A Comparison of The Methods of Estimating Missing Data Between Nearest Neighbor and Regression Imputation

Atsavin Saneechai
Department of Information Technology, Faculty of Science and Technology, Bangkok Suvarnabhumi College,
Bangkok, E-mail: atsavin555@hotmail.com

Abstract— **Missing data are observations that we need to know but cannot acquire. This is a problem that is often found in survey researches. Researchers must find a way to solve this data analysis problem. If the missing data are handled by an inappropriate method, the result of the analysis might be distorted and, consequently, the analysis of the rest of valid data could be deviated. This research studies 2 methods of solving the problem of missing data, which are Nearest Neighbor Imputation and Regression Imputation using the criterion of the root mean square error (RMSE) in the comparison of errors. It is found out from the comparison that one method may be appropriate for particular data or things that need to be considered only. However, the method of estimating missing data by Nearest Neighbor Imputation (NNI) is the appropriate means when considering from the criterion of the root mean square error (RMSE) since it yields minimum values in all levels of missing values. The sample size equals 40.**

Keywords- Missing Data; Nearest Neighbor Imputation; Regression Imputation

I. INTRODUCTION

Missing data are observations that we need to know but cannot acquire. This is a problem that is often found in survey researches. Researchers must find a way to solve this problem in the data analysis. If the missing data are handled by an inappropriate method, the result of the analysis might be distorted and, consequently, the analysis of the rest of valid data could be deviated. In general sample surveys, missing data can occur in 3 manners [3], i.e., a) non-coverage, for example, a question about the number of appointments with doctors in the last 12 months. Those who have not had any appointment with a doctor in the last 12 months may not answer this question, b) unit non-response, which may derive from the lack of understanding of the meanings of words that are used to create questionnaires and c) item non-response. In most cases, the imputation of the missing data will be used to solve the problem of missing data from a refusal to answer some questions or some variables.

There are several methods of handing missing data. The consideration to choose a particular method depends on the manner of the missing data. If an improper method is chosen, the deviation might be increased and it could affect the result of the analysis. The methods of handling missing data from item non-response are divided into 2 main categories [2], i.e.

a) Model-donor imputation: the estimation of data which comes directly from the model, which are mean imputation, regression imputation, ratio imputation, etc.

b) Real-donor imputation: the estimation of data which comes from the data set of observations, which are cold deck imputation, hot deck imputation, nearest neighbor imputation, etc.

This research will study the methods of estimating missing data by Nearest Neighbor Imputation and Regression Imputation.

II. THE OBJECTIVE OF THE RESEARCH

To study the methods of estimating missing data by Nearest Neighbor Imputation and Regression Imputation

III. RELATED THEORIES AND RESEARCHES

In this research, the researcher has studied related theories, which are the Principles of Estimating Missing Data, the Estimation of Missing Data by the Method of Nearest Neighbor Imputation and the Estimation of Missing Data by the Method of Regression Imputation.

A. The Principles of Estimating Missing Data [5]

Consider samples with 2 related variables:
$(x_1, y_1), (x_2, y_2), ..., (x_r, y_r), (x_{r+1}, y_{r+1}), ..., (x_n, y_n)$
The values of y_1 to y_r are y-values that are observable, y_{r+1} to y_n are values of missing data and all x-values are observable. The values of y missing data must be estimated as demonstrated in picture 1[1]

X	Y	
x_1	y_1	
x_2	y_2	
⋮	⋮	
x_r	y_r	
x_{r+1}	y_{r+1}	Missing
⋮	⋮	
x_n	y_n	

Figure 1 demonstrates the pattern of missing data

B. The Estimation of Missing Data by the Method of Nearest Neighbor Imputation

If $$y_j^* = y_i \quad (1)$$

Whereas if $|x_i - x_j| = \min_{1 \le i \le r} |x_i - x_j|$

For $i, 1 \le i \le r$ and $j = r+1, ..., n$

This is the method of consideration in choosing the sampling unit of which characters are most similar to those of the sampling unit with missing value. Afterwards, the missing value is replaced by the value of the similar sampling unit.

C. The Estimation of Missing Data by the Method of Regression Imputation

If $$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j \quad (2)$$

for $i, 1 \le i \le r$ and $j = r+1, ..., n$

Whereas $$\hat{\beta}_1 = \sum_{i=1}^{r} (x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum_{i=1}^{r} (x_i - \bar{x})^2$$

$, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\bar{y} = \frac{1}{r} \sum_{i=1}^{r} y_i$ and $\bar{x} = \frac{1}{r} \sum_{i=1}^{r} x_i$

D. The Estimation of Variance of Regression Imputation Estimator

From the equation $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j$, therefore

$$\begin{aligned} V(\hat{y}_j) &= V(\hat{\beta}_0 + \hat{\beta}_1 x_j) \\ &= V(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_j) \\ &= V(\bar{y} + \hat{\beta}_1 (x_j - \bar{x})) \end{aligned}$$

$$= V(\bar{y}) + V(\hat{\beta}_1 (x_j - \bar{x})) + 2Cov(\bar{y}, \hat{\beta}_1 (x_j - \bar{x}))$$

$$= \frac{\sigma^2}{r} + (x_j - \bar{x})^2 V(\hat{\beta}_1)$$

Whereas $\hat{\beta}_1 = \sum_{i=1}^{r} (x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum_{i=1}^{r} (x_i - \bar{x})^2$

$$V(\hat{y}_j) = \frac{\sigma^2}{r} + (x_j - \bar{x})^2 V\left(\sum_{i=1}^{r} (x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum_{i=1}^{r} (x_i - \bar{x})^2 \right)$$

$$= \frac{\sigma^2}{r} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^{r} (x_i - \bar{x})^2} \sigma^2$$

$$V(\hat{y}_j) = \sigma^2 \left[\frac{1}{r} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^{r} (x_i - \bar{x})^2} \right] \quad (3)$$

E. The Estimation of Variance of Nearest Neighbor Imputation Estimator

y_j^* is the estimated value that is acquired from the method of NNI with the variables of interest y and x which are related to each other under the linear model.

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j \quad (4)$$

at $\varepsilon_j \sim NID(0, \sigma^2)$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$

for $i \ne j = 1, 2, ..., n$

$$V(y_j^*) = \sigma^2 \quad (5)$$

F. The Criterion Used in the Comparison

In survey researches, the target population can specify the members, and the complete data set is necessary. If there is no missing data in the sample, the

sample mean will be $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, which is an unbiased estimator of the population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$. However, $\bar{y}_I = \frac{1}{n}\left(\sum_{i=1}^{r} y_i + \sum_{j=r+1}^{n} \hat{y}_j\right)$ is a biased estimator when $y_{r+1},...,y_n$ are the values that come from the 2 methods of estimation in the comparison of the $\bar{y}$ estimators by using the estimated values without considering relative variance. [4]

The demonstration of the result of the comparison between 2 methods of estimation, which are Nearest Neighbor Imputation (NNI) and Regression Imputation (RI), using the criterion of the root mean square error (RMSE) reveals that the root mean square error of the $\hat{V}(\bar{y}_I)$ variance estimator is:

$$RMSE\left(\hat{V}(\bar{y}_I)\right) = \sqrt{E\left(\hat{V}(\bar{y}_I) - V_M(\bar{y})\right)^2}$$

V_M is the Monte Carlo variance and $\hat{V}(\bar{y}_I)$ is the value of variance estimator of $\bar{y}_I$ and $V_M(\bar{y}_I)$ is the variance value of $\bar{y}_I$ that has been through simulation.

IV. The Methods of Performing the Experiment

In this study, the samples are randomly selected. The sample size is n, which comes from the sampling without replacement. x_i and y_i are random variables that are acquired from simulation. The population size (N) equals 300. The differences of each method of estimating missing data are evaluated by the Monte Carlo simulation. The things that are fixed for this study are:

a) The population is limited to N size by sampling. Size n=40
b) The rates of the anticipated values of missing data that are used in the consideration are 12, 8 and 4.
c) The correlation coefficient between X and Y variables is 0.750

V. The Results of the Experiment

The descriptive statistics of the data as shown in Table 1 and Table 2 evaluates the differences of each method of estimating missing data using Monte Carlo simulation.

TABLE I demonstrates the descriptive statistics of the data of X and Y variables.

	min	max	mean	median	variance	skewness	N
x	1200	8500	6300	5000	9500	2.11	**300**
y	**4000**	**4500**	**3800**	**6000**	**8500**	**2.03**	**300**

*Consider **n** = 40 with the missing y-values of **m** = 12, 8 and 4

TABLE II demonstrates the results of the simulation of the 2 methods of estimating missing data when the population size (N) equals 300 and the number of rounds (M) equals 500.

n	m	Method	RMSE
40	**12**	**NNI**	**4.25**
		RI	**4.38**
	8	**NNI**	**3.06**
		RI	**3.19**
	4	**NNI**	**2.53**
		RI	**2.88**

n = sample size; m = number of missing values; RMSE = root mean square error; N = population size.

VI. The Conclusion of the Experiment

In the handling of missing data that are found in sample surveys when the missing data are caused by item non-response, which is the subject of our interest, the results of the experiment which simulates the problem in order to compare the 2 methods of estimating missing data using the criterion of the root mean square error (RMSE) show that it is not possible for us to decide which method is the best. This is because one method might be appropriate for particular data or things that need to be considered. Nevertheless, the method of estimating missing data by Nearest Neighbor Imputation (NNI) is the appropriate means when considering from the criterion of root mean square error (RMSE) since it yields minimum values in all levels of missing values. The sample size equals 40.

VII. RECOMMENDATIONS

There should be more samples of the experiment for greater diversity and the differences of missing data should be increased. Moreover, in order to obtain a more efficient result of the estimation, there should be many more criteria for the consideration.

ACKNOWLEDGMENT

Special thanks to Bangkok Suvarnabhumi College which helps to support this research and contributes to its success.

REFERENCES

[1] Chaimongkol , W. and Suwattee , P. Weighted Nearest Neighbor andRegression Imputation . 9th Asia-Pacific Decision Sciences Institute Conference. APDSI-KOPOMS Seoul. Korea. July 1-4, 2004. Seoul. Korea.

[2] Royall, R.M., and Eberhardt, K.R.,Variance Estimates for the Ratio Estimator. Sankhy*a* C. vol. 37, 1975, pp. 43-52.

[3] Kalton,G. and Kasprzyk,D. Imputing for Missing SurveyResponses. Proceeding Section of Survey Research Method. American Statistical Association, 1982, pp 22-33.

[4] Laaksonen,S., Regression-Based Nearest Neighbor Hot Decking, Computational Statistics. vol.15, 2000, pp. 65-71.

[5] Little, R.J.A. and Rubin, D.B. *Statistical* Analysis with Missing Data. New York : Wiley, 1987.