Principal Component Regression with Survey Data. Application on the French Media Audience

Goga, Camelia
IMB, Université de Bourgogne
9 Avenue Alain Savary,
Dijon (21000), France
camelia.goga@u-bourgogne.fr

Shehzad, Muhammad Ahmed
IMB, Université de Bourgogne
9 Avenue Alain Savary,
Dijon (21000), France
Muhammad-Ahmed.Shehzad@u-bourgogne.fr

Vanheuverzwyn, Aurélie
Médiamétrie,
Levallois, Paris (92532), France
Avanheuverzwyn@mediametrie.fr

1 Introduction

The Mediamat panel of Médiamétrie is the device used to collect the measurements of the television (Media) audiences in France. The Mediamat panel was installed in 1989 and composed of 2300 French representative households. Over the years, the panel has been enriched in two aspects: the sample size and the qualification criteria. Today the panel consists of 4200 households for which a large number of variables (socio-demographic, geographic) with known totals is available. At the time being, this information is partially used: the quota and the calibration variables are only used. On the same hand, French media landscape/scenario has changed significantly over the past years. In 1989, the majority of the households did not receive more than 6 channels, but it is not the case today due to the development in the paid digital offers (cable, satellite and ADSL) and the appearance of the TNT. The combination of these facts motivates us today to review the list of criteria used until now to structure the panel. In this paper, we aim at extending the actual estimation calibration methods for allowing to incorporate large auxiliary data sets in order to improve efficiency. We suggest a new class of estimators based on principal component analysis which reduces the dimension while keeping the maximum of information. This approach is even more motivated by the fact that the auxiliary information will be very soon enriched by the development of the digital television channels with "return way" (the "return ways" allow to cable and satellite operators to have an exhaustive measure of the number of alight meters).

2 Model-assisted and calibration estimators for finite population totals

We consider the population $U = \{1, \ldots, i, \ldots, N\}$ and the sample $s \in \mathcal{S}$ selected from U according to a sampling plan $p(s)$. Let $\pi_i = Pr(i \in s)$ and $\pi_{ij} = Pr(i, j \in s)$, $i \neq j$ be the probabilities of inclusion of first and second degree respectively. Let $\mathcal{Y}$ be the variable of interest and y_i be the value of $\mathcal{Y}$ for

the i-th individual. Let $\mathcal{X}_1, \ldots, \mathcal{X}_p$ be the auxiliary variables. We denote by $\mathbf{x}_i = (X_{1i}, \ldots, X_{pi})'$ the values of the auxiliary variables for the i-th individual and let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ be the $N \times p$ matrix having the vectors $\mathbf{x}_i'$ for all $i \in U$ as rows.
We want to estimate the total of $\mathcal{Y}$ over the whole population U,

$$t_y = \sum_U y_i$$

by taking into account the auxiliary information. For this, we introduce the superpoplatoin model ξ,

$$(1) \quad \xi : \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

and we use the *model-assisted-approach* (Särndal *et al.*, 1992). The regression coefficient β is estimated under the model ξ by ordinary least square method (OLS) and we obtain

$$(2) \quad \hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

provided that $\mathbf{X}'\mathbf{X}$ is a full-rank matrix. An estimator of $\hat{\beta}_{OLS}$ based on the sampling design is $\hat{\boldsymbol{\beta}}_{OLS,\pi} = (\mathbf{X}_s'{\boldsymbol{\Pi}_s}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'{\boldsymbol{\Pi}_s}^{-1}\mathbf{y}_s$ where $\mathbf{X}_s$, respectively $\mathbf{y}_s$, is the restriction of $\mathbf{X}$, respectively of $\mathbf{y}$, on the sample s and $\boldsymbol{\Pi}_s^{-1} = \mathrm{diag}(\frac{1}{\pi_i})_{i\in s}$. The GREG estimator for the total t_y is given by,

$$(3) \quad \hat{t}_{GREG} = \hat{t}_{y,\pi} - \left(\hat{t}_{x,\pi} - t_x\right)' \hat{\boldsymbol{\beta}}_{OLS,\pi} = \sum_s w_i y_i$$

where $\hat{t}_{y,\pi} = \sum_{i\in s} \dfrac{y_i}{\pi_i}$, respectively $\hat{t}_{x,\pi} = \sum_{i\in s} \dfrac{\mathbf{x}_i}{\pi_i}$, is the Horvitz-Thompson (HT) estimator of t_y, respectively of $t_x = \sum_U \mathbf{x}_i$. The GREG estimator is a weighted estimator with weights not depending on the study variable,

$$(4) \quad \mathbf{w}_s = (w_i)_{i\in s} = \mathbf{d}_s - {\boldsymbol{\Pi}_s}^{-1}\mathbf{X}_s(\mathbf{X}_s'{\boldsymbol{\Pi}_s}^{-1}\mathbf{X}_s)^{-1}(\mathbf{d}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})$$

where $\mathbf{d}_s = (d_i)_{i\in s}$ with $d_i = 1/\pi_i$ is the sampling weight vector and $\mathbf{1}_U'$ is the N-dimensional vector of ones.

An alternative approach is the *calibration approach*. This method proposed by Deville and Särndal (1992) consists in calculating the calibration weights $\mathbf{w}_s^{cal} = (w_i^{cal})_{i\in s}$ such that they are as close as possible to the sampling weights $\mathbf{d}_s$ and subject to the constraints of calibration on the auxiliary variables. For the chi-square distance, we obtain the $\mathbf{w}_s^{cal}$ as solution of the following optimization problem

$$\mathbf{w}_s^{cal} \quad = \quad \mathrm{argmin}_{\mathbf{w}} \sum_s \frac{(w_i - d_i)^2}{d_i q_i} \quad \text{et} \quad \mathbf{w}_s'^{cal}\mathbf{X}_s = \mathbf{1}_U'\mathbf{X}$$

where q_i is a constant used to control the variability of the observations. In most of applications, $q_i = 1$ and in this situation the weights $\mathbf{w}_s^{cal}$ are similar to the $\mathbf{w}_s$ given by relation (4).
It is well-known that the GREG or the calibration estimator improve the Horvitz-Thompson estimator if the relation between the variable of interest and the auxiliary information is well-explained by the model ξ. Nevertheless, this estimator is not robust if the $\mathbf{X}$ matrix is ill-conditioned or if there exists a nearly linear relation between the columns of X (multicollinearity). This can happen if a large number of explanatory variables are used or if the variables contain many zeros. In these cases, the estimation of β given by (2) is very instable and the weights $\mathbf{w}_s^{cal}$ or $\mathbf{w}_s$ may be negative or very large. From a calibration point of view, this results when a very large number of calibration equations is used.

Ridge Regression in Survey Sampling

In order to circumvent this problem, Bardsley and Chambers (1984) in a model-based setting and Rao and Singh (2009) in a model-assisted setting have proposed a class of penalized estimators. The Rao and Singh's estimator is obtained as a solution of the following minimization problem

$$\mathbf{w}^c = \text{argmin}_{\mathbf{w}}(\mathbf{w}-\mathbf{d}_s)'\tilde{\mathbf{\Pi}}_s(\mathbf{w}-\mathbf{d}) + (\mathbf{w}'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})\mathbf{D}(\mathbf{w}'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})'$$

where $\tilde{\mathbf{\Pi}}_s = \text{diag}(q_k)_{k\in s}^{-1}\mathbf{\Pi}_s$ and $\mathbf{D}$ is a positive diagonal cost matrix. In a model-based setting, the matrix $\tilde{\mathbf{\Pi}}_s$ is replaced by the variance-covariance matrix of errors $(\varepsilon_i)_{i\in s}$ (Bardsley and Chambers, 1984). In this way, we penalize the large values of $\sum_s w_i^c\mathbf{x}_i' - \sum_U \mathbf{x}_i'$ and we eliminate the possibility of having very large or negative weights. This approach is equivalent to construct a GREG estimator given in (4) with the regression coefficient estimated by a ridge estimator (Hoerl and Kennard, 1970). More precisely, this consists in adding a positive diagonal cost matrix to the diagonal of the matrix $\mathbf{X}_s'\mathbf{\Pi}_s^{-1}\mathbf{X}_s$, $\hat{\beta}_{MA,R} = \left(\mathbf{X}_s'\mathbf{\Pi}_s^{-1}\mathbf{X}_s + \mathbf{D}^{-1}\right)^{-1}\mathbf{X}_s'\mathbf{\Pi}_s^{-1}\mathbf{y}_s$. The ridge estimator of $\boldsymbol{\beta}$ is ξ-biased but is more stable in presence of multicollinearity.

3 Principal Components Regression in Survey Sampling

3.1 Model-assisted approach

We suppose without loss of generality that the auxiliary variables are standardized, namely $\mathbf{1}_U'\mathbf{X}_j = 0$ and $\mathbf{X}_j'\mathbf{X}_j = 1$ for all $j = 1,\ldots,p$ 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) (Jolliffe, 2002). The PCR consists in reducing the space spanned by the columns of $\mathbf{X}$ and consider the regression model ξ' over the reduced space. We consider the eigenvectors $\mathbf{v}_1,\ldots,\mathbf{v}_r$ corresponding to the largest eigenvalues $\lambda_1 \geq \ldots \geq \lambda_r > 0$ of the matrix $\mathbf{X}'\mathbf{X}$. Let $\mathbf{z}_j = \mathbf{X}\mathbf{v}_j = (z_{ji})_{i\in U}$ for $j = 1,\ldots,r$ be the first r principal components and $\mathbf{Z}_r = (\mathbf{z}_1,\ldots,\mathbf{z}_r)$. The new model consists in regressing $\mathbf{y}$ on $\mathbf{Z}_r$

$$(5) \quad \xi': \quad \mathbf{y} = \mathbf{Z}_r\boldsymbol{\eta} + \boldsymbol{\varepsilon}_r$$

The estimation of $\boldsymbol{\eta}$ is done by least squares, $\hat{\boldsymbol{\eta}} = (\mathbf{Z}_r'\mathbf{Z}_r)^{-1}\mathbf{Z}_r'\mathbf{y}$ and the estimator of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}}_{PC} = (\mathbf{v}_1,\ldots,\mathbf{v}_r)'\hat{\boldsymbol{\eta}}$. This estimator is ξ-biased but its ξ-mean squared error (MSE) is smaller than that of $\hat{\boldsymbol{\beta}}_{OLS}$. Jolliffe (2002) provides criteria for choosing the number of principal components. Let $\tilde{\boldsymbol{z}}_i' = (z_{1i},\ldots,z_{ri})$ be the vector containing the values of the r principal components for the i-th individual and $\mathbf{Z}_r = (\tilde{\boldsymbol{z}}_i')_{i=1}^N$.
The estimator $\hat{\boldsymbol{\eta}}$ can not be calculated since it contains the unknown population vector $\mathbf{y}$. The design-based estimator of $\hat{\boldsymbol{\eta}}$ is given by $\hat{\boldsymbol{\eta}}_\pi = (\mathbf{Z}_{r,s}'\mathbf{\Pi}_s^{-1}\mathbf{Z}_{r,s})^{-1}\mathbf{Z}_{r,s}'\mathbf{\Pi}_s^{-1}\mathbf{y}_s$ where $\mathbf{Z}_{r,s}$ is the restriction of $\mathbf{Z}_r$ on the sample s, namely $\mathbf{Z}_{r,s} = (\tilde{\boldsymbol{z}}_i')_{i\in s}$. We suggest to estimate the total t_y by

$$(6) \quad \hat{t}_{PC} = \hat{t}_{y,\pi} - \left(\hat{t}_{z,\pi} - t_z\right)'\hat{\boldsymbol{\eta}}_\pi$$

where $\hat{t}_{z,\pi} = \sum_s \frac{\tilde{\boldsymbol{z}}_i}{\pi_i}$ is the Horvitz-Thompson estimator of $t_z = \sum_U \tilde{\boldsymbol{z}}_i$. For standardized variables $\mathbf{X}_j$, $j = 1,\ldots,p$ 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 (6) becomes

$$(7) \quad \hat{t}_{PC} = \hat{t}_{y,\pi} - \hat{t}_{z,\pi}'\hat{\boldsymbol{\eta}}_\pi = \sum_s \frac{y_i - \tilde{\boldsymbol{z}}_i'\hat{\boldsymbol{\eta}}_\pi}{\pi_i}$$

which is the Horvitz-Thompson estimator for the sample fit residuals $y_i - \tilde{\boldsymbol{z}}_i'\hat{\boldsymbol{\eta}}_\pi$. 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.

Result 1 *We suppose that* $\hat{\boldsymbol{\eta}}_\pi - \hat{\boldsymbol{\eta}} = o_p(1)$. *The asymptotic variance of* $\hat{t}_{PC}$ *is the variance of* $\hat{t}_{diff} = \hat{t}_{y,\pi} - \left(\hat{t}_{z,\pi} - t_z\right)' \hat{\boldsymbol{\eta}} = \hat{t}_{y,\pi} - \hat{t}'_{z,\pi}\hat{\boldsymbol{\eta}}$,

$$AV(\hat{t}_{PC}) = \sum_U \sum_U (\pi_{ij} - \pi_i\pi_j) \frac{y_i - \tilde{\boldsymbol{z}}_i'\hat{\boldsymbol{\eta}}}{\pi_i} \frac{y_j - \tilde{\boldsymbol{z}}_j'\hat{\boldsymbol{\eta}}}{\pi_j}$$

3.2 Calibration with Principal Components

An estimator using the calibration approach can be given. The vector of auxiliary information is now composed of the first r principal components $\mathbf{z}_1, \ldots, \mathbf{z}_r$. More exactly, we construct the estimator $\hat{t}_w = \sum_s w_i^c y_i$ calibrated on the finite totals of the principal components $\mathbf{z}_j$, $j = 1 \ldots, r$ instead of $\mathbf{X}_j$, $j = 1, \ldots, p$ variables. So, the weights $\mathbf{w}^c = (w_i^c)_{i \in s}$ satisfy

$$\mathbf{w}^c = \text{argmin}_{\mathbf{w}} \sum_s \frac{(w_i - d_i)^2}{d_i q_i} \quad \text{and} \quad \mathbf{w}'^c \mathbf{Z}_{r,s} = \mathbf{1}_U' \mathbf{Z}_r \tag{8}$$

The calibration weights obtained in this way will not allow to find exact totals of the initial auxiliary variables. This property is verified in the projection space on $\mathbf{Z}_r$.

Calibration on second moment of the principal component variables

An interesting extension of the classical calibration approach can be obtained noting that the principal components variables have the following property:

$$\mathbf{z}_j'\mathbf{z}_j = \sum_{i \in U} z_{ji}^2 = \lambda_j, \quad \text{for all} \quad j = 1, \ldots, p$$

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

$$\begin{aligned} \mathbf{w}^c &= \text{argmin}_{\mathbf{w}} \sum_s \frac{(w_i - d_i)^2}{d_i q_i} \quad \text{and} \\ \mathbf{w}'^c \mathbf{Z}_{r,s} &= \mathbf{1}_U' \mathbf{Z}_r, \quad \mathbf{w}'^c \mathbf{Z}_{r,s}^2 = \mathbf{1}_U' \mathbf{Z}_r^2 \end{aligned}$$

where $\mathbf{Z}_{r,s}^2$ is the sample restriction of $\mathbf{Z}_r^2$. The solution is given by

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

where $\tilde{\boldsymbol{\Pi}}_s = \text{diag}(q_k)_{k \in s}^{-1} \boldsymbol{\Pi}_s$ and the $n \times 2r$ matrix $\mathbf{T}_{r,s} = (\mathbf{Z}_{r,s}, \mathbf{Z}_{r,s}^2)$. 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

$$(9) \quad \hat{t}_{PC}^{c} = \mathbf{w}^{c'}\mathbf{y}_s = \mathbf{d}_s'\mathbf{y}_s - \left(\mathbf{d}_s'\mathbf{T}_{r,s} - \mathbf{1}_U'\mathbf{T}_r\right)\left(\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s}\right)^{-1}\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s$$
$$= \hat{t}_{y,\pi} - \left(\sum_s \frac{\mathbf{t}_i}{\pi_i} - \sum_U \mathbf{t}_i\right)' \hat{B}_{z,z^2}$$

where $\mathbf{t}_i = (\tilde{\boldsymbol{z}}_i', \tilde{\boldsymbol{z}}_{\boldsymbol{i}}^{\mathbf{2}\prime})$ is the i-th row of $\mathbf{T}_r$ and $\hat{B}_{z,z^2} = \left(\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{T}_{r,s}\right)^{-1}\mathbf{T}_{r,s}'\tilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s$. The estimator derived in this way is expected to perform better that the estimator calibrated only on the first moment of the principal components.

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 build of 19 columns and is ill-conditioned. The GREG estimator does not always work because the $\mathbf{X_s'\Pi}_s\mathbf{X_s}$ matrix has the minimum eigen-value λ_{min} equal to zero for many samples. We have therefore, compared principal component and ridge estimators on 1000 samples through the relative-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. The r number of principal components were chosen in function of $\lambda_j / \sum \lambda_j$ and the constant k by using the ridge trace (Hoerl and Kennard, 1970). For $\hat{t}_{PC}$ with $r = 15$ principal components, we obtain $\frac{MSE(\hat{t}_{PC})}{MSE(\hat{t}_{HT})} = 0.56$. We trace in Figure 1 (b), the ratio between the MSE of $\hat{t}_{ridge}$ and the MSE of $\hat{t}_{HT}$ for many values of k 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 k, the $\hat{t}_{ridge}$ estimator approaches to $\hat{t}_{HT}$.

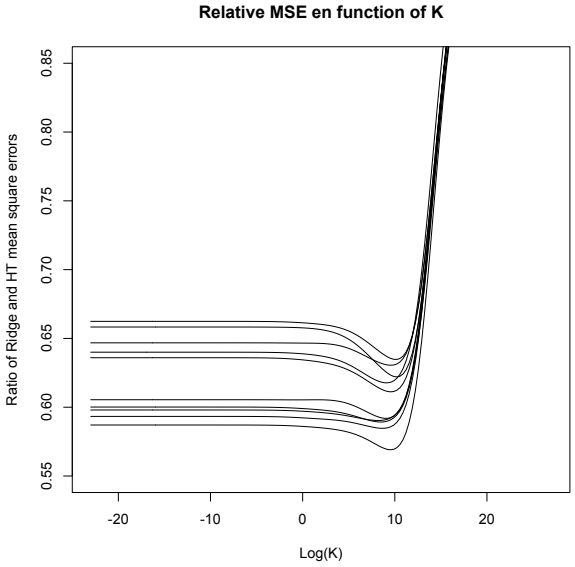

Figure 1: Ratio of the means square errors between the ridge and the Horvitz-Thompson estimators

REFERENCES (RÉFERENCES)

Bardsley Bardsley, P. and Chambers, R. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, **33**, 290-299.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association 87, 376382.

Hoerl, E., and Kennard, R. W. (1970). *Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics*, 12, 55-67.

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.

Jolliffe, I.T (2002). *Principal Components Analysis*, Second Edition, New York: Springer- Verlag.

Rao, J.N.K. and Singh, A. C. (2009). *Range Restricted Weight Calibration for Survey Data Using Ridge Regression.* Pakistan Journal of Statistics, 25(4), 371-384.

Särndal, C. E., Swensson, B., and Wretman J. (1992). *Model Assisted Survey Sampling.* Springer-Verlog, New York Inc.