REVISTA INVESTIGACION OPERACIONAL **VOL. xx , NO.x, xxx-xxx, 202x**

FORTHCOMING PAPER ·#62D05-20-11-01

IMPUTATION OF INDIVIDUAL VALUES OF A VARIABLE USING PRODUCT PREDICTORS

Carlos N. Bouza-Herrera* and Carmen E. Viada**
*MATCOM, Universidad de La Habana, Cuba
**Centro de Ingeniria Molecular, Cuba.

ABSTRACT
Missing data is a common problem present in almost every data collection. It is particularly important in sample survey research. The existence of missing observations (non-response) is solved in many sample surveys using some technique of imputation. They permit replacing the missing data. Several imputation techniques have been developed in the specialized literature. The capacity of them for predicting the mean or a total is the token for their evaluation. This paper presents an imputation rule with the main goal of predicting individual values. An auxiliary variable X is known for all the units and a product type predictor is developed for predicting the value of variable of interest in each non-respondent. The unbiasedness of the predictions is derived and the Mean Squared Errors (MSE`s). Some conclusions are pointed out.

KEYWORDS: Missing data, imputation, product type predictor, unbiasedness Mean Squared Errors.

MSC: 62D05

RESUMEN
La existencia de datos perdidos es un común problema, casi siempre presente al coleccionar datos. Esto es particularmente importante en investigaciones mediante encuestas por muestreo. La existencia de datos perdidos (no-respuestas) resuelta en muchas encuestas usando alguna técnica de imputación. Ellas permiten reemplazar esos datos perdidos. Muchas técnicas de imputación han sido desarrolladas en la literatura especializada. Su capacidad para predecir la media o el total es la piedra angular de su evaluación. Este paper presenta una regla de imputación cuyo objetivo fundamental es el predecir valores individuales. Una variable auxiliar X es conocida para todas las unidades y un predictor del tipo producto es desarrollado para predecir el valor de la variable de interés en cada non-respondiente. La insesgadez de las predicciones son derivadas así como los Errores Cuadráticos Medios (MSE`s). Algunas conclusiones son destacadas.

PALABRAS CLAVE: data faltante, imputación, predictor del tipo producto, insesgadez, Errores Cuadráticos Medios .

1. INTRODUCTION

Survey sampling is commonly used for collecting data and some sampled units may not respond. Missing data are present in many researches and may invalidate the results. Note that statistical methods do not consider the existence of missing data. Deletion of missing units is commonly used for dealing with item nonresponse. Missing values spoil the inference and mislead determining reliable conclusions. Nonresponse identifies that in the data does not provide values of the variable for all the sampled elements. Including only responses leads to biases , loss in accuracy and power. A solution is to follow up nonresponses and re-visits them. This method may be expensive and/or difficult to implement.
Some recent contributions in imputation are Ahmed, et al. (2006), who revised different imputation methods, Al-Omari & Bouza - Herrera, C. (2013) who considered Imputation methods when the correlation coefficient is known, Arnab & Singh (2006) who considered estimating the variance from imputed data, Nath & Singh (2018), Singh & Gogoi (2018) considered particular aspects of imputation rules, Bouza, et al. (2020) developed superpopulation model based procedures. Modern text books include chapters on imputation methods (see Wu& Thompson (2020)) and some others are devoted entirely to presenting the theory and practice of imputation in survey sampling, (see Little & Rubin (2002)). Commercial softwares are concerned with the implementation of imputation rules, see for example SAS (2015).
Section 2 presents a brief discussion on the generalities of imputation procedures.
Section 3 presents the proposed imputation rule. An auxiliary variable X is known for all the units. It uses a product type estimator for predicting the value of variable of interest in a non respondent. The unbiasedness of the predictions is derived and the Mean Squared Errors (MSE`s) are developed determining a proposition and a corollary. The sampling design used for modeling was simple random sampling with replacement (SRSWR)
Finally, some conclusions are pointed out.

2. SOME ISSUES ON IMPUTATION

Significant theoretical advances allow reducing the negative effects of non-response without revisiting non-respondents. See Al-Omari & Bouza - Herrera, C. (2013) , Nath & Singh (2018) Singh & Gogoi, U. (2018). Bouza, et al. (2020) . Imputation methods allow dealing with completing incomplete data. Typically, imputed values substitute missing item values, they are fabricated values. Statistical procedures use data sets which are structurally complete. Hence, imputation allows completing the dataset. Using imputation provides complete data for sustaining that data are consistent with statistical methodologies as is reduced the bias of using only the data coming from the respondents. Nevertheless, may be preferred imputing as doing so affects the statistical properties of the models at a minimum. Say, that the researcher expects obtaining adequate inferences.
The basic aims of imputation is to support obtaining adequate predictions of the missing values. This problem is common in clinical experiments studying effects of new a drug, as patients may abandon them, in agriculture, due to crops destroy due to the effect of natural, in socio-economic or demographic surveys non-response of some individuals in the sample, generated by various causes. When imputation is adequate the imputed values are used and the model treats them as if they were true-observed values. That is: imputed values are treated as if they were observed. The researcher should know that the corrections introduce an extra variability, and that the accuracy of the estimates is generally overstated as well as the power of tests.
Imputation methods may be based on a deterministic method or on a stochastic procedure. Deterministic methods use the same value if units are considered similar. Stochastic imputation assigns randomly the values. Stochastic imputation usually includes auxiliary variables which should be correlated with the variable of interest. Commonly imputation is carried out using ideas coming from ratio, product and regression procedures.
The model generating imputations (missing imputation mechanism) must be taken into account . Consider an independent uniform response mechanism with constant response probability, p. In practice p is rarely known. For each unit in the sample the Bernoulli distributed variable (Response indicator)

$$I(u_i) = \begin{cases} 1 \; if \; the \; unit \; responds \\ 0 \; otherwise \end{cases}$$

identifies the missing-data pattern as non-response occurs in a single response variable.
The missingness may be assumed as ignorable , see Rubin (1976), if the probability that a value is missing is independent of unobserved. Note that this does not mean that the missing values are occurred at random but they the missing values do not depend on Y.
It is non-ignorable when the probability of missingness depends on unobserved information. Rubin (1976) in his seminal paper explained the natures of missing values patterns.
The possible patterns of the missing mechanisms are:

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

MCAR and MAR belong to the class of ignorable missingness mechanism but MNAR is a non-ignorable type of mechanism. A detailed discussion on missing values was considered by Rubin (1987).
In this paper is considered that the missing is complete at random (MCAR).

3. THE PROPOSED PROCEDURE

Take a finite population $U = \{u_1, \dots, u_N\}$. Consider an auxiliary variable X and the variable of interest Y. They are correlated and missing observations of Y for some of the sampled units are present. A sample s of size n is drawn from U using simple random sampling with replacement (SRSWR) . The presence of non-response determines that s is partitioned into s_1 and s_2. Let n_1 be size of responding units' sample and n_2 the number of non-responding units. If $u_i \in s_1$ the values Y_i are observed, while for $u_i \in s_2$ they are missing and is needed to impute the unknown values. Different methods have been developed for imputation with the aid of the auxiliary variable X. Commonly is assumed a MCAR response mechanism. Different ratio , regression methods of imputation have been proposed as methods of imputation. They look for estimating the mean or the total but the interest in imputing the individual values has not been the goal of the existing papers, to our knowledge. In this paper the imputation of individuals is considered having in mind the need of obtaining information on the quality of the life, due to the use new drugs for of cancer post-surgery, of each patient.
Consider the relative errors

$$\varepsilon_y = \frac{y}{\bar{Y}} - 1$$

and

$$\varepsilon_x = \frac{x}{\bar{x}} - 1.$$

Their expectations are

$$E(\varepsilon_y) = 0\,;\;\; E(\varepsilon_x) = 0$$

The expectations of their square are the squares of Coefficients of Variation

$$,E(\varepsilon_y^2) = \frac{\sigma_y^2}{\bar{Y}^2} = C_y^2;\;\; E(\varepsilon_x^2) = \frac{\sigma_x^2}{\bar{X}^2} = C_x^2$$

The expectation of the cross product is

$$E(\varepsilon_x \varepsilon_y) = \frac{\rho \sigma_y \sigma_x}{\bar{Y}\bar{X}} = \rho C_y C_x$$

ρ is the coefficient of correlation between X and Y.

The imputation method proposed is based on the idea present in of the product estimator. The sample allows to compute the sample mean of the respondents.

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i \qquad (1)$$

The value of the auxiliary variable in a non-respondent is known. Therefore, may be computed

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \;\; \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

The existing ratio imputation procures use functions characterized in some sense by the basic idea of using

$$y_i^R = \frac{x_i \sum_{j=1}^{n_1} y_j}{\sum_{j=1}^{n_1} x_j}$$

Product imputation uses some transformation of the basic function

$$y_i^P = \frac{x_i\left(\sum_{j=1}^{n_1} x_j - \bar{x}\right)}{\bar{x}\sum_{j=1}^{n_2} x_j}$$

See a general imputation based on ratio-type exponential functions models developed in Prasad (2017) and a similar result on exponential product-type procedures Prasad, S. (2018). Some other contribution in this line are Singh& Gogoi (2018) and Shukla & Thakur (2008).
The proposed imputation procedure predicts as the value of Y in a non-respondent

$$\hat{y}_i = \frac{\bar{y}_1 x_i}{\bar{X}}$$

Note that now, the mean of the respondents may be written as

$$\bar{Y}\left(1 + \varepsilon_{\bar{y}_1}\right) = \bar{Y}\left[1 + \left(\frac{\bar{y}_1}{\bar{Y}} - 1\right)\right]$$

and the value of the axiliary variable as

$$1 + \varepsilon_{x_i} = 1 + \left(\frac{x_i}{\bar{X}} - 1\right)$$

Hence,

$$\frac{\bar{y}_1 x_i}{\bar{X}} = \bar{Y}\left(1+\varepsilon_{\bar{y}_1}\right)\left(1+\varepsilon_{x_i}\right) = \bar{Y}\left(1+\varepsilon_{\bar{y}_1}+\varepsilon_{x_i}+\varepsilon_{\bar{y}_1}\varepsilon_{x_i}\right)$$

Note that the non-response mechanism is MCAR is implied that $E(\bar{y}_1) = \bar{Y}$
The use of SRSWR sustains the independence between $u_i \in s_1 \; and \; u_j \in s_2$ and sustains that $\bar{y}_1$ *i*s independent of x_i. Therefore,

$$E(\hat{y}_i) = E\left(\frac{\bar{y}_1 x_i}{\bar{X}}\right) = EE\left[\bar{Y}\left(1+\varepsilon_{\bar{y}_1}\right)\left(1+\varepsilon_{x_i}\right)\right] = \bar{Y}E\left[1+\varepsilon_{\bar{y}_1}+\varepsilon_{x_i}+\varepsilon_{\bar{y}_1}\varepsilon_{x_i}\right] = \bar{Y} \qquad (2)$$

For determining the error is needed to calculate the model expectation

$$E(\hat{y}_i-\bar{Y})^2 = \bar{Y}^2E\left[\varepsilon_{\bar{y}_1}+\varepsilon_{x_i}+\varepsilon_{\bar{y}_1}\varepsilon_{x_i}\right]^2 = \bar{Y}^2E\left[{\varepsilon_{\bar{y}_1}}^2+{\varepsilon_{x_i}}^2+2\varepsilon_{\bar{y}_1}\varepsilon_{x_i}\right] \;= \bar{Y}^2\left[\frac{{\sigma_y}^2}{n_1\bar{Y}^2}+\frac{{\sigma_x}^2}{\bar{X}^2}\right] \qquad (3)$$

Using these results the following proposition may be stated.

Proposition: Consider that the sample design d is SRSWR is used for selecting a sample s of size n and n_1 and n_2 non-responses are obtained. The imputation of a missing value is made by using the procedure

$$\hat{y}_i = \frac{\bar{y}_1 x_i}{\bar{x}},\; \bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$$

The imputed value is design-model unbiased for y_i and its expected error is given approximately by

$$E_d[E(\hat{y}_i-\bar{Y}\lfloor s_1)^2] \cong \bar{Y}^2\left[\frac{{\sigma_y}^2}{\bar{Y}^2}\left(\frac{1}{np}+\frac{1-p}{n^2p^2}\right)+\frac{{\sigma_x}^2}{\bar{X}^2}\right]$$

Proof.

The model unbiasedness was obtained previously in (2). Using the approximation of Stephan (1945)

$$E\left(\frac{1}{n_1}\right) \cong \frac{1}{np}+\frac{1-p}{n^2p^2}$$

Substituting in (3) is obtained $E_d[E(\hat{y}_i-\bar{Y}\lfloor s_1)^2]$.•

An estimator of the population mean, using the proposed imputation procedures is

$$\bar{Y}_{imp} = \frac{n_1}{n}\bar{y}_1 + \frac{1}{n}\sum_{j=1}^{n_2}\hat{y}_j$$

From the proposition is derived easily the following corollary

Corollary: Under the model of the above stated proposition $\bar{Y}_{imp}$ is unbiased and

$$MSE\left(\bar{Y}_{imp}\right) \cong {\sigma_y}^2\left(\frac{1}{np}+\frac{1-p}{n^2p^2}\right)+\bar{Y}^2\left(\frac{{\sigma_x}^2}{\bar{X}^2}+\frac{pq}{n}\right)$$

Proof:

As $E(\hat{y}_j|s_1) = \bar{Y}$ and $E_d(\bar{y}_1) = \bar{Y}$ the unbiasedness of $\bar{Y}_{imp}$ *is* proved.
The structure of the error fixes the need of calculating

$$MSE\left(\bar{Y}_{imp}\right) = V_dE\left(\bar{Y}_{imp}\,|s_1\right)+E_dV\left(\bar{Y}_{imp}\,|s_1\right)$$

$V\left(\bar{Y}_{imp}\,|s_1\right)$ was derived, its design expectation is

$$E_d V\left(\bar{Y}_{imp}\,|s_1\right) = E_d[E(\hat{y}_i - \bar{Y}|s_1)^2] \cong \bar{Y}^2\left[\frac{{\sigma_y}^2}{\bar{Y}^2}\left(\frac{1}{np} + \frac{1-p}{n^2p^2}\right) + \frac{{\sigma_x}^2}{\bar{X}^2}\right]$$

From the model expectation of $\bar{Y}_{imp}$ and the fact that n_1 is a Binomial variable with parameters n and p is derived that

$$V_d E\left(\bar{Y}_{imp}\,|s_1\right) = V_d\left(\frac{n_1}{n}\bar{Y}\right) = \frac{\bar{Y}^2 pq}{n}$$

Summing these expressions is obtained the stated structure of the MSE.•

4. CONCLUSIONS

- The proposed imputation procedure grants that the prediction and $\bar{Y}_{imp}$ are unbiased .
- The proposed imputation procedure grants that $E_d[E(\hat{y}_i - \bar{Y}|s_1)^2]$ and $MSE\left(\bar{Y}_{imp}\right)$ decrease as n is increased and/or p tends to 1.
- The proposed imputation procedure grants that $E_d[E(\hat{y}_i - \bar{Y}|s_1)^2]$ and $MSE\left(\bar{Y}_{imp}\right)$ decrease as $\frac{{\sigma_x}^2}{\bar{X}^2}$ decreases.

RECEIVED: NOVEMBER, 2020.
REVISED: FEBRUARY, 2021.

REFERENCES

[1] AHMED, M.S., AL-TITI, O., AL-RAWI, Z. and ABU-DAYYEH, W. (2006): Estimation of population mean using different imputation methods, **Statistics in Transition** 7, 1247-1264.
[2] AL-OMARI, A. I. and BOUZA - HERRERA, C. (2013)Imputation methods of missing data for estimating the population mean using simple random sampling with known correlation coefficient, **Quality and Quantity**, 47, 353-365.
[3] ARNAB, R. and S. SINGH (2006): A new method for estimating variance from data imputed with ratio method of imputation. **Stat. Prob. Lett**., 76, 513-519.
[4] BOUZA C.N. (2008): Estimation of the population mean with missing observations using product type estimators, **Revista Investigación Operacional** 29, 207-233, 2008.
[5] BOUZA, C. N., C. VIADA and G. K. VISHWAKARMA (2020): Studying the total under missingness by guessing the value of a superpopulation model for imputation, **Revista Investigación Operacional** 41, 979-989
[6] LITTLE, R. J. A. and RUBIN, D. B. (2002): **Statistical Analysis with Missing Data**, Wiley: New York, 2nd ed.
[7] PRASAD, S., (2017): Ratio exponential type imputation in sample surveys, **Model Assisted Statistics and Application**, 12, 95–106.
[8] PRASAD, S. (2018): Product exponential method of imputation in sample surveys, **Statistics in Transition New Series,** 19, 159–166.
[9] NATH K. and B.K. SINGH (2018): Population Mean Estimation Using Ratio-cum Product Compromised-method of Imputation in Two-phase Sampling Scheme. **Asian J. Math. Stat**., 11, 27-39.
[10] RUBIN, R. B. (1976): Inference and missing data , **Biometrika** 63, 581-592.
[11] RUBIN, R. B. (1987): **Multiple imputation for non-response in surveys**, John Wiley, New
[12] SAS I NSTITUTE I NC (2015): **SAS/STAT 14.1 User's Guide—the SURVEYIMPUTE Procedure**. SAS Institute Inc., Cary, N.C.
[13] SINGH, B. K. and GOGOI, U. (2018): Estimation of Population mean using ratio-cum-product imputation techniques in sample survey, **Res. Rev. : J. Stat**., 7, 38-49.
[14] SHUKLA, D. and N.S. THAKUR (2008): Estimation of mean with imputation of missing data using factor-type estimator. **Stat. Trans**., 9: 33-48.
[15] WU, C. and THOMPSON, M. E. (2020): **Sampling Theory and Practice**, Springer Nature, Switzerland AG