STUDYING THE TOTAL UNDER MISSINGNESS BY GUESSING THE VALUE OF A SUPERPOPULATION MODEL FOR IMPUTATION

Carlos, N. Bouza[1]* , Carmen Viada ** and Gajendra K. Vishwakarma***
*Universidad de La Habana, Cuba.
** Dpto. Gestión de la Información Clínica, Dirección de Investigaciones Clínicas. Centro de Inmunología Molecular. Cuba
***Department of Mathematics & Computing, Indian Institute of Technology Dhanbad, India

ABSTRACT
We propose to use a superpopulation simple regression model. The regression coefficient is predicted by the researcher. The missing values of the variable of interest are predicted using the model. The behavior of the proposed predictor is evaluated by determining its the design expectations of the model bias and variance. A numerical study is developed using real life data.

KEYWORDS: superpopulation, imputation, predictor, missing data, assertiveness. Covid19, Body Index Mass.

MSC: 62D05

RESUMEN
Proponemos el uso de un modelo superpoblacional del tipo simple regresión. En la regresión el coeficiente es una predicción del investigador. Los valores de la variable de interés cuando se pierde la información son predichos usando el modelo. El compartimento del propuesto predictor es evaluado al determinar su esperanza bajo el diseño así como los sesgos y varianza. . Un estudio numérico es desarrollado usando datos de la vida real.

PALABRAS CLAVE: superpoblación, imputación, predictor, data faltante, asertividad, Covid19, Índice de Masa Corporal.

1. INTRODUCTION

Statistical inference is based upon and bounded by the theory developed on the basis of a certain approach; frequentist and Bayesian are the most popular. Sampling theory considers that the uncertainty in the data is introduced by the sampling design. The statistician determines the sampling design.
The selection of the sample determines a set of individuals. They should provide information on variables of interest. Commonly, some of the measurements are not obtained by different causes. List wise deletion is the default method implemented in most statistical software packages. But excluding some cases the statistical analysis are doubtfully correct. A solution is to substitute the missing data using imputation methods . Imputation provides complete data for developing subsequent analysis. As the imputed data incorporate the needed information is expected that the analysis would be statistically efficient and coherent. See Schenker and Raghunathan (2007) for a discussion of these aspects.
Commonly data are gathered from a finite population U. It may be referred to as being “generated by a stochastic model”. Hence we consider that a realization from a superpopulation model M.” The presentation of statistics uses the term infinite population but the discussions does not pictures to us clearly that it is a part of the enumeration of a finite population. Finite sampling theory deals with a well-defined finite population $U = \{u_1, \dots u_N\}$. Each unit u_i is perfectly identifiable in advance by the research. The variables involved changes, the population is dynamic. The concept of superpopulation is needed to differentiate between a finite population and an infinite superpopulation.
In sampling from a finite population, we often may find a reasonable probability model (“superpopulation model”) that characterizes relations among variables of the units of the population. For example, a physician with experience has knowledge of how each patient will recover from a certain viable in his records.

[1] bouza@matcom.uh.cu

We are going to use a superpopulation regression model, where is a guessed the regression coefficient, as an imputation mechanism to estimate totals. We hypothesize that the researcher may determine a plausible superpopulation model. This superpopulation is determined by fixing a value of the involved parameters. The value of them are to be predicted (guessed) in basis of the experience of the researcher. Hence, the method we are considering substitutes each missing value by a prediction, which depends on the parameter that is fixed by the researcher.
We analyze the behavior of our proposal for determining a total under missing observations. The effect of the accuracy off the guessed parameter in the statistical analysis is determined in terms of biases and means squared errors. We study them under the use of model inference at first. Afterwards is calculated the sampling design effect when random samples are considered as a new source of uncertainty. Therefore, the design expectations of the model bias and variance are computed. Finally are established regularities of the expectations under a particular non-response mechanism.
The next section presents some ideas on the philosophy of imputation procedure and superpopulation modeling. In the third section we develop a study of the prediction of the sample total under a MCMAR missing observation mechanism. The expectation of the measures of accuracy if the sample is selected y using an independent selection of the units (simple random sample with replacement, SRSWR). Considering that the response probabilities are constant are derived the overall expected values of biases and mean squared errors.
From the discussions is established that the main role in guarantying small biases and MSE´s is played by the accuracy of the guessed parameter. A numerical study is developed in the last section.

2. SOME ISSUES ON IMPUTATION AND SUPERPOPULATION PROCEDURES

In common statistical practice we deal with a finite population U of N units. Each unit j provides a pair of values X_j, Y_j . The X-variable is known or obtainable for any j, while for some units Y may be missing. The population total

$$T = \sum_{i \in U} Y_i$$

is of interest but a census may not be developed. Is usual that the researcher obtains a sample s for analyzing the behavior of Y. Different approaches to point estimation may be adopted in the presence of non-response. Some methods just ignore the non-response. In the case of unit non-response this will usually involve treating the set of responding units as if it were the selected sample
Having non responses generates three possible decisions

1. Use only the available data
2. Select a subsample among the missing observations
3. Impute the missing values of Y.

The first decision is very risky as the non-respondents may have a completely different behavior than the respondents. In such cases subsampling is the best solution form the statistical point of view, but it is costly. Having an adequate imputation method may solve the problem. Using the information on X and of the data obtained the statistician may consider that having the total of the data in the sample allows characterizing the problem under study. In many applications this is the main objective of the inquiry. We will consider that the researcher in principle does not need computing the total of the variably Y in the population. Hence we compute

$$t = \sum_{j=1}^{n} Y_j$$

and its knowledge allows to evaluate the behavior of the phenomena. For example the physician is evaluating the result of a medicament and the interest is evaluating if it is provides the expected improvement in the patients. Having t a function g (.) is evaluated. The number of evaluated persons may play an important role in the evaluation of it, as some methods are sensitive to the lack of units. That is the case in many medical researches. The investigation must be based in a certain minimum number of observed units for being credible or for having sufficiently large degrees of freedom.
Missing data is problem which arises in many real world application of statistics. Imputation the missing

values is a way deal with data set . Rubin (1976) fixed the existence of three possible types of missingness mechanism:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR).

MCAR and MAR are in class of ignorable missingness mechanism but MNAR is a non-ignorable type of mechanism. Though MCAR assumption is generally difficult to meet in reality in some particular studies where the sampling nits are under control in a laboratory or in particular populations as the patients in an experiment with a new drug. An experimented researcher may support that there is no statistically significant difference between incomplete and complete cases. He/she considers that in their study missingness is due to a chance mechanism.

Commercial softwares include some imputation methods in their library, see for example a: ICE, Imputation with Chained Equations (Stata) , SAS; IVEware: Imputation and Variance Estimation Software, R Packages (MICE, Amelia, missForest, Hmisc, mi). See Raghunathan, et al. (2016), Waljee, et al. (2013), Rickert, (2016) .

The evaluation of the sample determines whether a unit provides information or not. This collection of random variables is called a superpopulation. Considering that the behavior of the random variables is described by a certain probability structure. This structure is often stated in terms of the so called superpopulation model. Deming and Stephan (1941) introduced the term but Cochran in 1939 used this approach at first see Cochran (1946). They agree in considering that the finite population under study was drawn from a larger universe. In such cases the parameters of the superpopulation have a statistical meaning as different sets of N subjects will arise from the realization fo the superpopulation. Särndal (1992) used this concept of superpopulation to consider that it is an abstract representation of "a broader entity from which the population values are generated" . In this context the superpopulation represents a causal system. Then, the potentially observed random variables and the missing ones are both described by the superpopulation model. The researcher should be able to fix such probabilistic structure, the involved assumptions, and a consensus can be reached as to constitute the 'best' guess on the model to be assumed.

3. PREDICTING THE TOTAL UNDER MISSING DATA

Consider that we have a sample of n individuals selected from a population of size N, then the sample total of the variable of interest Y is given by

$$t=\sum_{i=1}^{n} y_i$$

Assume that an auxiliary variable X is known for all the individuals in the population. This variable is related with Y by means of a superpopulation model. That is commonly the case in medical research where X appears in the files of the patients. We will consider the superpopulation model

$$M\colon\ Y_i = BX_i + e_i, E(e_i e_j) = \begin{cases} \sigma^2\ if\ i=j \\ 0\ otherwise \end{cases}$$

Commonly the information of some sample units is missing and t is not computable. Some notation is needed. Take

$s_1=\{i\in s|y_i\ is\ obtained\}, s_2=\{i\in s|y_i\ is\ missing\}, n_j=|s_j|>0, j=1,2$

$$\bar{z}_j=\frac{1}{n_j}\sum_{i=1}^{n_j} y_i\,, j=1,2; z=x,y$$

We only can compute the mean of Y for the individuals in s_1, but the mean of X may be obtained for the whole sample $s=s_1+s_2$. The sample total t may be rewritten as

$$t=\sum_{i=1}^{n} y_i = n_1\bar{y}_1 + n_2\bar{y}_2 = t_1+t_2$$

Using the information obtained and the superpopulation model M we may impute the missing values of Y by using the predictor $\hat{y}_i = Bx_i$. Generally B is unknown. Consider that B may be approximated using information available for guessing an appropriate value B_0. For example in a longitudinal study the physician may fix a “guessed vale” B_0 . He/she may be considering that the value of B, obtained in the previous evaluation, say B_a , of the patients should be incremented in such a way that $B_0 = \gamma B_a$.

From these facts and using the superpopulation M what is possible is to use $\hat{y}_{i0} = B_0 x_i$ and compute

$$t_{02} = n_2 \bar{y}_{20} = \sum_{i=1}^{n_2} B_0 x_i$$

The respondents may be used for having an idea of the value of the residuals in the predictions made by using the guessed parameter B_0. The “guessed” residual is $e_{0i} = y_i - B_0 x_i$. They may be computed only if $i \in s_1$. The $e_{0i}`s$ may be used for centering the predictor t_{02}. Our proposal is using as predictor of t

$$t^* = n_1 \bar{y}_1 + n_2 \left[\frac{1}{n_1} \left(\sum\nolimits_{i=1}^{n_1} y_i - B_0 x_i \right) + \frac{1}{n_2} \sum_{i=1}^{n_2} B_0 x_i \right]$$

Let us consider the effect of using this predictor of t.

$$D = t^* - t = n_2 \left[\frac{1}{n_1} \left(\sum\nolimits_{i=1}^{n_1} (y_i - B_0 x_i \right) + \frac{1}{n_2} \sum_{i=1}^{n_2} B_0 x_i \right] - n_2 \bar{y}_2$$

is the difference obtained when predicting t. The model bias is

$$E_M(t^* - t) = E_M \left\{ n_2 \left[\frac{1}{n_1} \left(\sum\nolimits_{i=1}^{n_1} B x_i + e_i - B_0 x_i \right) + \frac{1}{n_2} \sum_{i=1}^{n_2} B_0 x_i \right] - n_2 B \bar{x}_2 - n_2 \bar{e}_2 \right\}$$

$$= \delta \left[\frac{n_2}{n_1} \sum\nolimits_{i=1}^{n_1} x_i - \sum_{i=1}^{n_2} x_i \quad \right], \delta = B - B_0$$

Hence we may write the model bias as $B_M = n_2 \delta [\bar{x}_1 - \bar{x}_2]$. Then we have that the proposed predictor is approximately unbiased if:

1. $B \cong B_0$, say that the guessed value of the parameter is close to the true B.
   $or\ if$
2. $\bar{x}_1 \cong \bar{x}_2$, say that the strata of non respondents and of respondents have a similar value of the auxiliary variable

For a fixed sample

3. $E(E_M(t^* - t)|s) = n_2 \square (\mu_{X_1} - \mu_2)$

Hence accepting that the respondent and non-respondent strata have the same mean of the auxiliary variable the expected bias is zero. Note that the second condition may be checked once the sample is evaluated. Clearly n_2 is distributed according to the Binomial $B(n, W_2)$, then

$$E[E_d(E_M(t^* - t)|s)] = n W_2 \delta (\mu_{X_1} - \mu_2) = Bias(t^*)$$

Doing some calculus is derived the model ´s mean squared error, MMSE

$$E_M(t^* - t)^2 = E_M \left\{ n_2 \left[\frac{1}{n_1} \left(\sum\nolimits_{i=1}^{n_1} B x_i + e_i - B_0 x_i \right) + \frac{1}{n_2} \sum_{i=1}^{n_2} B_0 x_i \right] - n_2 B \bar{x}_2 - n_2 \bar{e}_2 \right\}^2$$

$$= n_2^2 E_M \left\{ \left[\frac{}{n_1} \left(\sum\nolimits_{i=1}^{n_1} \delta x_i + e_i \right) - \frac{\delta}{n_2} \sum_{i=1}^{n_2} x_i \right]^2 - 2 \left[\frac{}{n_1} \left(\sum\nolimits_{i=1}^{n_1} \delta x_i + e_i \right) - \frac{\delta}{n_2} \sum_{i=1}^{n_2} x_i \right] \bar{e}_2 \right.$$

$$\left. + \bar{e}_2^2 \right\}$$

let us denote

$$H = \left[\left(\frac{1}{n_1} \right)^2 \left(\sum\nolimits_{i=1}^{n_1} \delta x_i + e_i \right)^2 + \left(\frac{\delta}{n_2} \sum_{i=1}^{n_2} x_i \right)^2 - \frac{2}{n_1 n_2} \left(\sum\nolimits_{i=1}^{n_1} \delta x_i + e_i \right) \sum_{i=1}^{n_2} \delta x_i \right] = A + B - C$$

$$A = \left(\frac{1}{n_1}\right)^2 \left\{\sum\nolimits_{i=1}^{n_1} \delta x_i + \sum\nolimits_{i=1}^{n_1} \quad e_i\right\}^2$$

$$= \left(\frac{1}{n_1}\right)^2 \left[\delta^2\{\sum_{i=1}^{n_1} x_i^2 + \sum\nolimits_{i\neq j=1} x_i x_j\} + \left(\sum_{i=1}^{n_1} e_i^2 + \sum\nolimits_{\neq j=1} e_i e_j\right) + 2\sum\nolimits_{i=1}^{n_1} \delta x_i \sum\nolimits_{i=1}^{n_1} \quad e_i\right]$$

As the expected value of the errors are null and they are independent:

$$E_M(A) = \left(\frac{1}{n_1}\right)^2 E_M \left[\sum_{i=1}^{n_1} \delta^2 x_i^2 + \sum\nolimits_{i\neq j=1} \delta^2 x_i x_j + \sum_{i=1}^{n_1} e_i^2 + \sum\nolimits_{i\neq j=1} e_i e_j + 2\,\delta \quad \sum\nolimits_{i=1}^{n_1} x_i \sum\nolimits_{i=1}^{n_1} \quad e_i\right]$$

$$= \left(\frac{\delta}{n_1}\right)^2 [\sum_{\square=1}^{n_1} x_i^2 + \sum\nolimits_{i\neq j=1} x_i x_j] + \frac{\sigma^2}{n_1}$$

B is model constant, and we have:

$$E_M(B) = \left(\frac{\delta}{n_2}\sum_{i=1}^{n_2} x_i\right)^2$$

Developing C is obtained that

$$C = \frac{2}{n_1 n_2}\left(\delta^2 \sum\nolimits_{i=1}^{n_1} x_i \sum_{i=1}^{n_2} x_i + \delta \quad \sum_{i=1}^{n_2} x_i \sum\nolimits_{i=1}^{n_1} e_i\right)$$

Its model expectation is

$$E_M(C) = \frac{2\delta^2}{n_1 n_2}\left(\sum\nolimits_{i=1}^{n_1} x_i \sum_{j=1}^{n_2} x_j\right)$$

Then

$$E_M(H) = \left(\frac{\delta}{n_1}\right)^2 [\sum_{i=1}^{n_1} x_i^2 + \sum\nolimits_{i\neq j=1} x_i x_j] + \frac{\sigma^2}{n_1} + \left(\frac{\delta}{n_2}\sum_{i=1}^{n_2} x_i\right)^2 - \frac{2\delta^2}{n_1 n_2}\left(\sum\nolimits_{i=1}^{n_1} x_i \sum_{i=1}^{n_2} x_i\right)$$

$$= \delta^2[(\bar{x}_1 - \bar{x}_2)^2] + \frac{\sigma^2}{n_1}$$

Now we may compute the MMSE. It is

$$E_M(t^* - t)^2 = n_2^2 \delta^2[(\bar{x}_1 - \bar{x}_2)^2] + n_2^2 \frac{\sigma^2}{n_1} + n_2\,\sigma^2 = n_2^2\delta^2[(\bar{x}_1 - \bar{x}_2)^2] + \frac{n_2 n \sigma^2}{n_1}$$

Note that having a guessed parameter close to B or being similar the means of the two samples the MMSE depends only of the ratio of the subsample sizes and the model variance. For a relatively large number of respondents, $n_1 \gg n_2$, the model error will be smaller.

The Decision Maker may decide to use only the Model criteria for predicting t. In such case if $\bar{x}_1 - \bar{x}_2 \cong 0$ the error is only a function of the last term.

As the studied population U as a stratified one, we have that:

$U = U_1 \cup U_2;\ U_1 = \{i \in U | a\ response\ is\ obtained\ \}, U_2 = \{i \in U | a\ response\ is\ not\ obtained\}$,

$N_j = |U_j| > 0, j = 1{,}2; N = N_1 + N_2$.

Then we have that $W_j = \frac{N_j}{N}, j = 1{,}2$, are the response and non-response probabilities.

Consider that the sample was selected from U using as sampling design simple random sampling with replacement. We detect non-responses after selecting the sample and we deal with a post stratification sampling design. Note that the means of the auxiliary variables are random. The design-model bias is now a function of the strata parameters of X : $\mu_{X_j},\ \sigma_{X_j}^2;\ j = 1{,}2$. Therefore the design expected bias and design model MSE are easily derived as

$$E_d[E_M(t^* - t)|s] = E_d E_M\left\{n_2\delta\left[\frac{\delta}{n_1}\sum_{i=1}^{n_1} x_i - \frac{\delta}{n_2}\sum_{i=1}^{n_2} x_i\right]\right\} = n_2\delta(\mu_{X_1} - \mu_{X_2})$$

$$E_d[E_M(t^* - t)^2|s] = n_2^2\delta^2\left[(\mu_{X_1} - \mu_{X_2})^2 + Var(\bar{x}_1 - \bar{x}_2)\right] + \frac{n_2 n\sigma^2}{n_1}$$

If we use SRSWR

$$E_d[E_M(t^* - t)^2|s] = n_2^2\delta^2\left[(\mu_{X_1} - \mu_{X_2})^2 + \frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}\right] + \frac{n_2 n\sigma^2}{n_1}$$

Note that we really do not know the parameters of X within the strata.

Usually the interest is in fixing an approximate value of the population mean of Y, μ_Y. From the derived results we recommend using

$$\bar{y}^* = \frac{t^*}{n} = w_1\bar{y}_1 + w_2\left[\frac{1}{n_1}\left(\sum_{i=1}^{n_1}(y_i - B_0 x_i)\right) + \frac{1}{n_2}\sum_{i=1}^{n_2} B_0 x_i\right];\ w_j = \frac{n_j}{n}, j = 1,2$$

Note that it mimics the estimator proposed by Hansen-Hurvitz (1946) when using a subsample from the non-respondent stratum.

An analysis of $EE_d[E_M(t^* - t)^2|s]$ is a more complicated as we need to obtain an approximation to $E(n_2^t/$ *n1, t=1,2.* The usual expansion in Taylor Series for ratios permits fixing that

$$E\left(\frac{n_2^t}{n_1}\right) \cong \frac{E(n_2^t)}{E(n_1)} + V(n_1)\left(\frac{E(n_2^t)}{(E(n_1))^3}\right) - 2\frac{Cov\ (n_2^t, n_2)}{(E(n_1))^2}$$

For t=1 we have

$$E\left(\frac{n_2}{n_1}\right) \cong \frac{W_2(nW_1 + W_1W_2 + 2)}{nW_1^2} = \lambda(n) \to_n \frac{W_2}{W_1}$$

For t=2

$$E\left(\frac{n_2^2}{n_1}\right) \cong \frac{n^2W_1W_2\left[nW_1^2W_2(n+1) + W_1W_2(n+1) - W_1((2-3n)W_2 - nW_1) + 2\right]}{n^3W_1^2}$$

$$= \frac{n^2W_1W_2\left[W_1W_2((n+1)(W_2n+1)) - W_1((2-3n)W_2 - nW_1) + 2\right]}{n^3W_1^3} = \nu(n)$$

Then substituting the expectations, variance and covariance we have, as an approximation to the overall MSE, for n large,

$$E_d[E_M(t^* - t)^2|s] = n_2^2\delta^2\left[(\mu_{X_1} - \mu_{X_2})^2 + \frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}\right] + \frac{n_2 n\sigma^2}{n_1}$$

$$EE_d[E_M(t^* - t)^2|s] = \delta^2(\nu(n)\sigma_{X_1}^2 + nW_2\sigma_{X_2}^2) + \delta^2\nu(n)(\mu_{X_1} - \mu_{X_2})^2 + \lambda(n)n\sigma^2$$

Hence the predictor of the mean has as expected MSE

$$E(MSE(\bar{y}^*)) = \frac{\delta^2(\nu(n)\sigma_{X_1}^2 + nW_2\sigma_{X_2}^2) + \delta^2\nu(n)(\mu_{X_1} - \mu_{X_2})^2 + \lambda(n)n\sigma^2}{n^2}$$

Note that, as

$$\frac{\lambda(n)\sigma^2}{n} \to_n \frac{W_2}{nW_1}\sigma^2$$

$$\frac{\nu(n)}{n^2} \to_n \frac{W_2^3}{nW_1}$$

$$\frac{W_2(nW_1 + W_1W_2 + 2)}{nW_1^2}$$

$$E(MSE(\bar{y}^*)) \to_n \delta^2\left(\frac{W_2^3}{nW_1}\sigma_{X_1}^2 + \frac{W_2\sigma_{X_2}^2}{n} + \right) + \delta^2\frac{W_2^3}{nW_1}(\mu_{X_1} - \mu_{X_2})^2 + \frac{W_2}{nW_1}\sigma^2$$

As a result, having $\delta \cong 0$ grants that the first two terms are negligible. If $\mu_{X_1} - \mu_{X_2} \cong 0$ we will have a small value of the second term of the error. If both relations hold the EMSE only depends of the variance of the model error and the relative size of the non-response stratum.

4. SOME APPLICATIONS

In this section we will discuss 3 real life applications of this model for imputing missing dat.

4.1. A study of assertiveness

Psychologists consider that assertiveness of individuals is a function of their ability to express needs, sentiments, opinions, beliefs, and needs clearly and with honesty, to other persons without affecting their rights. Assertiveness is measured by means applying special psychological tests.

Let us consider an item A_t , t=1,…, T. The test is applied and each interviewed individual reports

$$D(A_t) = \begin{cases} 1 \ \ if\ responds\ yes \\ \ \ 0\ otherwise \end{cases}$$

The simplest way of measuring assertiveness in a person is computing the total of “yes”:

$$I(A) = \sum_{t=1}^{T} D(A_t)$$

In the basic questionnaire for measuring assertiveness T>250 items. In practice it is too long. commonly are used an smaller number of questions. We conducted an experiment with adolescents and adults with ages between 30 and 40 years. The psychologists chose to use two short versions from the scale, which measure assertiveness. After applying them the interviewed were questioned if they considered if short forms were reliable compared with the larger instrument.

Commonly psychologists use I(A) or some variant of it. This procedure difficults the comparison of assertiveness when different questionnaires are used. A naïve solution is computing its sample mean (proportion of yes)

$$\frac{1}{T}\sum_{t=1}^{T} D(A_t)$$

This measure takes values in [0,1].

The research was conducted with:

- A group of 52 adolescents , 24 boys and 28 girls.
- A group of 49 persons aged between 30 and 40, 29 men and 20 women.

A classification of their assertiveness was made with a questionnaire of 60 questions the persons received psychological treatment regularly. Afterwards the persons received orientations from the psychologists in 5 consults. In each one they filled a questionnaire with 20 items and another one with 13.

Take

$$x_j\big(q(k)\big) \quad = \frac{1}{T(q)} \sum_{t=1}^{T(q)} D(A_t(q(k)))$$

as the index of assertiveness obtained in the classification in the consult k by the individual j-th with questionnaire q=1,2. Then $x_j(q(k))$ was obtained in the consult k-th.

The specialists assumed that the superpopulation model, described previously, was adequate. That is

$$Y_j\big(q(k)\big) = B_{q(k)} X_j(q(k)) \quad + \varepsilon_{jq(k)}$$

Using the information provided in the previous consult was determined the value of B_{0k}. The auxiliary variable is the value of the index calculated in the visit k-1. In the visit k was obtained a value of the index for some patients, as missing observations are due to the failure of assisting to the previously concerted consult. In any case they were interviewed afterwards. Hence was computed

$$\bar{Y}(q(k)) = \frac{1}{n}\sum_{j=1}^{n} Y_j(q(k))$$

A question is if the index may be predicted with accuracy. Say, if they were able to predict the results of the missing data fixing a certain function $\vartheta(B_{k-1}) = \tau B_k = B_{0k}$. The psychologists considered that, at k=5,

$$\delta_{q(k)} = B_{q(k)} - B_{0q(k)} \cong 0$$

Say, that they have learned which was the change of the super-parameter . Negative values indicate that the psychologists overestimate the super population parameter .

The results of the research produced table 1.

Table 1: results of $\frac{\delta_{q(k)}}{\bar{Y}_{q(k)}}$ in the consults

		q=	1			q=	2	
Consult	Boys	Girls	Men	Women	Boys	Girls	Men	Women
1	-0,190	0,198	-0,161	-0,181	0,179	0,186	0,151	-0,154
2	-0,185	0,189	-0,165	0,167	0,171	0,185	0,182	0,152
3	0,170	0,186	-0,097	-0,159	0,168	0,218	0,128	-0,136
4	0,040	-0,151	0,041	-0,058	0,053	0,064	-0,087	-0,171
5	0,035	-0,079	0,016	0,050	-0,034	-0,054	-0,071	0,126

From table 1 we have that the efficiency of the predictions of the experts was improved with the experience, as for the last consult we observed a considerably small relative bias, except for women. The index of boys and men seem to be better predictable than girls and women. The adolescents are fairly less predictable than adults. The model was generally better described with questionnaire 2 than by questionnaire 1.

The differences between the means of the respondent and non respondent was measured by

$$\gamma(q(k)) = \frac{\bar{x}_1(q(k)) - \bar{x}_2(q(k))}{\bar{Y}(q(k))}$$

Table 2 presents the obtained values of $\gamma(q(k))$. Note that the values observed do not vary too much among the consults. Men have the largest values while women exhibited general the smallest ones. Adolescents had a very similar pattern

Table 2: results of $\gamma(q(k))$ in the consults

		q=	1			q=	2	
Consult	Boys	Girls	Men	Women	Boys	Girls	Men	Women
1	0,485	0,577	0,480	0,388	0,500	0,419	0,300	0,155
2	0,438	0,740	0,520	0,322	0,505	0,520	0,254	0,195
3	0,581	0,750	0,401	0,360	0,567	0,440	0,276	-0,065
4	0,657	0,757	0,341	-0,154	0,448	0,308	0,214	-0,043
5	0,670	0,878	0,500	0,225	0,316	0,713	0,350	0,050

From table 2 we have that the adolescents non-respondents tend to have the smallest indexes of assertiveness with both questionnaires. Men have the same behavior with both instruments and higher values of the difference than women. Questionnaire 1 was generally associated with larger differences in all the cases. The results suggest that, in general, missing observations were associated with having lower results of the index. Men with small indexes seem to tend to avoid the consults.

4.2. Predicting a risk coefficient for COVID19 in prevalence studies.

Physicians consider that the risk of being positive to COVID-19 of individuals is a function φ of the factors determined by evaluating age, chronic diseases and sex, say $\varphi(a, c, s)$. These factors determines different groups. For example for age we have

$$\begin{aligned} Age: a(0) = baby, a(1) = children, a(2) = adolescents, a(3) \\ = persons\ aged\ beween\ 19\ and\ 30\ years, a(4) \\ = persons\ aged\ between\ 31\ and\ 60, a(5) \\ = persons\ aged\ beween\ 61\ and\ 70\ years, a(6) = older\ than\ 70\ years \end{aligned}$$

$$\begin{aligned} Chronic\ diseases : c(0) = no\ disease, c(1) = being\ diabetic, c(2) = being\ allergic, c(3) \\ = being\ hypertense, c(4) = being\ cardiopatic, c(5) \\ = having\ psycollogic\ disorders, c(6) = cancer\ disturbs, c(7) = other\ diseases \end{aligned}$$

$$Sex: s(0) = transgener, s(1) = man, s(2) = woman$$

The risk is measured by using a function of these factors and of the current contacts with other persons G.

Let us consider the factors A_{tf} , t=0,…, T; f=1,..,F. Then in a visit is evaluated

$$D(A_{tf}) = \begin{cases} 1 \; if\ the\ individual\ is\ positive\ for\ the\ factor\ A_{tf} \\ 0\ otherwise \end{cases}$$

The simplest way of measuring risk in an individual is given by :

$$R(u_i) = \frac{\sum_{f=1}^{F} \quad \sum_{t=1}^{T} D(A_{tf})\varphi_i(a,c,s) + G_i}{(F+1)(T+1)+1}$$

It may be calculated form the file of patient u_i . In a visit the physician evaluates the condition and fix the response

$$Y(u_i) = \frac{\sum_{f=1}^{F} \quad \sum_{t=1}^{T} W(A_{tf})\varphi_i(a,c,s) + G_i}{(F+1)(T+1)+1},$$

$$W(A_{tf}) = \begin{cases} a\ value\ in\ (0, M_{tf}) \;\; if\ the\ indivual\ is\ positive\ for\ the\ factor\ A_{tf} \\ 0\ otherwise \end{cases}$$

The protocols establish whether u_i is to be controlled or not.

This value weights the importance of A_{tf} in terms of tests or subjective criteria. For example if $u_i \in a(j)$ the file permits to establish the value of $D(A_{tf})$ but during the visit is possible to establish if it characterizes the status or through $W(A_{tf})$ change the importance. Consider a person classified in c(4) but being overweighed and with two chirurgical operations the condition is worse and this fact is modeled through $W(A_{tf})$. A person who is not interviewed generates a missing data .The specialists approved that the superpopulation model, described previously, was adequate and that was usable for predicting a variable. It is a consideration of the specialists evaluating the patient. It is given as for an age group c(k) by

$$Y(u_i) = B_k R(u_i) + \varepsilon_i$$

Statistics from 7 specialized hospitals was obtained and the risk of 2056 controlled persons was measured . Using the information provided by them was used for fixing the value of B_{0k} and the missing observations are predicted by

$$\hat{Y}(u_i) = B_{0k} R(u_i)$$

The missing observations are due to the failure of evaluating $Y(u_i)$ when visiting the patient. In any case they were interviewed afterwards and was computed the parameter

$$\bar{Y}(c(k), a(t)) = \frac{1}{\|c(k) \cap a(t)\|} \sum_{j \in c(k) \cap a(t)} Y\ (u_j)$$

A question is the how good is the accuracy of $\hat{Y}(u_i)$. Say if for any patient $\Delta(u_i) = \hat{Y}(u_i) - Y(u_i) \cong 0$. We evaluate the overall accuracy by computing

$$\bar{\delta}(c(k), a(t)) = \frac{1}{\|c(k) \cap a(t)\|} \sum_{j \in c(k) \cap a(t)} \left| \frac{\Delta(u_j)}{\bar{Y}(c(k) \cap a(t))} \right|$$

The results of the research produced table 3.

Table 3: results of $\bar{\delta}(c(k))$ in the analyzed patients

			Chronic	Disease					
Age Group	0	1	2	3	4	5	6	7	Consensus Ranking Of Age
0	0,03	0,02	0,02	0,05	0,09	0,10	0,13	0,22	1
1	0,18	0,24	0,26	0,22	0,17	0,14	0,28	0,28	3
2	0,09	0,05	0,03	0,13	0,19	0,25	0,23	0,23	2
3	0,16	0,12	0,30	0,42	0,47	0,35	0,26	0,33	4
4	0,20	0,42	0,54	0,61	0,50	0,15	0,59	0,49	5
5	0,73	0,56	0,69	0,61	0,58	0,36	0,95	0,70	7

6	0,10	0,75	0,24	0,11	0,81	0,82	0,81	0,77	6
Consensus Ranking of Chronic Disease	2	1	3	4	5	6	7	8	

The consensus allows establishing the behavior of the prediction within the groups of age and Chronic Disease (smaller the better) . As expected by physicians the younger the patient the better the predictions excepting when comparing persons with psychological problems and those with other diseases (c(5) and c(6)). For age a similar behavior was observed but adolescent were better predicted than children and in the third age persons with more than 70 years obtained more accurate predictions than those within 61 an 70 years .

4.3. A study of the Bio Mass Index.

A real data base with diabetics patients controlled by public medical institutions was studied by Bouza-Herrera et al (2019). The population under study was of size N=274 349 persons. The Body Mass Index (BMI) of them were measured when they started to be controlled. Missing values in a sample of n=2 700 patients were observed. The percentage of patients who did not attend the second visit was approximately 22%. They were visited at home for measuring their BMI after a month with treatment. The auxiliary variable X is the BMI in the first visit and Y it in the second one. The physicians accepted that

$$Y_i = BX_i + e_i, E(e_ie_j) = \begin{cases} \sigma^2 \; if \; i = j \\ 0 \; otherwise \end{cases}$$

We imputed the missing values of Y by using the predictor $\hat{y}_i = Bx_i$. The physicians used the responses and the information in the files of the patients for “guessing ” $B_0 = \gamma B_a$.
From these facts and using the superpopulation M what is possible is to use $\hat{y}_{i0} = B_0x_i$ and compute

$$t_{02} = n_2\bar{y}_{20} = \sum_{i=1}^{n_2} B_0x_i$$

The accuracy of the proposal was measured by computing

$$DR = \left|\frac{t^* - t}{t}\right|$$

We performed an experiment with 5 physicians. Each one proposed a regression coefficient B_{0e} .e=1,…,5. See the results of $DR_e = \left|\frac{t_e^* - t}{t}\right|, e = 1, \ldots, 5$ in table 4. We also computed a sample source of the relative model bias. It is $S(x) = \frac{[\bar{x}_1 - \bar{x}_2]}{t} = 1{,}04$.

Table 4: results of $100DR_e$ of BMI in the analyzed patients

Physician (e)	1	2	3	4	5	mean	Standard deviation
$100DR_e$	4,94	3,28	17,8	3.21	3,66	6,578	5,645

Therefore it seems that the fourth physician provided the best-guessed value.

Acknowledgements: The research producing this paper was supported partially by the project cu2018JOt0044104 (A Cuban-Flemish Training and Research Program in Data Science and Big Data Analysis)

RECEIVED: MAY, 2020.
REVISED: OCTOBER, 2020.

[1] REFERENCES

[2] ANDRIDGE, R. R. (2009): **Statistical methods for missing data in complex sample surveys**. Phd thesis, The University of Michigan.

[3] BOUZA-HERRERA, C. N., ALLENDE-ALONSO, S. M., G. K. VISHWAKARMA and N. SINGH (2019): Estimation of optimum sample size allocation: An illustration with body mass index for evaluating the effect of a dietetic supplement. **International Journal of Biomathematics** 12, 108-120.

[4] COCHRAN, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. **Ann. Math. Statist**. 17, 164-177.

[5] DEMING, W. E. and STEPHAN, F. F. (1941). On the interpretation of censuses as samples. **J. Amer. Statist. Assoc**. 36, 45-49.

[6] DORFMAN, A. H. and R. VALLIANT (2005): Superpopulation Models in Survey Sampling. **Encyclopedia of Biostatistics,** 1 Book Editor(s): Peter Armitage and Theodore Colton .

[7] FOX, R, (2016): **Applied regression analysis and generalized linear models,** (3rd ed.), Dage Publications Inc, Thousand Oaks, CA .

[8] GIRA, A. A. (2015): Estimation of Population Mean with a New Imputation Method . **Applied Mathematical Sciences**, 9, 1663 - 1672 .

[9] LITTLE, R. J. A. and RUBIN, D. B. (2002), Statistical Analysis with Missing Data, Wiley: New York, 2nd ed. Kabacoff, R. I. (2011) R in Action: Data analysis and graphics with R, In "**Advanced methods for missing data** Editor Manning, M. . (chapter 15, 352-372).

[10] OMURA, M., MAGUIRE J., LEVETT-JONES T. and STONE T. E. (2017): The effectiveness of assertiveness communication training programs for healthcare professionals and students: A systematic review. **Int J Nurs Stud.**, 76, 120-128.

[11] PALAREA-ALBALADEJO, J., and J.A. MARTÍN-FERNÁNDEZ (2015) : Compositions — R package for multivariate imputation of left-censored data under a compositional approach. **Chemometrics and Intelligent Laboratory Systems,** 143, 85–96

[12] PARHAM, J. B., C. C. LEWIS, C. E. FRETWELL, J. G. IRWIN and M. R. SCHRIMSHER, (2015): Influences on assertiveness: gender, national culture, and ethnicity, Journal of Management Development, 34, 421-439.

[13] RAGHUNATHAN, T., P. SOLENBERGER, P. BERGLUND and J. VAN HOEWYK (2016): **IVEware: Imputation and Variance Estimation Software** (Version 0.3) . Survey Research Center, Institute for Social Research University of Michigan Ann Arbor, Michigan

[14] RICKERT, J. (2016): Missing Values, **Data Science and R**.

[15] SARNDAL, C., SWENSSON, B. and WRETMAN, J. (1992): **Model Assisted Survey Sampling**. Springer-Verlag, New York.

[16] ROYSTON, P. (2004): Multiple imputation of missing values. **The Stata Journal**, 4, 227- 241.

[17] SCHENKER, N. and T. RAGHUNATHAN (2007): Combining information from multiple surveys to enhance estimation of measures of health. **Statist. Med**. 26, 1802–11.

[18] STATACORP (2009): **Stata Statistical Software: Release 11**. Stata Press: College Station, TX, 2009.

[19] WALJEE, A.K., A. MUKHERJEE, A. G SINGAL, Y. ZHANG, J. WARREN, U. BALIS, J. MARRERO, J. ZHU and P. D. R. HIGGINS (2013): Comparison of imputation methods for missing laboratory data in medicine **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3733317/** (Last consulted April 10, 2018)