RESEARCH ARTICLE OPEN ACCESS

Estimation of Missing Data Based on Vector Correlation Coefficient

Hadeel Darweesh [1], Mohamed Dribate [2], Mounzer Boubou [3]
PhD Student [1], Assistant Professor [2], Assistant Professor [3]
Department of Mathematics
Tishreen University - Syria

ABSTRACT
The objective of this research is to propose a new method for estimating missing data during data analysis, which is based on the use of the multi-variable structure of available data by maximizing the value of the vector correlation coefficient expressed by a vector-based on all available data, and then demonstrating the effectiveness of the method by comparing results to the results of the previous studies.

Keywords :— Missing data, vector correlation coefficient, data analysis.

I. INTRODUCTION

In a multi-variable statistical inference, we often encounter missing data, where missing data is: not responding to some items, data entry errors, or a lack of understanding of what the answer is.
This missing data is a problem of analysis or inference, some statisticians find different solutions to this problem, the most important of which are: the removal of forms or individuals from the database, the deletion of the variable with the greatest loss, or other statistical techniques.

II. SOME STATISTICAL TECHNIQUES FOR PROCESSING MISSING DATA:

First, the statisticians have to determine precisely the nature of this missing data whether it is for individual or variable, to determine whether there is independence between not answering the variable or the individual responding, that is, we want to distinguish between the complete loss of data or the loss by chances because the statistical processing varies by the nature of the estimate, we see that in (Simon et Simonoff 1986, Little 1988, Little and Rubin 2002, Dauide. Hawell 2007, Christyn E. Tannenbaum 2009)
That means that we distinguish between two types of unresponsive:

1-**Full nonresponse:**
This means that the individual has not responded to any paragraph of the test, which often occurs when an individual does not exist or refuses to participate for fear or not to appear because of age and health.

2-failure to respond to the paragraph:
the individual responded to some paragraphs and left some of them unresponsive, so we have some responsive partial data and missing partial data, which means that the individual is involved but does not respond to some paragraphs.
We have to choose between two approaches:

The basic procedure for this method is one of the following methods:

a) Based on the arithmetic mean: missing values are replaced by the arithmetic mean of data (Bemaards and Sijtsma 2000, Banyawwad 2011).
b) Based on simple or multiple regression: so, after we find the regression, we predict the value of missing data (Seber 1984, Little 2002, and Banyawwad 2011).
c) By modeling: Data is modeled by the normal distribution rules (Srivastava 1985, Little and Rubin 2002).
d) Based on factor analysis (Kamakura and Wedel 2000).

III. PREVIOUS STUDIES:

When you go back to previous studies of the missing data problem, the first one to look at this problem is (Hancen and Horwitz 1946), then Frane (1976) suggested some solutions of a practiced research nature, returning to non-responsive individuals and changing the requirements for data collection, the little (2002) suggested a way of predicting linear regression.
And Rubin (2002) replaced the missing data with an M value where (M>1) by creating an M table, and through these tables, an M estimate is created, and then estimates are analyzed and their impact is calculated on the missing data.

In(2006), Allison did a study aimed at figuring out the effect of using different compensatory value calculation methods to address the missing data on the metadata, applying the entire random loss mechanism (MCAR) and the random loss mechanism (MAR).

Frinch (2008) did a study that was aimed at showing the efficiency of the different methods of treating the missing data to estimate paragraph parameters in the paragraph response theory.

In this article, we will present a new estimation technique for missing data: it uses the multivariate structure of the data and is based on the maximization of the RV coefficient introduced by Escoufier in 1973. The definition and properties of the RV and the new estimation method shall be considered, and then we will compare on a sample basis the main methods of imputation with the RV method to using two criteria defined by Gleason and Staelin in **1975.**

Probably the most commonly used method is the Buck method, which assigns to the missing value the prediction provided by regression of this variable on the other variables. It can be shown to be equivalent to estimating the missing data by the value that minimizes the distance of Mahalanobis between the individual vector containing the missing data and the mean vector.

This minimization of The function of a vector standard gave us the idea of minimizing a function of a matrix standard to place itself in a general multivariate context.

IV. VECTOR CORRELATION COEFFICIENT:

Let's have $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ is a random vector so it is:

$X^{(1)}$: is a vector $(p \times 1)$

$X^{(2)}$: is a vector $(q \times 1)$

Splitting the vector X into two parts is done naturally for example (Males, Females) (works, does not work) or any other special that divides it into two parts.

Where:

$\mu = E(X) = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}$ is the mean

$\sigma = E(X - \mu)(X - \mu)^{\setminus} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ is covariance matrix.

We define the vector correlation coefficient ρV :

$$\rho V = \rho V(X^{(1)}, X^{(2)}) = \frac{\operatorname{cov}(\sigma_{12}\sigma_{21})}{\sqrt{\operatorname{cov}(\sigma_{11}^2)\operatorname{cov}(\sigma_{22}^2)}}$$

(Escoufier, 1973).

V. BASIC PROPERTIES OF THE VECTOR CORRELATION COEFFICIENT:

1) If we have $p = q = 1$ then $\rho V = \rho^2$ so, it is the simple linear correlation squared.
2) If we have $0 \le \rho V \le 1$ then:
   a) $\rho V = 0$ if and only if $\sigma_{12} = 0$
   b) If $X^{(2)} = AX^{(1)} + b$ then $\rho V = 1$

   Where A is a matrix (p*q) so that $A'A = KI$ where (K) a positive number.
   c) $\rho V(AX^{(1)}, X^{(2)}) = \rho V(X^{(1)}, X^{(2)})$.....(*)

If we have a sample $X_1, X_2, ..., X_n$, we'll define the sample vector correlation coefficient by within (*) replacing coefficients.

With normal estimates in (*) in order to get:

$$RV = RV(X^{(1)}, X^{(2)}) = \frac{\operatorname{cov}(s_{12}s_{21})}{\sqrt{\operatorname{cov}(s_{11}^2)\operatorname{cov}(s_{22}^2)}}$$

Where $s_{ij} = \frac{1}{n-1}\sum\left(X_{\alpha}^{(i)} - \overline{X}_{\theta}^{(i)}\right)\left(X_{\alpha}^{(j)} - \overline{X}_{\theta}^{(j)}\right)'$

$i, j = 1, 2$

$\overline{X}^{(i)}, \overline{X}^{(j)}$: is a vector mean for both (i) and (j) values calculated from values $X_{\alpha}^{(i)}, X_{\alpha}^{(j)}$ where $(1 \le \alpha \le n)$.

We define the matrix Y_1 :

$$Y_1 = \left(X_1^{(1)} - \overline{X}^{(1)}, X_2^{(1)} - \overline{X}^{(1)}, ..., X_n^{(1)} - \overline{X}^{(1)}\right) : p \times n$$

$$Y_2 = \left(X_1^{(2)} - \overline{X}^{(2)}, X_2^{(2)} - \overline{X}^{(2)}, ..., X_n^{(2)} - \overline{X}^{(2)}\right) : q \times n$$

Using Norm $\|E\|$ where $\|E\| = \sqrt{\operatorname{cov} E'E}$

We get $dist(Y_1, Y_2) = \sqrt{1 - RV(X^{(1)}, X^{(2)})}$

$$dist(Y_1, Y_2) = \left\| \frac{Y_1'Y_1}{\sqrt{\operatorname{cov}(Y_1'Y_1)^2}} - \frac{Y_2'Y_2}{\sqrt{\operatorname{cov}(Y_2'Y_2)^2}} \right\|$$

VI. THE METHOD IT DEPENDS ON RV :

Let'shave

$$X = (X_1, X_2, ..., X_n) = \begin{pmatrix} X_1^{(1)}, X_2^{(1)}, ..., X_n^{(1)} \\ X_1^{(2)}, X_2^{(2)}, ..., X_n^{(2)} \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

Is a data matrix formed by a vector of individuals of dimensions (n).

Now we are going to assume that part of the first set of a variable for the vector X_n contain missing data, we are going to call this unknown vector (x) , then we get $X_n' = (x', y', z')$ where y' is the full part of the first set of variables.

The suggested method is to replace (x) with a vector that reduces distance $dist(Y_1, Y_2)$, which means reducing distance $RV(X^{(1)}, X^{(2)})$

$$S_n = \frac{n-2}{n-1} S_{n-1} + \frac{1}{n}(X_n - \overline{X}_{n-1})(X_n - \overline{X}_{n-1})'$$

Let's have: $\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \frac{1}{\sqrt{n}}(X_n - \overline{X}_{n-1})$

$$\frac{n-2}{n-1} S_{n-1} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}$$

So:

$$S_n = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} + \begin{pmatrix} uu' & uv' & uw' \\ vu' & vv' & vw' \\ wu' & wv' & ww' \end{pmatrix}$$

$$S_{11} = \begin{pmatrix} A_{11} + uu' & A_{12} + uv' \\ A_{21} + vu' & A_{22} + vv' \end{pmatrix}$$

$$S_{12} = \begin{pmatrix} A_{13} + uw' \\ A_{23} + vw' \end{pmatrix}$$

$$S_{22} = A_{33} + ww'$$

Since $RV(X^{(1)}, X^{(2)})$ depends on unknown (x) , we can rely on (u) writing:

$$RV(u) = \frac{2w'A_{31}u + w'wu'u + \alpha}{\sqrt{\gamma}\sqrt{(u'u)^2 + 2u'(A_{11} + Iv'v)u + 4v'A_{21}u + \beta}}$$

$$\alpha = \operatorname{cov}\{A_{13}A_{31} + (A_{23} + uw')(A_{32} + wv')\}$$

$$= \operatorname{cov}(A_{13}A_{31} + A_{32}A_{23} + 2w'A_{32}v + w'wuu')$$

$$\beta = \operatorname{cov}\left(\left(A_{11}^2 + 2A_{12}A_{21} + A_{22}^2\right) + 2v'A_{22}v + (v'v)^2\right)$$

$$\gamma = \operatorname{cov}\left\{(A_{33} + ww')^2\right\} = \operatorname{cov}\left\{\left(A_{33}^2\right) + 2w'A_{33}w + (w'w)^2\right\}$$

Example:

We are going to show the method through an example (5 individuals and 4 variables).

We will use a structure of Y_2 to estimate the missing data in the last individual.

$$X = \begin{pmatrix} 1 & 4 & 6 & 5 & \\ 3 & 4 & 3 & 1 & -1 \\ 1 & 4 & 8 & 5 & 7 \\ 5 & 6 & 5 & 3 & 1 \end{pmatrix}$$

In this case:

$$RV_u = \frac{4.062u^2 + 11.739u + 41.563}{9.076\sqrt{u^4 + 12 - 625u^2 + 3.35u^4 + 28.752}}$$

Where $u = \frac{(x-4)}{\sqrt{5}}$

$u_{\max} = 0.9257$

$x_{\max} = 6.07$

$RV_{\max} = 0.935$

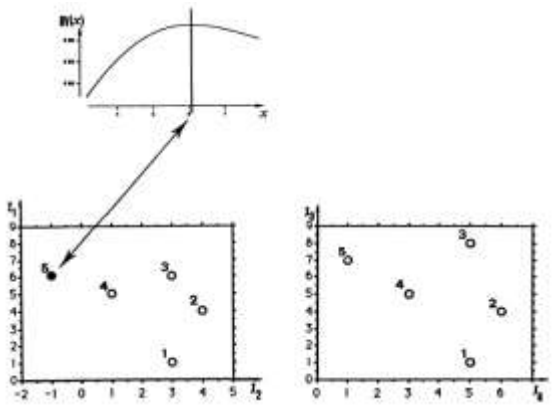

Fig 1 results of the example

VII. NOTES:

1- If individual vectors have missing data, each incomplete vector is treated separately and sequentially with the complete $n - r$ vector.

2- If the separation of the data into two groups of variables is not imposed by the context, a

separation may be chosen so that the missing data appear only in the first group and possibly even so that the entire missing vector represents the first group. The formulas are simplified.

$$S_n = \begin{pmatrix} A_{11} + uu' & A_{13} + uw' \\ A_{31} + wu' & A_{33} + ww' \end{pmatrix}$$

$$RV_{(u)} = \frac{2w'A_{31}u + w'wu'u + \alpha}{\sqrt{\gamma}\sqrt{(u'u)^2 + 2u'(A_{11})u + \beta}}$$

With: $\alpha = \text{cov}(A_{13}A_{31})$, $\beta = \text{cov}(A_{11}^2)$

And: $\gamma = \text{cov}(A_{33}^2) + 2w'A_{33}w + (w'w)^2$

If on other hand, separation is imposed a prior and unfortunately data is missing in both sets of variables, the principle of maximizing RV is still applicable. Only the expression of the RV in relation to missing parts is more complicated.

In the particular case of bivariate data $(p = q = 1 and X_{n-1} = (c_1, c_2))$, $RV(u)$ becomes:

$$RV(u) = \frac{(wu + A_{13})^2}{(w^2 + A_{33})(u^2 + A_{11})}$$

And by the property (a) of σ_v is equal to the square of the simple correlation.

The maximization of this function provides the estimates.

$$\hat{x} = c_1 + \frac{A_{11}}{A_{13}}(z - c_3)$$

Whereas Buck's method estimates by:

$$\hat{x} = c_1 + \frac{A_{13}}{A_{33}}(z - c_3)$$

RV method therefore uses the regression of z on x and Buck's method of x on z.

Properties (b) and (c) make it possible to ensure the invariance of the RV method for orthogonal transformations.

VIII. EXAMPLE AND COMPARISON:

We will deal with the "Heads" data collected by Frets and processed in the book by Mardia, Kent, and Bibby (1979). They form an x data matrix of 25 individuals and 4 variables (2 groups of 2 variables each).

We randomly decree 4 missing values:

$X(23,1)$, $X(24,2)$, $X(25,1)$ and $(25,2)$

For the direct estimation approach of missing data, we compare the RV method with four methods each using as much information as possible. The first method estimates the missing data by the mean (MEAN).The next three are based on regression: multiple regression mostly variables (REGR), simple regression on the most correlated variable (SINGLE), and step-by-step regression (STEP).

For the parameter estimation approach, we will compare the estimate of the correlation matrix after the estimates of missing values of the previous five methods and that of three new methods.

The first is based on the EM algorithm (ML), the second uses all the information available for calculation of the correlation matrix (ALLVALUE) and the last takes into account only the complete data (COMPLETE).

After from the RV method, all result was obtained with BMDP.

We can therefore refer to the BMDP manual for a precise description in parentheses brackets correspond to BMDP terminology. The comparison will be made using two criteria defined by Gleason and Staelinen 1975:

The first Q_α represents a distance between the real value and imputed value, the second D_α represents a distance between the real correlation and estimated correlation.

Direct estimation approach:

$$Q_\alpha = \sqrt{\sum \frac{(X_{ij}^{(\alpha)} - X_{ij})^2}{\sigma_j^2 np\pi}}$$

Parameter estimation approach:

$$D_\alpha = \sqrt{\sum \frac{(R_{ij}^{(\alpha)} - R_{ij})^2}{p(p-1)}}$$

In this formulas, p is the number of variables, n the number of individuals, π the percentage of missing data, $R_{ij}(resp.X_{ij})$ the correlation matrix (resp. Real data), $R_{ij}^{(\alpha)}(resp.X_{ij}^{(\alpha)})$ the correlation matrix (resp. of the data) obtained by method α and σ_j^2 the real variance of variable j.

The table below makes it possible to make the following observations:

1- The RV method gives a good results which can be explained by a fairly high RV value on the real data $(RV = 0.5998)$.

2- The percentage of missing data in our example is low $(\pi = 4\%)$. Despite this, the Mean method is much less efficient than the Buck type methods (REGR, SINGLE, STEP). The importance of variables being an optimality criterion of Buck type methods, their good behavior here is not surprising: indeed we have max $R_{ij} = 0.8392$ and min $R_{ij} = 0.6932$.

3- Surprisingly, specific methods of parameter estimation have a poorer D_α than those that estimate parameters after imputation.

Can this be seen as a condemnation of the first approach??

4- To see the influence of the hypothesis of random distribution of the missing data, we created non-DMCH data (the data is missing if $X_1 > 200$ or $X_2 > 160$).

The Q_α coefficient increases by a factor ranging from 1.2 to 1.8 compared to the DMCH data but the classification of the methods remain identical RV method at the head followed by regression type methods than by method based on the mean.

Thus the RV method showed in this example that it resisted the violation of this hypothesis better than the other methods.

Processing of other real data confirmed these observations and corroborated the good imputation quality of the RV method.

TABLE 1
Comparison Results

Method	D_α	Q_α
RV	0.01034	0.73959
MEAN	0.04280	1.56021
REGR	0.01043	1.11445
SINGLE	0.00896	1.21326
STEP	0.01053	1.12060
ML	0.01172	
ALLVALUE	0.02398	
COMPLETE	0.02208	

IX. CONCLUSION

in this research, we introduced a new method to estimate missing data by taking advantage of the vector correlation coefficient, and we made sure that it's effective by comparing it to previous standard methods and studies, which in turn helps improve the sample estimates for a better representation for the population, which gives us more accurate and closer results to reality.

REFERENCES

1. Allison, P. D. (2006). **Imputation of categorical variables with PROC MI**. Paper presented at the annual meeting of the SAS Users Group International, San Francisco, CA.
2. BaniAwad, Ali Mohamed. (2011). **Comparing Methods of Dealing with Missing Data in Estimating Items and Persons Parameters**. PhD Thesis unpublished, Yarmouk University, Irbid, Jordan.
3. Christyn E.Tannenbaum.(2009).**The empirical nature and statistical treatment of missing data**, Journal of University of Pennsylvania.
4. Der Megreditchian G.(1988**).Problems with missing data in statistica l practice. Working Note of the Meteorological Studies and Research Institution**
5. Finch, H. (2008).**Estimation of item response theory parameters in the presence of missing data**. Journal of Educational.
6. Hussein, Ali Nasser. (2012**). Estimating of lost value of Responding variable in the Multi Regression Model**.Journal of Economic Sciences, Vol. 8, No. 30.
7. Kariya T, Krishnaiah P.R. et Rao C.R (1983) . **Inference on parameters ofmultivariate normal populations when some data is missing**. Developments in Statistics, 4.
8. Khare, B.B, Srivastava, S. (1997). **Transformed ratio type estimators for the population mean in presence of non-response**. *Comm. Statist. Theory Methods*.
9. Little R.J.A. (1988) .**A test of missing completely at random for multivariate data with missing values**. Journal of the American Statistical Association, 83.
10. Little R.J.A. et Rubin D.B. (2000) . **Statistical analysis with missing data,(second edition).**Wiley & Sons,lnc.
11. Murray L.W. (1986) **.Estimation of missing cells in randomized block and latin square designs**. The American Statistician, 40.
12. Pigott, T. D. (2001). **A review of methods for missing data**. Educational Research and Evaluation, 7.
13. Rubin D.B. et Schenker N. (1986) . **Multiple imputation for interval estimation from simple random samples with ignorable nonresponse**. Journalof the American Statistical Association, 81.
14. Schafer, J. L., & Graham, J. W. (2002). **Missing data: Our view of the state of the art.** Psychological Methods, 7.
15. Simon G.A. et Simonoff J.S. (1986) . **Diagnostic plots for missing data in least squares regression**. Journal of the American Statistical Association, 81.
16. Srivastava M.S. (1985). **Multivariate data with missing observations**. Commun. Statis. - Theor. Meth. , 14.
17. Witta, E.L. (2000). **Effectiveness of four methods of handling missing data using samples from a national database**. Teacher Education Yearbook, 28.