Jacknife and Regression Approaches to Missing Data Imputation.

Jumlong Vongprasert
Applied Statistics Department, Ubon Ratchathani Rajabhat University, Ubon Ratchathani, 34000, THAILAND, jumlong.v@ubru.ac.th

Abstract

Missing data imputation is an important task in cases where it is crucial to use all available data and does not discard records with missing values. The purposes of this study was first to develop the Average of Jackknife Mean and Regression (AJRI) for missing data estimation and second is to compare its efficiency of estimation with other methods, namely; Mean Imputation (MI) Regression Imputation (RI) Regime Switching Regression Imputation (RSRI) EM Algorithm (EM) and Multiple Imputation (MUL). By using simulation data, the comparisons were made with the following conditions: (i) Four sample sizes (100, 200 500 and 1,000) (ii) three types of missing data MCAR, MAR, and NMAR. The best imputation under MSE., The best imputation under MSE of mean variance, and correlation classified by sample sizes and by percentage of missing data obtained using RSRI, EM, and RI respectively. classified by missing data types which were obtained using RSRI for mean and EM for variance and correlation. The best imputation under MSE of regression coefficient was obtained using EM and RSRI for model 1 and model 2 respectively. The best imputation under MSE of R^2 was obtained using EM.

Keywords: Missing Data, Imputation, Jackknife.

1. Introduction

Missing data issues exist when a sampled unit does not respond in survey research and when individuals drop out or withdraw before intended completion in a longitudinal study. The existence of missing values may have significant influence on the analysis of the data and therefore on the conclusion of the data analysis. When missing data are present, we may power variability and bias issues. Researchers frequently deal with missing data in survey sampling, longitudinal data analysis and multivariate analysis.

The most serious concern is that missing data can introduce bias into estimates derived from a statistical model [1] [2] [3]. If the responses are not ignorable, however, estimation of the propensity scores is complicated and often requires additional surrogate [4] or instrumental variables [5] to estimate the model parameters consistently.

Furthermore, missing data result in a loss of information and statistical power [6][7]. The elimination of subjects with missing information on one or more variables from the statistical analysis in listwise deletion decreases the error degrees of freedom (df) in statistical tests such as the t. This decrease in turn leads to reduced statistical power and larger standard errors compared to those obtained from complete random samples. [8]

Missing data analysis is importance since an inference based ignoring the missingess may not only misleading conclusions, but also lose efficiency and lead to biased results. [9]

Let $\mathbf{y} = (y_1, y_2, \cdots y_n)'$ denote the complete set of the outcome variables, and $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots \delta_n)'$ be the vector of missing data indicators such that $\delta_i = 1$ when y_i is observed and $\delta_i = 0$ when y_i is missing. We note that each y_i and the corresponding δ_i can also be vectors. Let $\mathbf{y}_{\text{obs}}$ denote the observed and $\mathbf{y}_{\text{mis}}$ missing components of $\mathbf{y}$. With the above notation, the missing data mechanisms are characterized by the conditional distribution of $\boldsymbol{\delta}$ given $\mathbf{y}$, say $f(\boldsymbol{\delta}|\mathbf{y},\phi)$, where ϕ denotes some unknown parameters.

Three major types of missing data are [9]:

Missing completely at random (MCAR) denotes the mechanism that missingness does not depend on the values of the data $\mathbf{y}$, missing or observed.

$$f(\boldsymbol{\delta}|\mathbf{y},\phi) = f(\boldsymbol{\delta}|\phi)\,\forall \mathbf{y},\phi \quad (1)$$

Missing at random (MAR) denotes the mechanism that missingness only depends on the components y_{obs} of $\mathbf{y}$ that are observed, and not on the components that are missing.

$$f(\boldsymbol{\delta}|\mathbf{y},\phi) = f(\boldsymbol{\delta}|\mathbf{y}_{\text{obs}},\phi)\,\forall \mathbf{y}_{\text{mis}},\phi \quad (2)$$

Not missing at random (NMAR) denotes the one that the distribution of $\boldsymbol{\delta}$ does depend on the missing values in the data $\mathbf{y}$.

$$f(y_i|x_i,\delta_i = 0,\phi) \neq f(y_i|x_i,\delta_i = 1,\phi) \quad (3)$$

2. Research Methodology

Data Set

In this section, we introduced and described the data set. We generated random variables (X,Y) from, Model 1 : $y_i = 0.8 + x_i + e_i$ and, Model 2 : $y_i = 0.4x_i + 0.6x_i^2 + e_i$ where $x_i \sim Uniform(-2,2)$ and $e_i \sim N(0,0.5)$ used by Im. [10], with sizes of 100, 200, 500 and 1,000 units. The sample data sets are denoted as $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$. We used three types of missing data MCAR, MAR and NMAR to generate the missing data with the percentages of the samples at 5, 10, 15 and 20, and termed, complete data set, $\{(x_1,y_1),(x_2,y_2),\ldots,(x_r,y_r)\}$ whereas the missing data set are denoted as $\{(x_{r+1},y_{r+1}),(x_{r+2},y_{r+2}),\ldots,(x_{n-r},y_{n-r})\}$

Methods

In this section, the methods applied to impute the original incomplete data set and describe the imputation method used based on AJRI are introduced and described. The subsequent subsections were organized as follows. First, several general considerations were made to explain how the imputation methods had been implemented. Then, the five imputation techniques applied were described: MI RI EM MUL and RSRI. Finally, the JRI method to impute missing value was described together with statistical methods commonly used in methods accuracy evaluation.

Average of Jackknife Mean and Regression (AJRI)

We first measured means from both data suits, and then performed cross validation using Jackknife conjectured by Quenouille [11]. The means were later calculated with Yu [12], and Sahinler and Topuz [13], which was derived from:

$$\bar{y}_j = \frac{\sum_{i=1}^{r} \bar{y}_i}{r};\ j = 1,2,\ldots r \quad (4)$$

In the next step, we used regression under Draper, Norman, and Smith [14], which was derived from:

$$\hat{y}_j^{*} = \hat{\beta}_0 + \hat{\beta}_1 x_j;\ j = 1,2,\ldots,r \quad (5)$$

where

$$\hat{\beta}_0 = \bar{y}_r - \hat{\beta}_1\bar{x}_r,\ \hat{\beta}_1 = \frac{\sum_{i=1}^{r}(x_i - \bar{x}_r)(y_i - \bar{y}_r)}{\sum_{i=1}^{r}(x_i - \bar{x}_r)^2},\ \bar{y}_r = \frac{\sum_{i=1}^{r} y_i}{r}; j = 1,2,\ldots r,\ \bar{x}_r = \frac{\sum_{i=1}^{r} x_i}{r}; j = 1,2,\ldots r$$

From Equations (4) and (5), the AJIR was derived from:

$$\hat{y}_j^* = \frac{\bar{y}_j + \hat{y}_j}{2} \quad (6)$$

and then replaced missing data with $\hat{y}_j^*$.

Mean Imputation (MI) [15]

In the general approach to mean imputation, the mean value of each non-missing variable is used to fill in missing values for all observations by equation 7.

$$\hat{y}_j^* = \frac{\sum_{i=1}^{r} \bar{y}_i}{r}; j = 1,2,\ldots r \quad (7)$$

Regression Imputation (RI)

The completed data set was used to construct regression equation for impute missing data by equation 8:

$$\hat{y}_j^* = \hat{\beta}_0 + \hat{\beta}_1 x_j; j = 1,2,\ldots,r \quad (8)$$

where

$$\hat{\beta}_0 = \bar{y}_r - \hat{\beta}_1\bar{x}_r,\ \hat{\beta}_1 = \frac{\sum_{i=1}^{r}(x_i - \bar{x}_r)(y_i - \bar{y}_r)}{\sum_{i=1}^{r}(x_i - \bar{x}_r)^2},\ \bar{y}_r = \frac{\sum_{i=1}^{r} y_i}{r}; j = 1,2,\ldots r,\ \bar{x}_r = \frac{\sum_{i=1}^{r} x_i}{r}; j = 1,2,\ldots r$$

Regime Switching Regression Imputation (RSRI)

RSRI proposed by Jumlong and Bhusana [6]: The completed data set in each group $\{(x_{1_{st}}, y_{1_{st}}), (x_{2_{st}}, y_{2_{st}}), \ldots, (x_{r_{st}}, y_{r_{st}})\}$ was used to construct regression equation for impute missing data in each group by equation 9:

$$\hat{y}_{j_{st}}^* = \hat{\beta}_{0_{st}} + \hat{\beta}_{1_{st}} x_{j_{st}}; j_{st} = 1,2,\ldots,r_{st} \quad (9)$$

EM Algorithm (EM) [17]

Suppose there is a likelihood or posterior distribution which is difficult to deal with. We augmented the data with some latent data $\mathbf{Z}$, to make the likelihood or posterior tractable and proceed as usual. When the goal is to maximize the posterior distribution, the EM algorithm is often a reasonable method. The EM algorithm is comprised of two steps: an Expectation step (E-step) and a Maximization step (M-step).

Let μ_i denote the current estimation for the mode of the observed posterior $p(\theta|\mathbf{Y})$, $p(\theta|\mathbf{Y},\mathbf{Z})$ be the augmented posterior and $p(Z|\theta_i,\mathbf{Y})$ be the predictive distribution of $\mathbf{Z}$ conditioned on μ_i and $\mathbf{Y}$.

E-step: Compute $Q(\theta,\theta_i) = \int_Z \ln\left[p(\theta|\mathbf{Z},\mathbf{Y})\right] p(\mathbf{Z}|\theta_i,\mathbf{Y}) d\mathbf{Z}$. This is the expectation, $E\left[\ln p(\theta|\mathbf{Z},\mathbf{Y})\right]$ with respect to $p(\mathbf{Z}|\theta_i,\mathbf{Y})$.

M-step: Maximize Q with respect to θ to obtain θ_{t+1}.

These two steps are iterated until convergence is achieved. As with any maximization problem, if there are many critical points for the posterior distribution, the algorithm may not converge to a global maximum. In this work, we used Amelia II in R for EM imputation method.

Multiple Imputation (MUL) [18]

Suppose there is data Y_m with m missing values and we create K new data sets $Y_{c1}, Y_{c2}, \cdots Y_{cK}$ each with values imputed for the missing values where Y_c denotes the corresponding imputed values. Using each of the K data sets we perform the desired analysis and store the result. Thus K separate analyses are generated.

With m imputations, you can compute m different sets of the point and variance estimates for a parameter Q. Let $\hat{Q}_i$ and $\hat{U}_i$ be the point and variance estimates from the ith imputed data set, $i = 1, 2, \ldots, m$ i-1, 2, ..., m. Then the point estimate for Q from multiple imputations is the average of the m complete data estimates: $\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$

In this work, we used Amelia II in R for MUL imputation method.

Model Evaluation

The accuracy of missing data imputation methods is evaluated by Mean Square Error (MSE). To evaluate more precisely the difference in prognosis accuracy among the missing data imputation methods, mean square error of mean variance correlation regression coefficient and R^2 were evaluated in this study.

Structural Flow of the Work

Referring to Section 2, the random variables were generated with sampled the data sets at 100, 200 500 and 1,000 with two models, following extraction of the missing data from the incomplete data sets by the MCAR MAR and NMAR method with the percentages of the samples at 5, 10, 15 and 20. As a result, each data set was complete and ready for simulation using the missing data imputation techniques, namely MI, RI, RSRI, EM, MUL and IJR. For executing the tests, we wrote the codes in R-programming and retrieved some equations relating to those techniques from CRAN projects, and used 10,000 replicated for each condition. Next, the results from simulations with the estimators, namely MSE were tested. The simulations and results are described in the next section.

3. Research Results and Discussion

Missing data imputation methods: MI RI RSRI EM MUL and AJRI were applied to impute missing data. The goal was to analyse the improvements in accuracy when different algorithms were applied to impute missing data values. Table 1-3 indicates the average of MSE of mean variance and correlation classified by sample sizes percentage of missing data and missing type respectively. Table 4-5 indicates the average of MSE of regression coefficient model 1 and model 2 respectively. Table 6 indicates the average of MSE of R^2.

Table 1. Average MSE of mean variance correlation classified by sample sizes

	Sample Size	Method					
		MI	RI	RSRI	EM	MUL	AJRI
Mean	**100**	0.03620359	0.00730347	**0.00729249**	0.00798543	0.01119550	0.00730894
	300	0.03582724	0.00706599	**0.00704602**	0.00729022	0.00915398	0.00706660
	500	0.03572795	0.00701823	**0.00700638**	0.00714394	0.00859795	0.00701846
	1000	0.03569955	0.00698824	**0.00698267**	0.00705796	0.00807938	0.00698830
		0.03586458	0.00709398	**0.00708189**	0.00736939	0.00925670	0.00709557
Variance	**100**	0.17491619	0.06703392	0.06533937	**0.05370747**	0.07584659	0.06708664
	300	0.17170454	0.06596075	0.06537240	**0.04945047**	0.06264433	0.06596650
	500	0.17091817	0.06572458	0.06536907	**0.04858726**	0.05887354	0.06572664
	1000	0.17065553	0.06564991	0.06546722	**0.04810538**	0.05535780	0.06565042
		0.17204861	0.06609229	0.06538702	**0.04996264**	0.06318057	0.06610755
Correlation	**100**	0.01818101	**0.00093019**	0.00099168	0.00154019	0.00247434	0.00093042
	300	0.01749898	**0.00056733**	0.00056985	0.00086625	0.00130424	0.00056735
	500	0.01731664	**0.00049853**	0.00049897	0.00074160	0.00105928	**0.00049853**
	1000	0.01726016	**0.00044987**	0.00045006	0.00065210	0.00087011	**0.00044987**
		0.01756420	**0.00061148**	0.00062764	0.00095004	0.00142699	0.00061155
Overall		0.07515913	0.02459925	0.02436552	**0.01942736**	0.02462142	0.02460489

Table 2. Average MSE of mean variance and correlation classified by percentage of missing data

	Percentage of Missing	Method					
		MI	RI	RSRI	EM	MUL	AJRI
Mean	**5**	0.00659391	**0.00110414**	0.00110592	0.00120992	0.00162872	0.00110441
	10	0.02188109	0.00379691	**0.00379457**	0.00400001	0.00519375	0.00379782
	15	0.04371598	0.00832763	**0.00831367**	0.00865886	0.01092877	0.00832954
	20	0.07126735	0.01514724	**0.01511341**	0.01560876	0.01927557	0.01515052
		0.03586458	0.00709398	**0.00708189**	0.00736939	0.00925670	0.00709557
Variance	**5**	0.04648586	0.01746464	0.01738905	**0.01467748**	0.01954825	0.01746837
	10	0.12438570	0.04611757	0.04575851	**0.03612306**	0.04670337	0.04612800
	15	0.21247075	0.08079100	0.07994919	**0.06093261**	0.07698428	0.08080964
	20	0.30485211	0.11999595	0.11845130	**0.08811743**	0.10948637	0.12002420
		0.17204861	0.06609229	0.06538702	**0.04996264**	0.06318057	0.06610755
Correlation	**5**	0.00388532	0.00018804	0.00019108	0.00026864	0.00039231	**0.00018807**
	10	0.01230981	**0.00047311**	0.00048080	0.00068948	0.00102729	0.00047319
	15	0.02208557	**0.00075036**	0.00076548	0.00115941	0.00174391	0.00075045
	20	0.03197609	**0.00103440**	0.00107320	0.00168262	0.00254447	0.00103448
		0.01756420	**0.00061148**	0.00062764	0.00095004	0.00142699	0.00061155
Overall		0.07515913	0.02459925	0.02436552	**0.01942736**	0.02462142	0.02460489

Table 3. Average MSE of mean variance and correlation classified by missing type.

	Missing type	Method					
		MI	RI	RSRI	EM	MUL	AJRI
Mean	MCAR	0.00091610	0.00030373	0.00031204	0.00060352	0.00056293	**0.00030372**
	MAR	0.00693953	**0.00091550**	0.00092847	0.00114102	0.00147863	0.00091580
	NMAR	0.09973812	0.02006272	**0.02000516**	0.02036362	0.02572855	0.02006720
		0.03586458	0.00709398	**0.00708189**	0.00736939	0.00925670	0.00709557
Variance	MCAR	0.04795130	0.00616646	0.00608642	**0.00314171**	0.00382489	0.00617159
	MAR	0.03156517	0.00370328	0.00365008	**0.00262293**	0.00311840	0.00370609
	NMAR	0.43662935	0.18840712	0.18642454	**0.14412329**	0.18259841	0.18844497
		0.17204861	0.06609229	0.06538702	**0.04996264**	0.06318057	0.06610755
Correlation	MCAR	0.00264834	0.00033146	0.00033117	0.00022833	**0.00021994**	0.00033136
	MAR	0.00227097	0.00023256	0.00023518	0.00020123	**0.00019887**	0.00023249
	NMAR	0.04777328	**0.00127041**	0.00131657	0.00242055	0.00386216	0.00127079
		0.01756420	**0.00061148**	0.00062764	0.00095004	0.00142699	0.00061155
Overall		0.07515913	0.02459925	0.02436552	**0.01942736**	0.02462142	0.02460489

Table 4. Average MSE of Regression Coefficient Model 1

		Method					
Alpha		MI	RI	RSRI	EM	MUL	AJRI
Sample Size	100	0.01685491	0.00237062	**0.00236524**	0.00236677	0.00391711	0.00237404
	300	0.01687338	0.00241399	0.00241205	**0.00240464**	0.00327980	0.00241436
	500	0.01683147	0.00242103	**0.00242011**	0.00242336	0.00308151	0.00242116
	1000	0.01682872	0.00242657	**0.00242612**	0.00242736	0.00288951	0.00242660
		0.01684711	0.00240800	0.00240582	**0.00240547**	0.00328105	0.00240900
Percentage of Missing	5	0.00333581	0.00049659	**0.00049653**	0.00049849	0.00074770	0.00049678
	10	0.01162976	0.00164206	0.00164115	**0.00163653**	0.00229049	0.00164272
	15	0.00880073	0.00110041	**0.00109932**	0.00110270	0.00163338	0.00110120
	200	0.01444413	0.00185228	0.00184886	**0.00184788**	0.00266314	0.00185363
		0.00900571	0.00120786	**0.00120667**	0.00120712	0.00174774	0.00120855
Missing data type	MCAR	0.00000006	**0.00000002**	**0.00000002**	0.00000003	**0.00000002**	**0.00000002**
	MAR	0.00539306	**0.00036080**	0.00036114	0.00036256	0.00067017	0.00036109
	NMAR	0.07425702	0.01249309	0.01247841	**0.01245980**	0.01606144	0.01249743
		0.01331525	0.00190374	0.00190222	**0.00190218**	0.00259327	0.00190452
Overall		0.01684082	0.00250704	0.00250445	**0.00250274**	0.00338493	0.00250809
Beta		MI	RI	RSRI	EM	MUL	JRI
Sample Size	100	0.03500239	0.00174884	**0.00173951**	0.00173989	0.00457773	0.00175271
	300	0.03492384	0.00178997	0.00178127	**0.00178650**	0.00330514	0.00179048
	500	0.03489642	0.00179812	**0.00179001**	0.00179682	0.00293558	0.00179819
	1000	0.03490093	0.00180642	**0.00179849**	0.00180694	0.00258742	0.00180646
		0.03493088	0.00178577	0.00177725	**0.00178245**	0.00331191	0.00178690

Table 4. Average MSE of Regression Coefficient Model 1 (Continued)

Alpha		MI	RI	RSRI	EM	MUL	AJRI
					Method		
Percentage of Missing	5	0.00749790	0.00047277	**0.00047205**	0.00047615	0.00099601	0.00047309
	10	0.02510753	0.00134852	0.00134482	**0.00134160**	0.00256997	0.00134897
	15	0.04943260	0.00243857	**0.00242737**	0.00243054	0.00444835	0.00244087
	20	0.04943260	0.00243857	0.00242737	**0.00243054**	0.00444835	0.00244087
		0.02973149	0.00154507	**0.00153929**	0.00154154	0.00290646	0.00154624
Missing type	MCAR	0.01575492	**0.00000001**	**0.00000001**	0.00000003	**0.00019444**	**0.00000001**
	MAR	0.00977164	**0.00000146**	0.00000095	0.00000146	0.00018030	0.00000150
	NMAR	0.11311379	0.01574163	0.01572416	**0.01569584**	0.02110509	0.01574800
		0.03493088	0.00178577	0.00177725	**0.00178245**	0.00331191	0.00178690
Overall		0.03725769	0.00268954	**0.00268236**	0.00268239	0.00430440	0.00269101
Overall		0.02704925	0.00259829	0.00259341	**0.00259256**	0.00384467	0.00259955

Table 5. Average MSE of Regression Coefficient Model 2

Alpha		MI	RI	RSRI	EM	MUL	AJRI
					Method		
Sample Size	100	0.00773886	0.00040611	**0.00040089**	0.00040399	0.00092310	0.00040618
	300	0.00777817	0.00042442	0.00042343	**0.00042327**	0.00070244	0.00042443
	500	0.00777671	0.00043318	0.00043371	**0.00043111**	0.00064269	0.00043317
	1000	0.00780810	0.00043483	0.00043647	0.00043613	0.00058085	**0.00043483**
		0.00777546	0.00042463	**0.00042362**	0.00042363	0.00071227	0.00042465
Percentage of Missing	5	0.00285456	0.00022040	**0.00021994**	0.00022012	0.00036048	0.00022045
	10	0.00695804	0.00045512	**0.00045317**	0.00045686	0.00072497	0.00045523
	15	0.01043511	0.00055940	**0.00055801**	0.00055892	0.00092063	0.00055953
	200	0.01293189	0.00051022	0.00050971	**0.00050482**	0.00090510	0.00050998
		0.00829490	0.00043629	0.00043521	**0.00043518**	0.00072779	0.00043630
Missing type	MCAR	0.01005656	**0.00000002**	**0.00000002**	0.00000003	0.00020332	**0.00000002**
	MAR	0.00096961	0.00096361	0.00097415	0.00097448	**0.00049399**	0.00096325
	NMAR	0.01771942	0.00093921	**0.00092359**	0.00092141	0.00187218	0.00093961
		0.00958186	0.00063428	**0.00063259**	0.00063197	0.00085650	0.00063429
		0.00845700	0.00048605	**0.00048483**	0.00048465	0.00075725	0.00048606
Beta 1		**MI**	**RI**	**RSRI**	**EM**	**MUL**	**JRI**
Sample Size	100	0.01628848	0.00115737	**0.00112439**	0.00112565	0.00286860	0.00115748
	300	0.01650100	0.00122696	**0.00121678**	0.00122123	0.00208101	0.00122706
	500	0.01647726	0.00123860	0.00123338	**0.00123300**	0.00184800	0.00123863
	1000	0.01655097	0.00125052	0.00124892	**0.00124802**	0.00163153	0.00125052
		0.01645443	0.00121836	**0.00120587**	0.00120697	0.00210728	0.00121842

Table 5. Average MSE of Regression Coefficient Model 2 (Continued)

		Method					
Alpha		MI	RI	RSRI	EM	MUL	AJRI
Percentage of Missing	5	0.00519422	0.00039012	**0.00038834**	0.00038839	0.00072523	0.00039020
	10	0.01400157	0.00099436	0.00098621	**0.00098270**	0.00171931	0.00099443
	15	0.02253428	0.00164066	**0.00162008**	0.00162721	0.00275046	0.00164071
	200	0.02978505	0.00228624	**0.00225750**	0.00226041	0.00380918	0.00228620
		0.01787878	0.00132785	**0.00131303**	0.00131468	0.00225104	0.00132788
Missing type	MCAR	0.00252327	**0.00000001**	0.00000002	0.00000002	0.00004728	**0.00000001**
	MAR	0.00328128	0.00000856	0.00000881	**0.00000845**	0.00003282	0.00000853
	NMAR	0.07689971	0.01033401	**0.01021223**	0.01023371	0.01545389	0.01033629
		0.02756809	0.00344753	**0.00340702**	0.00341406	0.00517800	0.00344828
		0.02000337	0.00186613	**0.00184515**	0.00184807	0.00299703	0.00186637
Beta 2		MI	RI	RSRI	EM	MUL	JRI
Sample Size	100	0.02348000	0.00166504	**0.00163124**	0.00163546	0.01962197	0.00166561
	300	0.02350221	0.00172730	**0.00171550**	0.00171591	0.02157921	0.00172732
	500	0.02346001	0.00174479	0.00173886	**0.00173582**	0.02220101	0.00174481
	1000	0.02352180	0.00175715	0.00175532	**0.00175445**	0.02298758	0.00175715
		0.02349100	0.00172357	**0.00171023**	0.00171041	0.02159744	0.00172372
Percentage of Missing	5	0.00661682	0.00043332	**0.00043159**	0.00043198	0.02898329	0.00043341
	10	0.01870986	0.00124723	0.00123792	**0.00123619**	0.02323550	0.00124747
	15	0.03237578	0.00236793	**0.00234683**	0.00234938	0.01859162	0.00236840
	20	0.04622947	0.00375289	0.00372107	**0.00372096**	0.01652917	0.00375254
		0.02598298	0.00195034	**0.00193435**	0.00193463	0.02183490	0.00195046
Missing type	MCAR	0.00566473	**0.00000001**	**0.00000001**	**0.00000001**	0.03588752	**0.00000001**
	MAR	0.00533408	0.00000628	0.00000661	**0.00000642**	0.03665120	**0.00000624**
	NMAR	0.09703473	0.01486882	0.01473157	0.01474130	**0.00357726**	0.01487180
		0.03601118	0.00495837	**0.00491273**	0.00491591	0.02537199	0.00495935
Overall		0.02781177	0.00268825	**0.00266514**	0.00266617	0.02271321	0.00268862

Table 6. Average MSE of R^2

		Method					
		MI	RI	RSRI	EM	MUL	AJRI
Sample Size	100	0.02282108	0.00071763	0.00072148	**0.00040967**	0.00112891	0.00071763
	300	0.02315524	0.00073410	0.00073312	**0.00041815**	0.00077343	0.00073410
	500	0.02483480	0.00073810	0.00073898	**0.00045581**	0.00072927	0.00073811
	1000	0.02326706	0.00074301	0.00074373	**0.00042628**	0.00058683	0.00074301
		0.02351954	0.00073321	0.00073433	**0.00042748**	0.00080461	0.00073321

Table 6. Average MSE of R^2 (Continued)

		Method					
		MI	RI	RSRI	EM	MUL	AJRI
Percentage of Missing	5	0.00660031	0.00016511	0.00016398	**0.00010004**	0.00022420	0.00016509
	10	0.01780242	0.00052670	0.00052603	**0.00029265**	0.00056752	0.00052673
	15	0.03233435	0.00103031	0.00103166	**0.00059281**	0.00106831	0.00103033
	200	0.04790646	0.00163309	0.00164281	**0.00097521**	0.00169403	0.00163308
		0.02616088	0.00083880	0.00084112	**0.00049018**	0.00088851	0.00083881
Missing type	MCAR	0.00681481	0.00085183	0.00082675	**0.00000048**	0.00001265	0.00085146
	MAR	0.00608364	0.00052883	0.00051739	**0.00000355**	0.00001051	0.00052859
	NMAR	0.08968299	**0.00084387**	0.00088772	0.00353293	0.00603432	0.00084455
		0.03419381	0.00074151	0.00074395	**0.00117899**	0.00201916	0.00074153
Overall		0.02739119	0.00077387	0.00077579	**0.00065524**	0.00116636	0.00077388

4. Conclusions and Recommendations

Six imputation methods were applied to treat the problem of missing data. We reviewed and provided technical details of the different methods used included MI RI RSRI EM MUL and AJRI. As depicted in Table 1-6, all imputation methods led to an improvement in prediction accuracy, as measured by MSE of mean variance correlation regression coefficient and R^2. The best imputation under MSE of mean variance and correlation classified by sample sizes and by percentage of missing data were obtained using RSRI EM and RI respectively, classified by missing data type was obtained using RSRI for mean and EM for variance and correlation. The best imputation under MSE of regression coefficient was obtained using EM and RSRI for model 1 and model 2 respectively. The best imputation under MSE of R^2 was obtained using EM.

For AJRI, the best imputation is under MSE of correlation with sample size 500 and 1000, 5 percentage of missing data, under MSE of mean and regression coefficient with MCAR.

5. Acknowledgements

This work is fully achieved by the collaborations of the Applied Statistics Department, Faculty of Science, Ubon Ratchathani Rajabhat University, Thailand.

6. References

[1] Becker, W. E., and W. B. Walstad. Data loss from pretest to posttest as a sample selection problem. Review of Economics and Statistics. 1990; 72 (1): 184-188.

[2] Becker, W., & Powers, J. Student performance, attrition, and class size given missing student data. Economics of Education Review. 2001. 20, 377-388.

[3] Rubin, D.B. Multiple imputation for nonresponse in surveys, New York: John Wiley & Sons, Inc.; 1987.

[4] Chen, S. X., Leung, D. H. Y. and Qin, J. Improving semiparametric estimation by using surrogate data. Journal of Royal Statistical Society: Series B. 2008; 70: 803-823.

[5] Kott, P. S. and Chang, T. Using calibration weighting to adjust for nonignorable unit nonresponse. Journal of the American Statistical Association. 2010; 105, 1265-1275.
[6] Anderson, A.B., Basilevsky, A. & Hum, D. P. J. Missing data: A review of the literature. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), Handbook of survey research (pp. 415-494). San Diego: Academic Press.; 1983.
[7] Kim, J.O. and Curry, J. The Treatment of Missing Data in Multivariate Analysis. Sociological Methods Research. 1977; 6:215-240.
[8] Cohen, J., & Cohen, P. Missing data. In J. Cohen & P. Cohen, Applied multiple regression: Correlation analysis for the behavioral sciences (pp. 275-300).; 1983.
[9] Little, R.J. and D.B. Rubin, Statistical Analysis with Missing Data. 2nd Edn., John Wiley and Sons, New York, ISBN: 978-0-471-18386-0, pp: 408. 2002.
[10] Im, Jongho. Some methods for handling missing data in surveys. PhD [Dissertations]. Iowa State University; 2015.
[11] Quenonuille, M. H. Notes on Bias in Estimation. Biometrika. 1956;43, 353-360.
[12] Yu, C. H., Resampling Methods: Concepts, Applications, and Justification, Practical Assessment, Research and Evaluation. 2003: 8(19).
[13] Sahinler S., & Topuz, D. Bootstrap and Jackknife Resampling Algorithms for Estimation of Regression Parameters. Journal of Applied Qualitative Methods. 2007;2(2) Summer, 188-199.
[15] Little, R.J.A. Regression with missing X's:a review, Journal of the American Statistical Association. 1992;87: 1227–1237.
[14] Draper, Norman R., & Smith, H.. Applied Regression Analysis. 3rd ed. John Willey & Sons, Inc., New York; 1998.
[16] Vongprasert, J. and B. Premanode, Missing data imputation using weighted of regime switching mean and regression. J. Math. Stat. 2014; 10: 255-261.
[17] Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B. 1977;39 (1): 1–38.
[18] Rubin, D.B. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.