k$ is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952 [16]) and $\hat{\beta}_{PLS} = \left(\sum_{k \in s} d_k q_k l_k l_k^{'} \right)^{-1} \left(\sum_{k \in s} d_k q k l_k y_k \right)$.
Since the PLS-calibration estimator is asymptotically equivalent to the regression estimator ξ returns to

$$\xi^{'} : \quad y = L_v \beta_{PLS} + \epsilon_v = X \alpha_{PLS} + \epsilon_v \tag{9}$$

with

- $L_v = (l_k^{'})_{k \in U}$
- $\beta_{PLS} = \left(\sum_{k \in U} l_k l_k^{'} \right)^{-1} \left(\sum_{k \in U} l_k y_k \right)$ estimated by $\hat{\beta}_{PLS}$
- $\epsilon_v = (\epsilon_{v,1}, ..., \epsilon_{v,N})$ the residual vector
  $\alpha_{PLS} = M_v \beta_{PLS}$
- and $M_v = (m_1, ..., m_v)$ the matrix of the PLS coefficients.

2.2 Statistical properties

Due to the asymptotic equivalence between the PLS-calibration estimator and the regression estimator that uses the PLS regression, both estimators share the same statistical properties.

Proposition 1

The PLS-calibration estimator is unbiased under the model ξ if the sum of the PLS components are perfectly estimated on s using the Horvitz-Thompson estimator (i.e $\sum_{k \in s} d_k l_k = \sum_{k \in U} l_k$), or if the matrix of PLS coefficients $M_{v,s}$ provided by the PLS regression between $y_s = (y_1, ..., y_n)^{'}$ and $X_s = (x_k^{'})_{k \in s}$ is the same as the M_v in $\xi^{'}$ because

$$E_\xi \left(\hat{Y}_{PLS} - t_y \right) = \left(1_U^{'} L_v - d_s^{'} L_{v,s} \right) \left(M_{v,s}^{'} - M_v^{'} \right) \beta \tag{10}$$

with $1_U = (\underbrace{1, ..., 1}_{N \quad times})^{'}$ and $d_s = (d_1, ..., d_n)^{'}$.

Proof:

Using the formula (8) of $\hat{Y}_{PLS}$, we get

$$\begin{aligned}\hat{Y}_{PLS} - t_y &= \hat{Y}_{HT} + \left(1_U^{'} L_v - d_s^{'} L_{v,s}\right)\hat{\beta}_{PLS} - t_y \\ &= \sum_{k\in s}\frac{y_k}{\pi_k} + \sum_{k\in U} l_k^{'}\hat{\beta}_{PLS} - \sum_{k\in s}\frac{l_k^{'}\hat{\beta}_{PLS}}{\pi_k} - \sum_{k\in U} y_k \\ &= \sum_{k\in s}\frac{l_k^{'}\beta_{PLS}+\varepsilon_k}{\pi_k} - \sum_{k\in s}\frac{l_k^{'}\hat{\beta}_{PLS}}{\pi_k} + \sum_{k\in U} l_k^{'}\hat{\beta}_{PLS} - \sum_{k\in U}\left(l_k^{'}\beta_{PLS}+\varepsilon_k\right) \\ &= \sum_{k\in s}\frac{l_k^{'}}{\pi_k}\left(\beta_{PLS} - \hat{\beta}_{PLS}\right) + \sum_{k\in U} l_k^{'}\left(\hat{\beta}_{PLS} - \beta_{PLS}\right) + \sum_{k\in s}\frac{\varepsilon_k}{\pi_k} - \sum_{k\in U}\varepsilon_k \\ &= \left(\sum_{k\in U} l_k^{'} - \sum_{k\in s}\frac{l_k^{'}}{\pi_k}\right)\left(\hat{\beta}_{PLS} - \beta_{PLS}\right) + \sum_{k\in s}\frac{\varepsilon_k}{\pi_k} - \sum_{k\in U}\varepsilon_k \qquad (11)\end{aligned}$$

Then,

$$\begin{aligned}E_\xi\left(\hat{Y}_{PLS} - t_y\right) &= \left(\sum_{k\in U} l_k^{'} - \sum_{k\in s}\frac{l_k^{'}}{\pi_k}\right) E_\xi\left(\hat{\beta}_{PLS} - \beta_{PLS}\right) \\ &= \left(\sum_{k\in U} l_k^{'} - \sum_{k\in s}\frac{l_k^{'}}{\pi_k}\right) Bias\left(\hat{\beta}_{PLS}\right). \qquad (12)\end{aligned}$$

To calculate the bias of $\hat{\beta}_{PLS}$, we have to start by computing $E_\xi\left(\hat{\beta}_{PLS}\right)$.
Since $\hat{\beta}_{PLS} = \left(L_{v,s}^{'}\Pi_s^{-1}L_{v,s}\right)^{-1} L_{v,s}^{'}\Pi_s^{-1} y_s$ and $X_s = L_{v,s}M_{v,s}^{'}$,
we obtain that

$$\begin{aligned}E_\xi\left(\hat{\beta}_{PLS}\right) &= \left(L_{v,s}^{'}\Pi_s^{-1}L_{v,s}\right)^{-1} L_{v,s}^{'}\Pi_s^{-1} E_\xi\left(y_s\right) \\ &= \left(L_{v,s}^{'}\Pi_s^{-1}L_{v,s}\right)^{-1} L_{v,s}^{'}\Pi_s^{-1} X_s\beta \\ &= \left(L_{v,s}^{'}\Pi_s^{-1}L_{v,s}\right)^{-1} L_{v,s}^{'}\Pi_s^{-1} L_{v,s}M_{v,s}^{'}\beta \\ &= M_{v,s}^{'}\beta \qquad (13)\end{aligned}$$

So, the bias of $\hat{\beta}_{PLS}$ is given by

$$E_\xi\left(\hat{\beta}_{PLS} - \beta_{PLS}\right) = E_\xi\left(\hat{\beta}_{PLS}\right) - \beta_{PLS}$$

$$= M_{v,s}^{'}\beta - M_{v}^{'}\beta$$
$$= \left(M_{v,s}^{'} - M_{v}^{'}\right)\beta \tag{14}$$

Finally,

$$E_{\xi}\left(\hat{Y}_{PLS} - t_y\right) = \left(1_U^{'}L_v - d_s^{'}L_{v,s}\right)\left(M_{v,s}^{'} - M_{v}^{'}\right)\beta.$$

Proposition 2:

We now use theorem 1 of H. Chun and S. Keleş (2010, page 6)[6]. If $\frac{m}{n} \to 0$, then $\hat{\beta}_{PLS} - \beta_{PLS} = O\left(\sqrt{\frac{m}{n}}\right)$. Therefore, the asymptotic variance of $\hat{Y}_{PLS}$ is expressed by

$$AV_p\left(\hat{Y}_{PLS}\right) = \sum_{i\in U}\sum_{k\in U}\left(\pi_{ki} - \pi_k\pi_i\right)\frac{y_k - l_k^{'}\beta_{PLS}}{\pi_k}\frac{y_i - l_i^{'}\beta_{PLS}}{\pi_i} \tag{15}$$

with $\beta_{PLS} = \left(L_v^{'}L_v\right)^{-1}L_v^{'}y$.

Since $AV_p\left(\hat{Y}_{PLS}\right)$ is unknown, it is suggested to estimate it by

$$\hat{Var}\left(\hat{Y}_{PLS}\right) = \sum_{i\in s}\sum_{k\in s}\frac{\left(\pi_{ki} - \pi_k\pi_i\right)}{\pi_{ki}}\frac{y_k - l_k^{'}\hat{\beta}_{PLS}}{\pi_k}\frac{y_i - l_i^{'}\hat{\beta}_{PLS}}{\pi_i}. \tag{16}$$

Proof:

Let $\hat{\alpha}_{PLS}$ be the estimate of α_{PLS} the PLS regression estimator such that $\hat{\alpha}_{PLS} = M_v\hat{\beta}_{PLS}$ and $\alpha_{PLS} = M_v\beta_{PLS}$.

Next we use the closed form of $\hat{\alpha}_{PLS}$ given by Helland (1990)[14]

$$\hat{\alpha}_{PLS} = \hat{R}\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}\hat{R}^T S_{XY} \tag{17}$$

where $\hat{R} = \left(S_{XY}, ..., S_{XX}^{v-1}S_{XY}\right)$, S_{XX} and S_{XY} are the estimations of the variance $\sum_{XX}$ of X and of the covariance σ_{XY} of X and Y respectively.
We get,

$$\|\hat{\alpha}_{PLS} - \alpha_{PLS}\|_2 = \|\hat{R}\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}\hat{R}^T S_{XY} - R(R^T\sum_{XX}R)^{-1}R^T\sigma_{XY}\|_2$$

$$= \|\hat{R}\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}\hat{R}^T S_{XY} + R\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}\hat{R}^T S_{XY}$$

$$-R\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}\hat{R}^T S_{XY} - R\left(R^T\Sigma_{XX}R\right)^{-1}R^T\sigma_{XY}\|_2$$

$$= \|\left(\hat{R}-R\right)\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}\hat{R}^T S_{XY}+R\left[\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}\hat{R}^T S_{XY} - \left(R^T\Sigma_{XX}R\right)^{-1}R^T\sigma_{XY}\right]\|_2$$

$$= \|\left(\hat{R}-R\right)\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}\hat{R}^T S_{XY}+R\left[\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}\hat{R}^T S_{XY} - \left(R^T\Sigma_{XX}R\right)^{-1}\hat{R}^T\sigma_{XY}\right]\|_2$$

$$+R\left[\left(R^T\Sigma_{XX}R\right)^{-1}\hat{R}^T\sigma_{XY} - \left(R^T\Sigma_{XX}\right)^{-1}R^T\sigma_{XY}\right]\|_2$$

$$\leqslant \|\hat{R}-R\|_2\|\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}R^T S_{XY}\|_2+\left[\|R\|_2\|\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1} - (R^T\sum_{XX}R)^{-1}\|_2\|\hat{R}^T S_{XY}\|_2\right]$$

$$+\|R\|_2\|(R^T\sum_{XX}R)^{-1}\|_2\|\left(\hat{R}^T S_{XY}-R^T\sigma_{XY}\right)\|_2 \quad (18)$$

because $\|\hat{R}-R\|_2 = O\left(\sqrt{\frac{m}{n}}\right)$, $\|\left(\hat{R}^T S_{XX}\hat{R}\right)-\left(R^T\sum_{XX}R\right)\|_2 = O\left(\sqrt{\frac{m}{n}}\right)$ and $\|\hat{R}^T S_{XY}-R^T\sigma_{XY}\|_2 = O\left(\sqrt{\frac{m}{n}}\right)$ due to lemmas 2 and 3 from H. Chun and S. Keleş (2010, page 21)[6].

Actually, by using the definition of a matrix norm, we have

$$\|\hat{R}-R\|_2 \leqslant \sqrt{v}\ max_{1\leqslant k\leqslant r}\|S_{XX}^{k-1}S_{XY} - \sum\nolimits_{XX}^{k-1}\sigma_{XY}\|_2 \quad (19)$$

Because of the lemma 3 from H. Chun and S. Keleş (2010, page 21)[6], we have

$$\|\hat{R}-R\|_2 = O\left(\sqrt{\frac{m}{n}}\right) \quad (20)$$

Refering to Golub and Van Loan (1987)[13] we apply

$$\|\left(A+E\right)^{-1}-A^{-1}\|_2 \leqslant \|E\|_2\|A^{-1}\|_2\|\left(A+E\right)^{-1}\|_2$$

to $\|\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}-(R^T\sum_{XX}R)^{-1}\|_2$ where $A = R^T\sum_{XX}R$ and $E = \hat{R}^T S_{XX}\hat{R}-R^T\sum_{XX}R$
Then, we obtain

$$\|\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1} - (R^T\sum\nolimits_{XX}R)^{-1}\|_2$$

$$\leqslant \|\hat{R}^T S_{XX}\hat{R} - R^T\sum\nolimits_{XX}R\|_2\|(R^T\sum\nolimits_{XX}R)^{-1}\|_2\|\left(\hat{R}^T S_{XX}\hat{R}\right)^{-1}\|_2 \quad (21)$$

We know that

$$\|\hat{R}^T S_{XX}\hat{R} - R^T\sum\nolimits_{XX} R\|_2 = \|\hat{R}^T S_{XX}\hat{R} - R^T S_{XX}\hat{R} + R^T S_{XX}\hat{R} - R^T\sum\nolimits_{XX} R\|_2$$

$$= \| \left(\hat{R}^T + R^T\right)\left(S_{XX}\hat{R}\right) - R^T(S_{XX}\hat{R} - \sum\nolimits_{XX} R)\|_2$$

$$= \| \left(\hat{R}^T + R^T\right)\left(S_{XX}\hat{R}\right) - R^T(S_{XX}\hat{R} - \sum\nolimits_{XX}\hat{R} + \sum\nolimits_{XX}\hat{R} - \sum_{XX} R)\|_2$$

$$= \| \left(\hat{R}^T + R^T\right)\left(S_{XX}\hat{R}\right) - R^T[(S_{XX} - \sum\nolimits_{XX})\hat{R} + \sum\nolimits_{XX}(\hat{R} - R)]\|_2$$

$$\leqslant \underbrace{\|\hat{R}^T - R^T\|_2}_{=O\left(\sqrt{\frac{m}{n}}\right)} \|S_{XX}\hat{R}\|_2 + \|R^T\|_2 \underbrace{\|S_{XX} - \sum\nolimits_{XX}\|_2}_{=O\left(\sqrt{\frac{m}{n}}\right)} \|\hat{R}\|_2 + \|R^T\|_2 \|\sum\nolimits_{XX}\|_2 \underbrace{\|\hat{R} - R\|_2}_{=O\left(\sqrt{\frac{m}{n}}\right)}$$

$$= O\left(\sqrt{\frac{m}{n}}\right) \tag{22}$$

$\|S_{XX} - \sum_{XX}\|_2 = O\left(\sqrt{\frac{m}{n}}\right)$ since the lemma 2 from H. Chun and S. Keleş (2010, page 21) [6]. By formula (21), we get

$$\| \left(\hat{R}^T S_{XX}\hat{R}\right)^{-1} - \left(R^T\Sigma_{XX}R\right)^{-1} \|_2 = O\left(\sqrt{\frac{m}{n}}\right) \tag{23}$$

Finally,

$$\|\hat{R}^T S_{XY} - R^T\sigma_{XY}\|_2 = \|\hat{R}^T S_{XY} - R^T S_{XY} + R^T S_{XY} - R^T\sigma_{XY}\|_2$$

$$\leqslant \underbrace{\|\hat{R}^T - R^T\|_2}_{=O\left(\sqrt{\frac{m}{n}}\right)} \|S_{XY}\|_2 + \|R^T\|_2 \underbrace{\|S_{XY} - \sigma_{XY}\|_2}_{=O\left(\sqrt{\frac{m}{n}}\right)}$$

$$= O\left(\sqrt{\frac{m}{n}}\right) \tag{24}$$

$\|S_{XY} - \sigma_{XY}\|_2 = O\left(\sqrt{\frac{m}{n}}\right)$ due to lemma 2 from H. Chun and S. Keleş (2010, page 21) [6].
By (20), (23) and (24), we get

$$\|\hat{\alpha}_{PLS} - \alpha_{PLS}\| = O\left(\sqrt{\frac{m}{n}}\right). \tag{25}$$

Since

$$\hat{\beta}_{PLS} - \beta_{PLS} = M_v^{'}\hat{\alpha}_{PLS} - M_v^{'}\alpha_{PLS}$$
$$= M_v^{'}\left(\hat{\alpha}_{PLS} - \alpha_{PLS}\right) \quad (26)$$

we get,

$$\hat{\beta}_{PLS} - \beta_{PLS} = O\left(\sqrt{\frac{m}{n}}\right) \quad (27)$$

Consequently,

$$\left(\hat{Y}_{PLS} - t_y\right) = \hat{Y}_{HT} + \left(1_U^{'}L_v - d_s^{'}L_{v,s}\right)\hat{\beta}_{PLS} - t_y$$

$$= \hat{Y}_{HT} + \left(1_U^{'}L_v - d_s^{'}L_{v,s}\right)\beta_{PLS} - t_y - \left(1_U^{'}L_v - d_s^{'}L_{v,s}\right)\beta_{PLS} + \left(1_U^{'}L_v - d_s^{'}L_{v,s}\right)\hat{\beta}_{PLS}$$

$$= \left[\hat{Y}_{HT} + \left(1_U^{'}L_v - d_s^{'}L_{v,s}\right)\beta_{PLS} - t_y\right] + \left(1_U^{'}L_v - d_s^{'}L_{v,s}\right)\left(\hat{\beta}_{PLS} - \beta_{PLS}\right)$$

$$= \left[\hat{Y}_{HT} + \left(1_U^{'}L_v - d_s^{'}L_{v,s}\right)\beta_{PLS} - t_y\right] + O_P\left(\sqrt{\frac{m}{n}}\right) \quad (28)$$

so,

$$AV_p\left(\hat{Y}_{PLS}\right) = Var\left(\hat{Y}_{HT} + \left(1_U^{'}L_v - d_s^{'}L_{v,s}\right)\beta_{PLS}\right)$$

$$= \sum_{i\in U}\sum_{k\in U}\left(\pi_{ki} - \pi_k\pi_i\right)\frac{y_k - l_k^{'}\beta_{:PLS}}{\pi_k}\frac{y_i - l_i^{'}\beta_{PLS}}{\pi_i}.$$

3 Simulation

This section is composed of three subsections. The first one is dedicated to the direct application of the PLS-calibration on real data given by Marocmetrie (a Moroccan company specialized in the TV audience measurement). The second one uses a sample from the same real data considered as a population in order to be able to evaluate the bias. In the third subsection, a simulation on non-real-data will be made to confirm the previous results and to be able to use the bias expression given by the formula (10). Finally, it is necessary to mention that self programming functions in the R software were adopted for all the calibration methods used.

3.1 Application on real data

The data contains 23 calibration variables (6 qualitative variables and 17 quantitative variables) observed on 10121 individuals. By using the condition number (see Erkel-Rousse, 1995[9]) it was verified that the data suffers from severe multicollinearity. Therefore, the ordinary calibration couldn't work. So, only the PLS-calibration and the Horvitz and Thompson estimator were calculated with their statistical properties. We recall that the Horvitz-Thompson Estimator variance is given by

$$V(\hat{Y}_{HT}) = \sum_{k\in U} x_k^{'} x_k \left(\frac{1-\pi_k}{\pi_k}\right) + \sum_{k\neq l\in U} x_k^{'} x_l \left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\right) \tag{29}$$

and estimated by

$$\hat{V}(\hat{Y}_{HT}) = \sum_{k\in s} x_k^{'} x_k \left(\frac{1-\pi_k}{(\pi_k)^2}\right) + \sum_{k\neq l\in s} x_k^{'} x_l \left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}\pi_k\pi_l}\right) \tag{30}$$

The table hereafter summarizes the results:

Table 1: The application on real data results

	Direct application	
	PLS-calibration	Horvitz-Thompson
Total estimation	75774351.5	78639148.1
Estimated variance	759143945.9	611866918446.0

As table 1 shows, the PLS-calibration is doing better than the Horvitz-Thompson estimator according to the variance estimation. However, the bias is unknown. So, the Mean Square Error (MSE) can't be calculated to judge the performance of our method. This is why, the following subsections are necessary.

3.2 Simulation using real data

In this simulation the data introduced before is considered as the whole population and a sample s of size 1013 is drawn using the stratified random sampling with proportional allocation strategy to insure the representability

(Gerville-Rache and Couallier, 2011[10]). The sample size was chosen to have a sample that represents 10% of the universe as the real data do. The multicollinearity problem still persists in the sample s. In consequence, only the PLS-calibration estimator and the Horvitz-Thompson estimator with their statistical properties were calculated.
In order to make our results reliable, we need to use the bootstrap procedure (Efron and Tibshini, 2000 [8]). As we are using the stratified random sampling with proportional allocation strategy we work with the stratified bootstrap technique with $B = 1000$.

Table 2: The simulation with real data results

Total=114338	Direct application		Bootstrap	
	PLS-cali.	Horvitz-Thompson	PLS-cali.	Horvitz-Thompson
Total estimation	114475.9	114477.4	114350.9	114359.8
Asymptotic var.	9861094.3	33230717.3	9861094.3	33230717.3
Estimated var,	9495278.3	33292453.4	9492193.9	33238682.3
Observed bias	137.9	139.4	12.9	21.8
MSE	9880110.7	33250149.7	9861260.7	33231192.5

As table 2 shows, the estimated variance given by formulas (16) and (30) approaches the real variance calculated through formulas (15) and (29). So, they reflect the real precision. On the other hand, the PLS-calibration estimator gives an estimation with lower bias and much lower variance than the Horvitz-Thompson estimator. In other terms, The Mean Square Error (MSE=Asymptotic variance + Observed Bias^2) allows us to say that the PLS-calibration estimator is the best. Finally, the Bootstrap procedure confirms all the results and shows the consistency of the PLS-calibration estimator.

3.3 Simulation with non-real-data

To double check the previous results (see subsections 3.1 and 3.2), a simulation on non-real-data has been done. A data X of 15 variables (10 quantitative variables (Gaussian variables $\mathcal{N}(0,1)$) and 5 qualitative variables) and 10000 individuals suffering from strong multicollinearity was built by

using the Cholesky decomposition (Cholesky, 2005[4], Angeletti and Bernay, 2010[2]) . Furthermore, the coefficient vector β was generated using the uniform distribution ($\mathcal{U}(0,1)$) and the error vector e was generated using the Gaussian distribution ($\mathcal{N}(0,1)$). After that, the variable of interest Y was calculated through equation $Y = X\beta + e$. Then, a sample of size 1000 was drawn by the stratified random sampling with proportional allocation strategy. Finally, the PLS-calibration estimator and the Horvitz-Thompson estimator with their statistical properties were calculated. The ordinary calibration couldn't work due to the multicollinearity and the stratified bootstrap was used as in the previous subsection. The following table outlines the results:

As it can be seen in table 3, the PLS-calibration still performs better

Table 3: The simulation with non-real-data results

Real Total= 24283.4	Direct application		Bootstrap	
	PLS-cali.	Horvitz-Thompson	PLS-cali.	Horvitz-Thompson
Total estimation	24364.2	24455.1	24218.8	24287.8
Asymptotic variance	91003.9	3141533.6	91003.9	3141533.6
Estimated variance	86009.3	3064207.2	88333.8	3143618.8
Observed bias	80.7	171.6	-64.6	4.3
Bias through the formula	-14.8	NA	2.9	NA
MSE	97516.39	3170980.16	95177.06	3141552.09

than the Horvitz-Thompson one according to the MSE. Actually, the PLS-calibration reduces deeply the variance of the Horvitz-Thompson estimator which covered the little loss in the bias after using the bootstrap while the direct application shows that the PLS-calibration does very well and provides a nice gain for the precision and the bias at the same time. Finally, the bias calculated through formula (10) gives lower values than the observed ones as it is for the Horvitz-Thompson estimator that is supposed to be unbiased when the observed bias is different from 0.

Conclusion

In conclusion, the PLS-calibration is a nice remedy to overcome the multicollinearity problem when the ordinary calibration couldn't work. Moreover,

it conserves the benefits of the calibration method by insuring better precision than the Horvitz-Thompson estimator. So, it will be very beneficial to compare our method with the other remedies that are in the literature (the Ridge calibration (Bardsley and Chambers, 1984 [3]), the Principal Component calibration (Goga, Shehzad and Vanheuverzwyn, 2011 [12]) and the LASSO calibration (Chen, 2016[5])) in order to show its performance.

Acknowledgement
We would like to show our gratitude to MAROCMETRIE Company for sharing the data used in the application with us in order to assist our research.

References

[1] H. Akaike, Information theory and an extension of the maximum Likelihood Principle, Selected Papers of Hirotugu Akaike, Springer, 1973.

[2] C. Angeletti, B. Bernay, (2010), Génération de séquences aléatoires corrélées, Final year study project, Institut Supérieur d'Informatique de Modélisation et de leurs Applications, Aubiere.

[3] P. Bardsley, R. L. Chambers, Multipurpose estimation from unbalanced samples, Applied Statistics, **33,** (1984), 290–299.

[4] A. L. Cholesky, Sur la rsolution numrique des systmes dquations linaires, Bulletin de laSabix, **39,** (2005).

[5] K. T. Chen, Using LASSO to Calibrate Non-probability Samples using Probability Samples, A dissertation of the requirement for the degree of Doctor of Philosophy, (2016).

[6] H. Chun, S. Keleş Sparse partial least squares regression for simultaneous dimension reduction and variable selection, Journal of the Royal Statistical Society, **72,** no. 1, (2010), 3–25.

[7] J. C. Deville, C. E. and Särndal, Calibration estimators in survey sampling, Journal of American Statistical Association, **87,** (1992), 376–382.

[8] B. Efron, J. R. Tibshini, An Introduction to The Bootstrap, Florida, USA, 2000.

[9] H. Erkel-Rousse, (1995), Introduction l'économiétrie du modèle linéaire - Chapitre III: multicolinéarité dans le modèle linéaire ordinaire: définition, détection, propositions de solutions. 10.13140/RG.2.1.1693.1926.

[10] L. Gerville-Réache, V. Couallier, Echantillon reprsentatif (d'une population finie): définition statistique et propriétés, HAL archives-ouvertes, (2011).

[11] V. P. Godambe, V. M. Joshi, (1965), Admissibility and bayes estimation in sampling finite Population, The Annals of Mathematical Statistics, **36,** 1707–1722.

[12] C. Goga, M. A. Shehzad, A. Vanheuverzwyn, Principal component regression with survey data application on the French media audience, Proceedings of the 58th ISI World Statistics Congress, Dublin, (2011).

[13] G. H. Golub, C. F. Van Loan, Matrix Computations, Johns Hopkins University Press, Baltimore, MD, USA, 1987.

[14] I. S. Helland, Partial least squares regression and statistical models. Scand. J. Stat., **17,** (1990), 97–114.

[15] M. Husain, Construction of Regression Weights for estimation in sample surveys, Master Dissertation, Iowa State University, Iowa, USA, (1969).

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

[17] C. T. Ireland, S. Kullback, Contingency tables with given marginals, Biometrika, **55,** no. 1, (1968),179–188.

[18] S. Nahchel, J. Allal, Z. Zarrouk, PLS-calibration: a new calibration method, International journal of applied mathematics and statistics, **57,** (2018), 82–90.

[19] C. E. Särndal, B. Swensson, J. Wretman, Model-assisted Survey Sampling, Springer-Verlag, New York, 1992.

[20] H. Wold, Estimation of principal components and related models by iterative least squares, In P. R. Krishnaiaah (ed.) Multivariate Analysis, Academic Press, New York, 1966.