Recovery of Missing Data using Iterative Weight of Jackknife and Regression

Boonchom Srisa-ard[a]*, Ed.D. Jumlong Vongprasert[b], Ph.D.
Saner Piromjitpong[c], Ph.D. Nipaporn Shutiman[d], Ph.D.
Bhusana Premanode[e], Ph.D.

[a]Associate Professor Faculty of Education, Mahasarakham University.
[b]Assistant Professor Faculty of Science, Rajabhat Ubon Ratchathani University.
[c]Assistant Professor Faculty of Education, Rajabhat Ubon Ratchathani University.
[d]Assistant Professor Faculty of Science, Mahasarakham University, Thailand
[e]Professor Centre for Bio-Inspired Technology Imperial College.
*Corresponding Author. E-mail address: boonchom.c@msu.ac.th

Abstract

The purpose of this study were first to develop the iterated weighted Jacknife method and regression (IWJR) for missing data estimation, and secondly to compare its efficiency of estimation population mean, population variance, and population correlation under missing complete at random (MCAR) and simple random sampling with another four well-defined methods, namely; Listwise deletion (LD), Mean imputation (MI) and Regression imputation (RI). By using simulation data with 5 percent of the missing data, where i) sample sizes were 100, 200 and 500 units, and ii) low, moderate and high correlations, the tests of mean, variance, and correlation of the IWJR method outperformed the average of three methods; namely, Listwise Deletion (LD), Mean Imputation (MI), and Regression Imputation (RI). Under the constraint of the different sample sizes and degree of correlations, the proposed IWJR method performed remarkably better than the group average of LD, MI, and RI, by 22.65 and 21.34 percent respectively.

Keywords: Missing data, MCAR, Simple random sampling

Introduction

Missing data is a common problem that has been found in quantitative research (Heeringa. 2000) albeit there were controlled rigidly with preventive and corrective mechanism (Huisman. 1998: 271-278). Kim and Curry (2001) proved that

missing of the multivariate random variables by 10 percent provided analytical errors up to 59 percent. Estimation of missing data can vigorously improve the quality of research in education services (Peng et al. 2006: 31-78). For example, in examination papers, the impact of missing markings is crucial, in which it could cause errors in both Type-I and Type-II (Robitszsch, and Rupp. 2009: 18-34).

On the strong points of the missing data methods, Sentas and Angelis (2006) described that in Listwise deletion (LD), cases with missing values for any of the variables are omitted from the analysis. The procedure is quite common in practice because of its simplicity, but when the percentage of missing values is high, it results in a small complete subset of the initial data sets and, therefore, in difficulties in constructing a valid cost model. Moreover, the Mean Imputation (MI) method replaces the missing observations of a certain variable with the mean of the observed values in that variable. It is a simple method that performs well, especially when valid data are normally distributed. In Regression Imputation (RI) method, the missing values were estimated through the application of multiple regression where the variable with missing data was considered as the dependent one and all other variables as predictors (Sentas and Angelis. 2006: 404–414). On the weak points, Little and Rubin (2002), explained that the values of variance from the LD technique is underestimated. Rovine and Delaney (1990), Landerman, and Land & Pieper (1997), and Brockmeier (1998) tested that the variance from the MI technique is undervalued. Apparently, Little and Schenker (1995) showed that the RI method conceived the same undervalue, in which it exemplified to a problem of multicollinearity.

This paper presents a novel approach in recovery of missing data by employing Iterative Weight of Jackknife cross-validation and Regression (IWJR). The objectives of this paper is to compare its efficiency of estimation population mean, population variance and population correlation under both Missing Complete At Random (MCAR) and Simple Random sampling with another three well-defined methods, namely; Listwise Deletion (LD), Mean Imputation (MI), and Regression Imputation (RI) under the simulation data. Organization of this paper is: Section 1 for Introduction, Section 2 explains Theoretical Considerations of Jackknife whereas Step-by-Step of the Experiment is in Section 3.

Theoretical Consideration of Jackknife

Refaeilzadeh, Tang, and Hu (2008) verified that there are four methods of crossvalidation starting from Resubstitution, Hold-Out, K-Fold, and Leave-One-Out or Jackknife. The Jackknife's procedure, however, yielded the most accurate technique but consumed a large extend of computational time. The basis of Jackknife cross-validation is to analyse the statistical data by leaving out one or more observations at a time from the sample set. From this new set of replicates of the analysis, estimations for the bias and variance can be calculated. The following section is to describe theoretical and procedural aspects of the Jackknife technique.

Definitions and notations

The Jackknife is an iterative a process from which each element is dropped from the sample and then the rest of data set goes into statistical analysis. Right after the first analysis/estimation, the sample is added back and another sample is, in turn, being dropped and analysed again and again until it becomes a loop. The Jackknife technique can be used in the multivariate analysis, but is not suitable in time series data since the observations are assumed to be independent and identically distributed (i.i.d.).

Abdi and Williams (2010) defined an estimation of the population without the n-th observation to the n-*th* partial prediction, and it is denoted T_{-n} as follows:

$$T_{-n} = f\left(X_1,\ldots,X_{n-1},X_{n+1},\ldots,X_n\right) \quad (1)$$

A pseudo-value estimation of the nth observation is denoted T_n^*; it is computed as the difference between the parameter estimation obtained from the whole sample and the parameter estimation without the nth observation. Formally:

$$T_n^* = NT - \left(N-1\right)T_{-n} \quad (2)$$

The Jackknife estimation of θ is denoted T^*, and it is used as the mean pseudo-value, which is formulated by:

$$T^* = \overline{T}^* = \frac{1}{N}\sum_{n}^{N} X_i T_n^{8} \quad (3)$$

where T^* is the mean of the pseudo-values. The variance of the pseudo-value is denoted $\hat{\sigma}^2_{T_n^*}$; and it is formulated by:

$$\hat{\sigma}^2_{T_n^*} = \sum_{n}^{N} \frac{\left(T_n^* - \overline{T}^*\right)^2}{N-1} \quad (4)$$

Data structure

We first simulated Grade Point Average (GPA) by Monte Carlo, with the other suit being secondary data collected from 2,381 schools under the Minister of Education, Thailand.

Strategy

The comparison results between the IWJR method and the LD, MI, and RI methods were measured by Mean Square Error (MSE) (Hank, Reitsch, and Wichern. 2001), given by the following equation:

$$MSE = \sum_{i=1}^{n} \frac{\left(\theta_i - \hat{\theta}_i\right)^2}{n} \tag{5}$$

In the experiment where the missing data was 5 percent, we ran four methods in parallel; namely, LD, MI, RI, and IWJR with the sampling data of 100, 200, and 500 (Timm 1970). Moreover, the experimental plan was to measure three different levels of correlation, which are low, medium, and high (Frane 1976, Little and Rubin 2002, and Hegamin-Younger and Forsyth 1988).

Step-by-Step of the Experiment

We first measured means from both data suits, and then performed cross-validation using Jackknife conjectured by Quenouille (1956). The means were later calculated with Yu (2003), and Sahinler and Topuz (2007), which was derived from:

$$\overline{y}_j = \frac{\sum_{i=1}^{r} \overline{y}_i}{r}; j = r+1, r+2, \cdots, n \tag{6}$$

In the next step, we entered the results from Equation (2), into the regression under Draper, Norman, and Smith (1998), which was derived from:

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j \ ; \ r+1 \le j \le n \tag{7}$$

From Equations (6) and (7), the weighted Jackknife and regression was derived from:

$$\hat{y}_j^* = w_j \overline{y}_j + \left(1 - w_j\right) \hat{y}_j \tag{8}$$

where $$w_j = \frac{\hat{\sigma}^2_{Reg.}\left[\frac{1}{r}+\left(x_j-\bar{x}\right)^2 \Big/ \sum_{j=1}^{r}\left(x_j-\bar{x}\right)^2\right]}{\frac{\hat{\sigma}^2_{jackknife}}{r}+\hat{\sigma}^2_{reg}\left[\frac{1}{r}+\left(x_j-\bar{x}\right)^2 \Big/ \sum_{i=1}^{r}\left(x_i-\bar{x}\right)^2\right]}$$

where $\hat{\sigma}^2_{Reg.}$ is variance of regression, and $\hat{\sigma}^2_{jackknife}$ is variance of Jackknife.

We randomly sampled the missing data one by one, and then replaced them with the weighted Jackknife and regression, $\hat{y}^*_j$, which was derived from Equations (6) to (8). The next step was to measure the performance of the IWJR method by comparing with Listwise Deletion (LD), Mean Imputation (MI), and Regression Imputation (RI).

In lieu of accuracy testing, we examined effectiveness of mean, variance and correlation of the population.

Effectiveness of Mean, Variance, and Correlation of the Population

Figure 1 demonstrates effectiveness of Mean, Variance, and Correlation of the Population. It begins in an order of the following items: i) calculating mean, variance, and coefficient correlation, ii) simple random sampling with 100, 200, and 500 units of the population, iii) generating missing data using MCAR at 5 percent of the population, iv) estimating the missing data with four methods; namely, LD, MI, RI, and IWJR, v) calculating mean, variance and coefficient correlation from Item iv), and vi) calculating differences of mean, variance, and coefficient correlation of each missing data estimation of the population, vii) repeating item ii) to vi) until reaching 1,000 times, and finally viii) calculating Mean Square Error (MSE) of mean, variance, and coefficient correlation of each missing data estimation.

Results and Discussion

We demonstrated the effectiveness of the proposed IWJR method using parameters of mean, variance, and correlation estimations of the population. It concluded in Table 1 and Table 2 that with 5 percent of the missing data, where i) sample sizes were 100, 200 and 500 units, and ii) low, moderate and high correlations, the tests of mean, variance, and correlation of the IWJR method outperformed the average of three methods; namely, Listwise Deletion (LD), Mean Imputation (MI), and Regression Imputation (RI). Under the constraint of the different sample sizes and degree of correlations, the proposed IWJR method performed remarkably better than the group average of LD, MI, and RI, by 22.65 and 21.34 percent respectively.

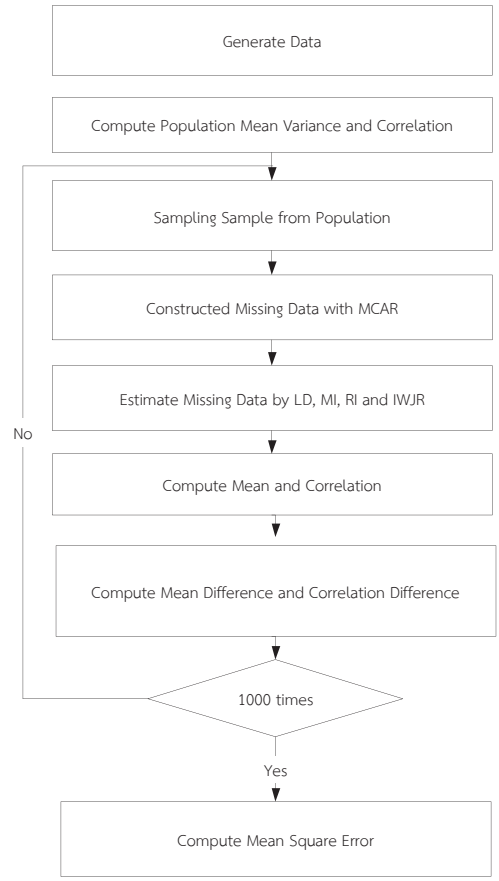

Figure 1 Step of the Experiment

Table 1 Means Square Error of LD MI RI and IWJR classified by sample sizes.

	Sample Size	Method					Percentage
		LD	MI	RI	Average	IWJR	
Mean	100	0.00061244	0.00061244	0.00060509	0.00060999	0.00049132	19.45%
	200	0.00030229	0.00030229	0.00030009	0.00030156	0.00024786	17.81%
	500	0.00011246	0.00011246	0.00011025	0.00011172	0.00008894	20.39%
Variance	100	0.00007535	0.00007548	0.00007466	0.00007516	0.00006160	18.05%
	200	0.00003550	0.00004014	0.00003775	0.00003780	0.00002889	23.56%
	500	0.00001474	0.00002242	0.00001893	0.00001870	0.00001169	37.48%
Correlation	100	0.00572699	0.00577014	0.00588620	0.00579444	0.00460155	20.59%
	200	0.00276313	0.00284179	0.00290709	0.00283734	0.00220152	22.41%
	500	0.00111020	0.00127885	0.00119057	0.00119321	0.00090579	24.09%
Average Percentage							22.65%

Table 2 Means Square Error of LD MI RI and IWJR classified by correlation between variables.

	Correlation	Method					Percentage
		LD	MI	RI	Average	IWJR	
Mean	Low	0.00034485	0.00034485	0.00034432	0.00034467	0.00027851	19.20%
	Moderate	0.00034628	0.00034628	0.00034238	0.00034498	0.00027927	19.05%
	High	0.00033606	0.00033606	0.00032872	0.00033361	0.00027034	18.97%
Variance	Low	0.00004236	0.00004724	0.00004626	0.00004529	0.00003440	24.03%
	Moderate	0.00003880	0.00004293	0.00004101	0.00004091	0.00003175	22.39%
	High	0.00004444	0.00004787	0.00004407	0.00004546	0.00003602	20.75%
Correlation	Low	0.00470831	0.00454879	0.00494511	0.00473407	0.00374121	20.97%
	Moderate	0.00343693	0.00349947	0.00355895	0.00349845	0.00280132	19.93%
	High	0.00145507	0.00184251	0.00147980	0.00159246	0.00116633	26.76%
Average Percentage							21.34%

References

Adbi, H., & Williams, L. J., (2010). Jackknife, *Encyclopedia of Research Design*, Thousand Oaks, CA: Sage.

Brockmeier, L.L., Kromrey, J.D., & Hines, C. V., (1998). Systematically missing data and multiple regression analysis: An empirical comparison of deletion and imputation techniques. *Multiple Linear Regression Viewpoints*. 25, 20-39.

Draper, Norman R., & Smith, H., (1998). *Applied Regression Analysis.* 3rd ed. John Willey & Sons, Inc., New York.

Frane, J.W., (1976). Some simple procedures for handling missing data in multivariate analysis. *Psychometrika*. 41, 409–415.

Hank J,E., Reitsch, A.G., & Wichern, W., (2001). *Business Forecasting*. 7th ed. Prentice Hall. New Jersey.

Hegamin-Younger, C., & Forsyth, R., (1998). A comparison of four imputation procedures in a two-variable prediction system. *Educational and Psychological Measurement*. 58(2), 197-210.

Heeringa, Steven George, (2010). *Multivariate Imputation of Coarsened Survey on Household Wealth*, Doctor's Dissertation, University of Michigan. Dissertation Abstract International. Retrieved May, 16 2010, from http://proquest.umi.com/pqdweb?did=731895771andsid =1andFmt= 2andclientId =73599andRQT =309andVName=PQD

Huisman, M., 1998. *Item Nonrespons : Occurrence cause, and Imputation of Missing Answers to Test Item*. DSWO Press, Lieden University, The Netherlands.

Kim, J. & Curry, J., (1977). The Treatment of Missing Data in Multivariate Analysis, *Sociological Methods & Analysis*. 6: 215-240.

Landerman, L.R., Land, K.C., & Pieper C.F., (1997). An empirical evaluation of the predictive mean matching method for imputing missing values. *Sociological Methods and Research*. 26(1), 3–33.

Little, R. J. A., & Rubin, D. B., (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York. John Wiley and Sons.

Little, R.J.A., & Schenker, N., (1995). *Missing data*. In G. Arminger, C. C. Clogg, & M.E. Sobel (Eds.), Handbook of statistical modeling for the social and behavioral sciences. New York.

Peng, C.-Y. J., et al., (2006). *Advances in missing data methods and implications for educational research in Sawilowsky*, S. (eds). Real data analysis. Greenwich, CT., Information Age Publishing Inc. 31-78.

Quenonuille, M. H. (1956). Notes on Bias in Estimation. *Biometrika*. 43, 353-360.

Refaeilzadeh, & Tang, Hu., (2008). Cross Validation. In *Encyclopedia of Database Systems*, Editors: M. Tamer Özsu and Ling Liu. Springer, 2009.

Robitszsch, A., & Rupp, A.A., (2009). Impact of Missing Data on the Detection of Differential Item Functioning: The Case of Mantel-Haenszel and Logistic Regression Analysis. *Educational and Psychological Measurement*. 69(1), 18-34.

Rovine, M.J., & Delaney, M., (1990). Missing data estimation in developmental research. In A. Von Eye (Ed.), *Statistical methods in longitudinal research*, Stanford: Academic Press. 1, 35–79.

Sahinler S., & Topuz, D., (2007). Bootstrap and Jackknife Resampling Algorithms for Estimation of Regression Parameters. *Journal of Applied Qualitative Methods*. 2(2) Summer 2007, 188-199.

Sentas, P. and Angelis, L., (2006). Categorical Missing Data Imputation for Software Cost Estimation by Multinomial Logistic Regression, *Journal of System and Software* (Elsevier), 79, 404-414.

Timm, N.H., (1970). The estimation of variance-covariance and correlation matrices from incomplete data. *Psychometrika*. 35(4), 417-437.

Yu, C. H., (2003). Resampling Methods: Concepts, Applications, and Justification, Practical Assessment, *Research and Evaluation* 8(19). Retrieved October, 13 2009, from http://pareonline.net/getvn.asp? v=8andn=19