*Journal of the Korean*
*Data & Information Science Society*
*2006, Vol. 17, No. 3, pp. 871~888*

# A Study on Imputation using Adjusted Cohen Method[1)]

Sung-Suk Chung[2)] · Young-Min Chun[3)] · Sun-Kyung Lee[4)]

Abstract

Many studies have been done to develop procedures to deal with missing values. Most common method is to reassign the other values to the missing data. The purpose of our study is to suggest adjusted Cohen methods and to compare the efficiency of them with other methods through a simulation study. The adjusted Cohen methods use an auxiliary variable to arrange ranking of the variable with missing values. It leads to a reduced mean square error(MSE) compared with the Cohen method.

***Keywords*** : Adjusted Cohen method, Imputation method, Missing data patterns, Missing mechanisms

## 1. Introduction

Most statistical analyses are performed by complete data but missing values almost always exist in real data. Missing data may occur due to different reasons such as death of patients, equipment malfunctions, refusal of respondents to answer certain questions, and so on. If we don't include all the observations with missing values, the loss of information might be significant. This approach also ignores the possible systematic difference between the complete cases and incomplete cases, and the resulting inference may not be appropriate, especially

---

1) This work was supported by grant No. R01-2005-000-10752-0 from Ministry of Science and Technology(Korea Science and Engineering Foundation).

2) Professor, Div. of Mathematics and Statistical Informatics (Institute of Applied Statistics), Chonbuk National University, Jeonju, Korea
E-mail : sschung@chonbuk.ac.kr

3) Doctoral Course, Dept. of Computer and Statistical Informatics, Chonbuk National University, Jeonju, Korea
E-mail : zzari@chonbuk.ac.kr

4) Master, Dept. of Statistical Informatics, Chonbuk National University, Jeonju, Korea
E-mail : bandee77@hotmail.com

when the missing rate of data is high. Therefore many studies have been done to develop procedures to deal with missing values.

There are two types of nonresponse: unit nonresponse, in which the entire observation unit is missing, and item nonresponse, in which more than one item is missing for an observation unit. One of the most common methods to deal with unit nonresponse is weighting adjustment which is to reassign the weights of the nonresponse to the response. Common methods for item nonresponse are imputation methods. Imputation is a general and flexible method for handling missing data problems. Imputation is a procedure that replaces the missing values in a data set by predicted or simulated values. The basic object of imputation is to allow end users to apply their existing analysis tools to any dataset with missing values using the same command structure and output standards as if there were no missing data.

Imputation methods are divided into single imputation and multiple imputation. Single imputation substitutes a value for each missing value, so it is easy to use and simple, but it has serious drawbacks like reduction variance. Multiple imputation replaces each missing value with more than one value to represent imputation uncertainty. It offers valid result but it is a burden to impute missing values and analysis for several times.

The purpose of this study is to suggest adjusted Cohen methods and to compare the efficiency of them with other methods through a simulation study. The adjusted Cohen methods use an auxiliary variable to arrange ranking of the variable with missing values. It leads to a reduced mean square error(MSE) compared with the Cohen method. Cohen(1996) proposed a new approach that complements the underestimation variance of mean imputation, but it tends to inflate the MSE. We are not concerned here with categorical variables involving missing values.

This study consists of five sections. Section 1 reviews the missing data problems. Section 2 explains missing data mechanisms and patterns. Missing data mechanisms are MCAR, MAR and NMAR and missing data patterns are monotone and arbitrary. Section 3 discusses several imputation methods that are mean imputation, cohen method, regression imputation, and multiple imputation. Section 4 suggests the adjusted Cohen methods and reports the results of a simulation study. Section 5 comments about conclusions and future directions.

# 2. Missing Mechanisms and Patterns

## 2.1 Missing Mechanisms

Missing mechanisms concern the relationship between missingness and the values of variables in the data matrix. These are MCAR, MAR and NMAR(Little and Rubin(2002)).

The missing data for a variable Y is "Missing Completely at Random(MCAR)", if the probability of having a missing value for Y is unrelated to the value of Y itself or any other variable in the data set.

For example, income is MCAR, if two conditions are satisfied. One is that people who do not report their income have, on the average, the same income as people who do report income. The other is that each of the other variables in the data set would have to be the same, on average, for the people who did not report their income and the people who did report their income. MCAR is the best situation to treat the missing data.

The missing data for a variable Y is "Missing At Random(MAR)", if the probability of the missing data of Y is unrelated to the value of Y, but to other variables. MCAR is a special type of MAR and MAR is much weaker assumption than MCAR.

For example, income is MAR, if the probability of missing data of income depends on martial status, but within each category of martial status, the probability of missing data on income is unrelated to the value of income. MAR and MCAR are ignorable missingness.

The missing data for a variable Y is "Not Missing at Random(NMAR)", if the probability of missing data of Y is related to the actual value of the missing data.

For example, if high income households are less likely to report their income even after adjusting for other variables, then the probability of missing income is not ignorable. This is the most difficult condition to modeling.

## 2.2 Missing Data Patterns

Missing data patterns describe which values are observed in the data matrix and which are missing(Little and Rubin(2002)). A data set is said to have monotone missing patterns when the missing particular variable $x_j$ implies that all subsequent variables $x_k$ are all missing, for $k > j$. Simpler imputation methods can be used, if the missing data patterns are monotone, however a monotone pattern is uncommon in most real data.

In an arbitrary pattern, missing data can occur anywhere. The MCMC algorithm, introduced in section 3.4.1, is appropriate for missing data which has an

arbitrary pattern. Another way to handle a missing data set with an arbitrary pattern is to use the MCMC algorithm to impute enough values to make the missing data patterns monotone. Then, a simpler imputation method can be used.

## 3. Imputation Methods

### 3.1 Mean Imputation

Mean imputation is the simplest and the oldest method. This method replaces missing values with the mean of observed values. The mean is formed in conditional or unconditional situation.

The mean of the observed values is given as $\bar{y}_{j(obs)} = \frac{1}{n_{obs}} \sum_{observed} y_{ij}$. With unconditional mean imputation, the mean estimator is given as $\hat{\mu_j} = \frac{1}{n}\left[\sum_{observed} y_{ij} + \sum_{missing} \bar{y}_{j(obs)}\right] = \bar{y}_{j(obs)}$ and the variance estimator is given as $\hat{\sigma_j^2} = \left[\frac{1}{n_{obs}-1} \sum_{observed} (y_{ij} - \bar{y}_{j(obs)})^2\right] \frac{(n_{obs}-1)}{n-1}$. The notation $Y_j$ is $j^{th}$ random variable, $n$ is sample size, $n_{obs}$ is a number of observed values of $Y_j$.

Under MCAR assumption, $\hat{\mu_j}$ is unbiased but biased in general and the variance estimator underestimates the variance by a factor of $(n_{obs}-1)/(n-1)$. The covariance is also underestimated by this method and if the variables are highly correlated, this method cannot be recommended.

Conditional mean imputation first uses some auxiliary variables to form adjustment classes, and then replaces missing values in each class with its sample mean.

### 3.2 Cohen Method

Cohen(1996) suggested an approach that makes use of imputed values distributed more diffusely than the observed data. For example, instead of imputing the mean for all the missing values, half of the missing values are imputed by $\bar{y}_{j(obs)} + \sqrt{\frac{n+n_{obs}-1}{n_{obs}-1}} D_{obs}$ and the other half by $\bar{y}_{j(obs)} - \sqrt{\frac{n+n_{obs}-1}{n_{obs}-1}} D_{obs}$, where $D_{obs}^2 = \frac{1}{n_{obs}} \sum_{observed} (y_{ij} - \bar{y}_{j(obs)})^2$. This approach is effective to adjust reduced variance of mean imputation.

## 3.3 Regression Imputation

The idea of this method is that missing values are replaced by predicted values derived from a regression model. In regression imputation, missing values of a variable $Y_j$ are imputed by predicted values from the regression of $Y_j$ on $Y_1, \cdots, Y_{j-1}, Y_{j+1}, \cdots, Y_k$ , based on the $n_{obs}$ complete cases, where $Y_1, \cdots, Y_{j-1}, Y_{j+1}, \cdots, Y_k$ are fully observed and $Y_j$ is observed for the $n_{obs}$ observations. The result equation is given as $\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 y_{i1} + \cdots + \hat{\beta}_{j-1} y_{ij-1} + \hat{\beta}_{j+1} y_{ij+1} + \cdots + \hat{\beta}_k y_{ik}$ .

This method requires a model and assumes that missing data are MAR. The drawback of regression imputation is that this inflates correlations and becomes difficult in multivariate data when more than one variable has missing values, i.e. a data set that has multiple missing.

The stochastic regression imputation method replaces a missing value by a value predicted by regression imputation plus a random error. The form of the equation is like $\hat{y}_{ij} + \epsilon_i$ where $\epsilon_i$ is a random value from a normal distribution with zero mean and the variance equal to the residual variance in the regression.

## 3.4 Multiple Imputation

Multiple imputation is one of the most attractive methods for general purpose handling of missing data in multivariate analysis. It is first proposed by Rubin(1976) and elaborated in his(1987) book. Rubin described multiple imputation as a three-step process.

- Step 1 : The missing data are imputed in m times to generate m complete data sets.
- Step 2 : The m complete data sets are analyzed by using standard statistical analysis.
- Step 3 : The results from the m complete data sets are combined to produce one overall analysis.

Multiple imputation requires MAR or MCAR assumption and represents a random sample of the missing values rather than attempting to estimate each missing value.

This process results in valid statistical inferences that precisely reflect the uncertainty due to missing values. Uncertainty is accounted by creating different versions of the missing data and observing the variability between imputed data

sets. The disadvantage of multiple imputation is that it takes more work to create the imputations and analyze the results than single imputation and the statistical principles behind multiple imputation are not trivial.

### 3.4.1 Imputation step using Markov Chain Monte Carlo(MCMC)

In the imputation step, a variety of imputation methods have been used. The method of choice relies upon the type of missing data patterns. For an monotone missing data pattern, simple methods have been proposed. Propensity methods or predictive mean matching is appropriate for continuous variables and discriminant analysis or logistic regression for discrete variables. For an arbitrary missing data pattern, the Markov Chain Monte Carlo(MCMC) that assumes multivariate normal distribution has been suggested.

A Markov chain is a sequence of random variables in which the distribution of each element depends on the value of the previous one.

The first step computes mean vector and covariance matrix from the data that does not have missing values to estimate the prior distribution. Next, the imputation step simulates values for missing values by randomly selecting a value from the available distribution of values. The posterior steps recomputes mean vector and covariance matrix with the imputed values from the imputation step. This is posterior distribution. Imputation step and posterior step are iterated until mean vector and covariance matrix are unchanging as we iterate.

When we denote the variables with missing values for observation i by $Y_{i\,(mis)}$ and the variables with observed values by $Y_{i\,(obs)}$, the imputation step draws $Y_{mis}^{(t+1)}$ from $p(Y_{mis}|Y_{obs},\theta^{(t)})$ with a current parameter estimate $\theta^{(t)}$ at $t^{(th)}$ iteration. The posterior step draws $\theta^{(t+1)}$ from $p(\theta|Y_{obs},Y_{mis}^{(t+1)})$. This creates a Markov chain $(Y_{mis}^{(1)},\theta^{(1)}),(Y_{mis}^{(2)},\theta^{(2)}),\cdots,(Y_{mis}^{(t)},\theta^{(t)}),(Y_{mis}^{(t+1)},\theta^{(t+1)}),\cdots$, which converges in distribution to $p(Y_{mis},\theta|Y_{obs})$.

### 3.4.2 Combination results

Combining inference from the imputed data sets is done using rules conformed by Rubin(1987). Rubin detailed that combining the estimates of the point and variance for a parameter of interest. When $\hat{\theta}_i$ and $\widehat{W}_i$ are the point and variance estimates from the $i$th imputed data set, $i=1,2,...,m$ , the point estimate for $\theta$ from multiple imputation is the average of the $m$ complete data estimates : $\bar{\theta}=\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$ .

The total variance is expressed by the formula $T = \overline{W} + (1 + \frac{1}{m})B$, where $\overline{W} = \frac{1}{m}\sum_{i=1}^{m} \widehat{W}_i$, $B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\theta})^2$. $\overline{W}$ is the "within imputation variance" and B is the "between imputation variance". The former means the natural variability and the latter estimates uncertainty caused by missing data.

### 3.4.3 Multiple imputation efficiency

When we generate $m$ complete sets, we have to determine how many. Then, we can consult the relative efficiency of an estimate based on $m$ imputation which is showed by Rubin(1987,p.114). The relative efficiency(RE) is approximately given as a function of $m$ and $\gamma$. $RE = (1 + \frac{\gamma}{m})^{-1}$ , where $\gamma = \frac{r + 2/(df+3)}{r+1}, r = \frac{(1+m^{-1})B}{\overline{W}}$ , $df = (m-1)\left[1 + \frac{m\overline{W}}{(m+1)B}\right]$. The ratio $r$ is called the relative increase in variance due to nonresponse and $\gamma$ is the rate of missing information for the quantity being estimated(Rubin, 1987). The following Table 1 shows the RE with different value of $m$ and $\gamma$. Surprisingly, for cases with little missing information, only 3 ~ 10 imputations are enough.

<Table 1> Relative efficiency

| m \ γ | .1 | .2 | .3 | .5 | .7 |
|---|---|---|---|---|---|
| 3 | 0.9677 | 0.9375 | 0.9091 | 0.8571 | 0.8108 |
| 5 | 0.9804 | 0.9615 | 0.9434 | 0.9091 | 0.8772 |
| 10 | 0.9901 | 0.9804 | 0.9709 | 0.9524 | 0.9346 |
| 20 | 0.9950 | 0.9901 | 0.9852 | 0.9756 | 0.9662 |

Besides there are hot deck imputation, cold deck imputation, ratio imputation, and EM algorithm, etc. With hot deck imputation, missing values are replaced by values from the responding units in the sample that are derived sequentially, hierarchically or via a distance function. In the cold deck method, missing values are replaced by values from an external source, such as a value from a previous result of the same survey. In ratio imputation, $\hat{y}_{ij} = \frac{\bar{y}_{j(obs)}}{\bar{y}_{j'(obs)}} y_{ij'}$ is used as imputed values for the $i$-th missing value. This ratio imputation may provide very precise imputation if the missingness of $y_j$ mainly depends on a highly correlated an auxiliary variable $y_{j'}$. The EM algorithm is a very general

iterative algorithm for maximum likelihood estimation in an incomplete data problem(Little and Rubin(1987)). This method replaces the missing values by using observed values and a parameter and then reestimates the parameter based on observed values and the imputed values until iterating converges.

## 4. Adjusted Cohen Methods and a Simulation Study

### 4.1 Adjusted Cohen Methods

We reviewed the imputation methods. According to Scheffer(2002), multiple imputation is always better than case deletion or single imputation. Also many statistical software packages support multiple imputation. However, Horton and Lipsitz(2001) mentioned that none of the packages are clearly superior and they remain in large part of a complicated black box whose output can be difficult to interpret it in their study. This ultimately makes end users shun its use.

Therefore, this study proposes a new approach that improves drawback of single imputation. The Cohen method was introduced in 3.2. This approach complements the underestimation of variance that is the chief drawback of mean imputation. However, this method has a result even worse than mean imputation when comparing mean square error(MSE). It leads to inflation of MSE. Moreover, it tends to overestimate variance under MCAR assumption and it is not effective to adjust estimates of means.

This study suggests adjusted Cohen methods. This approach uses an auxiliary variable to arrange ranking of the variable with missing values, under the assumption that the auxiliary variable is fully observed. The missing values are imputed by more diversified values after sorting the variable with missing values by an auxiliary variable. Following figures show adjusted Cohen methods.

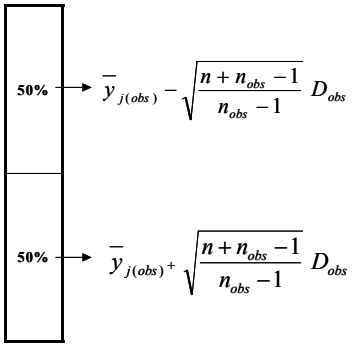
The first, two values can be used to impute like Cohen method. This method is diagrammed as figure 1. In the adjusted Cohen method 1, if a missing value is within the first 50% of ranking, the missing value is imputed by $\bar{y}_{j(obs)} - \sqrt{\frac{n+n_{obs}-1}{n_{obs}-1}} D_{obs}$.

The second, mean of observed values can be added. This makes the adjusted Cohen method 2. The adjusted Cohen method 3 and 4 use four different values to impute, but percentage of ranking is unlike. The adjusted Cohen method 5 and 6 use five different values involving the mean of the observed values to impute, but also percentage of ranking is unlike. However, we have to conform that the variable with missing values has positive correlation with the auxiliary variable; if negative correlation exists, we can convert a sign between $\bar{y}_{j(obs)}$ and

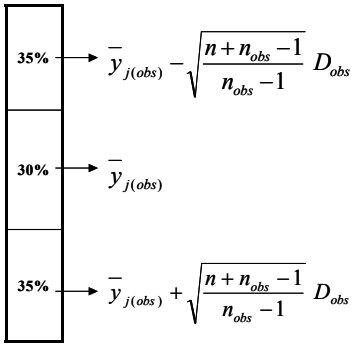
$$\sqrt{\frac{n+n_{obs}-1}{n_{obs}-1}}D_{obs} \text{ or } \frac{1}{2}\sqrt{\frac{n+n_{obs}-1}{n_{obs}-1}}D_{obs}.$$

The adjusted Cohen methods lower MSE and are useful to complement overestimation of variance under MCAR assumption. Mean estimates of the adjusted Cohen methods are mediated under MAR and NMAR, while the Cohen method has the same mean estimate with mean imputation. These methods don't need a model like regression imputation, and can also be used in multiple missing unless there is no variable to use as an auxiliary variable.

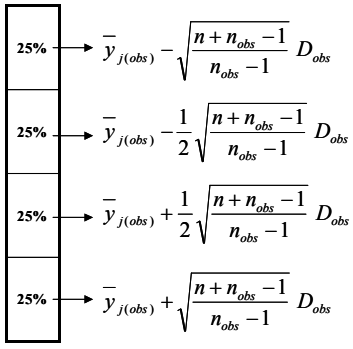
In the next section, a simulation study is performed to compare the efficiency of each adjusted Cohen method and with other imputation methods.



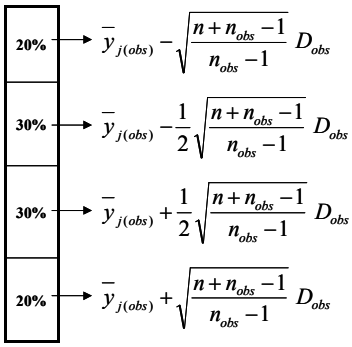

<Figure 1>
Adjusted Cohen method 1



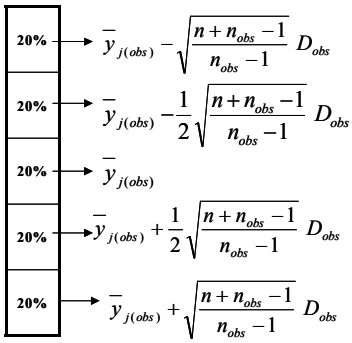

<Figure 2>
Adjusted Cohen method 2



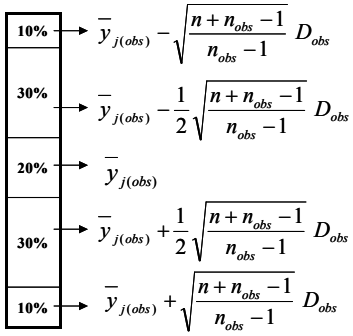

<Figure 3>
Adjusted Cohen method 3




<Figure 4>
Adjusted Cohen method 4




<Figure 5>
Adjusted Cohen method 5




<Figure 6>
Adjusted Cohen method 6

## 4.2 A Simulation Study

### 4.2.1 Simulation design

(1) Data

Simulation is performed by Iris data and generated data. Table 2 shows detailed information.

<Table 2> Simulation design of data

| data | number of unit | variables |
|---|---|---|
| Iris | 150 (50 units of each of 3 species ) | · sepal length(SL) · sepal width(SW)<br>· petal length(PL) · petal width(PW)<br>· species of iris (categorical variable) |
| Generated | 1000 | · $x_1 \sim N(3,2)$ · $x_2 \sim N(5,3)$ · $x_3 \sim N(2,1)$<br>· $y = x_1 + x_2 + x_3 + \epsilon$, where $\epsilon \sim N(0,1)$ |

(2) Missing mechanisms

Simulation is performed under MCAR, MAR, and NMAR assumptions. Sepal length of Iris data is missing in the first simulation, petal length is missing in the second and $y$ of generated data is missing in the third. Table 3 shows detail information. With notes, 'depends on PL' means sepal length of Iris data is first sorted by petal length and then omitted as missing rate. This simulation design is referred by the literature of Allison(2000), Horton and Lipsitz(2001) and Scheffer(2002). The simulation using Iris data is done separately, because the distributions of sepal length and petal length have different shapes.

<Table 3> Simulation missing mechanisms

| data | mechanisms | variable with missing values | notes |
|---|---|---|---|
| Iris | MCAR | · SL · PL | · missing randomly, each 1000 times |
| | MAR | · SL · PL | · depends on PL · depends on PW |
| | NMAR | · SL · PL | · depends on itself |
| Generated | MCAR | · $y$ | · missing randomly, each 1000 times |
| | MAR | · $y$ | · depends on $x_2$ |
| | NMAR | · $y$ | · depends on itself |

(3) Missing rates

The simulation is done with seven types of missing rates, 5, 10, 15, 20, 30, 40, 50 percent, for each of the missing mechanisms and data.

A simulation was done to compare results of adjusted Cohen methods. We compared the imputation methods in terms of the mean and standard deviation of data and mean square error of imputed values.

Table 4 and Table 5 presents the mean, standard deviation and MSE(mean square error) of adjusted Cohen methods of Iris sepal length data. In MCAR, adjusted Cohen6 has the best result – nearest mean, nearest standard deviation and smallest MSE. Also, in MAR and NMAR, adjusted Cohen1 gives the best result. Simulation of Iris petal length data and generated data were also performed and the results were similar to the result of Iris sepal length data. Therefore, in this study, we employed adjusted Cohen6 for MCAR and adjusted Cohen1 for MAR and NMAR.

(4) Imputation methods

Five imputation methods are used, which are multiple imputations using MCMC algorithm, regression imputation, mean imputation, Cohen method and adjusted Cohen method.

<Table 4> Mean and SD of adjusted Cohen methods

| | A-Cohen | Mean | | | | | | | | SD | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0% | 5% | 10% | 15% | 20% | 30% | 40% | 50% | 0% | 5% | 10% | 15% | 20% | 30% | 40% | 50% |
| MCAR | 1 | 5.843 | 5.842 | 5.843 | 5.842 | 5.842 | 5.843 | 5.842 | 5.838 | 0.828 | 0.847 | 0.872 | 0.895 | 0.922 | 0.986 | 1.065 | 1.167 |
| | 2 | 5.843 | 5.842 | 5.842 | 5.841 | 5.840 | 5.839 | 5.837 | 5.835 | 0.828 | 0.835 | 0.846 | 0.858 | 0.871 | 0.908 | 0.956 | 1.026 |
| | 3 | 5.843 | 5.842 | 5.842 | 5.842 | 5.841 | 5.841 | 5.839 | 5.837 | 0.828 | 0.832 | 0.840 | 0.848 | 0.858 | 0.886 | 0.927 | 0.989 |
| | 4 | 5.843 | 5.842 | 5.843 | 5.843 | 5.842 | 5.843 | 5.842 | 5.840 | 0.828 | 0.830 | 0.834 | 0.838 | 0.845 | 0.866 | 0.898 | 0.950 |
| | 5 | 5.843 | 5.842 | 5.843 | 5.842 | 5.842 | 5.843 | 5.842 | 5.841 | 0.828 | 0.828 | 0.829 | 0.832 | 0.836 | 0.851 | 0.877 | 0.922 |
| | 6 | 5.843 | 5.842 | 5.843 | 5.842 | 5.841 | 5.843 | 5.842 | 5.841 | 0.828 | 0.819 | 0.814 | 0.809 | 0.805 | 0.804 | 0.810 | 0.831 |
| MAR | 1 | 5.843 | 5.827 | 5.835 | 5.837 | 5.844 | 5.891 | 5.973 | 5.962 | 0.828 | 0.838 | 0.826 | 0.820 | 0.808 | 0.755 | 0.673 | 0.665 |
| | 2 | 5.843 | 5.827 | 5.835 | 5.837 | 5.844 | 5.891 | 6.019 | 6.113 | 0.828 | 0.838 | 0.826 | 0.820 | 0.808 | 0.755 | 0.665 | 0.648 |
| | 3 | 5.843 | 5.827 | 5.835 | 5.837 | 5.844 | 5.916 | 6.045 | 6.089 | 0.828 | 0.838 | 0.826 | 0.820 | 0.808 | 0.739 | 0.631 | 0.600 |
| | 4 | 5.843 | 5.827 | 5.835 | 5.837 | 5.844 | 5.945 | 6.071 | 6.116 | 0.828 | 0.838 | 0.826 | 0.820 | 0.808 | 0.718 | 0.612 | 0.582 |
| | 5 | 5.843 | 5.827 | 5.835 | 6.004 | 5.844 | 5.945 | 6.071 | 6.168 | 0.828 | 0.838 | 0.826 | 0.820 | 0.808 | 0.718 | 0.612 | 0.589 |
| | 6 | 5.843 | 5.827 | 5.835 | 5.864 | 5.900 | 5.998 | 6.121 | 6.219 | 0.828 | 0.838 | 0.826 | 0.797 | 0.761 | 0.675 | 0.573 | 0.544 |
| NMAR | 1 | 5.843 | 5.803 | 5.772 | 5.751 | 5.721 | 5.655 | 5.555 | 5.416 | 0.828 | 0.754 | 0.710 | 0.682 | 0.646 | 0.593 | 0.528 | 0.491 |
| | 2 | 5.843 | 5.803 | 5.766 | 5.732 | 5.685 | 5.605 | 5.495 | 5.371 | 0.828 | 0.754 | 0.706 | 0.672 | 0.630 | 0.568 | 0.490 | 0.427 |
| | 3 | 5.843 | 5.803 | 5.766 | 5.732 | 5.694 | 5.617 | 5.502 | 5.364 | 0.828 | 0.754 | 0.703 | 0.666 | 0.625 | 0.560 | 0.477 | 0.418 |
| | 4 | 5.843 | 5.803 | 5.766 | 5.732 | 5.691 | 5.605 | 5.492 | 5.353 | 0.828 | 0.754 | 0.703 | 0.666 | 0.622 | 0.551 | 0.470 | 0.410 |
| | 5 | 5.843 | 5.753 | 5.766 | 5.729 | 5.682 | 5.592 | 5.480 | 5.346 | 0.828 | 0.754 | 0.703 | 0.665 | 0.621 | 0.548 | 0.464 | 0.397 |
| | 6 | 5.843 | 5.803 | 5.759 | 5.716 | 5.661 | 5.561 | 5.445 | 5.312 | 0.828 | 0.754 | 0.697 | 0.653 | 0.602 | 0.521 | 0.434 | 0.367 |

<Table 5> MSE of adjusted Cohen methods

| | A-Cohen | 5% | 10% | 15% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|---|
| MCAR | 1 | 0.027 | 0.061 | 0.091 | 0.128 | 0.212 | 0.318 | 0.462 |
| | 2 | 0.015 | 0.033 | 0.051 | 0.070 | 0.116 | 0.174 | 0.251 |
| | 3 | 0.012 | 0.027 | 0.041 | 0.057 | 0.094 | 0.139 | 0.203 |
| | 4 | 0.011 | 0.024 | 0.036 | 0.051 | 0.082 | 0.120 | 0.173 |
| | 5 | 0.009 | 0.019 | 0.029 | 0.041 | 0.065 | 0.095 | 0.136 |
| | 6 | 0.007 | 0.016 | 0.024 | 0.033 | 0.050 | 0.068 | 0.091 |
| MAR | 1 | 0.017 | 0.018 | 0.023 | 0.030 | 0.043 | 0.093 | 0.133 |
| | 2 | 0.017 | 0.018 | 0.023 | 0.030 | 0.043 | 0.133 | 0.213 |
| | 3 | 0.017 | 0.018 | 0.023 | 0.030 | 0.064 | 0.157 | 0.199 |
| | 4 | 0.017 | 0.018 | 0.023 | 0.030 | 0.080 | 0.184 | 0.231 |
| | 5 | 0.017 | 0.018 | 0.023 | 0.030 | 0.080 | 0.184 | 0.273 |
| | 6 | 0.017 | 0.018 | 0.034 | 0.056 | 0.125 | 0.246 | 0.346 |
| NMAR | 1 | 0.035 | 0.060 | 0.078 | 0.105 | 0.185 | 0.314 | 0.511 |
| | 2 | 0.035 | 0.071 | 0.107 | 0.159 | 0.254 | 0.392 | 0.567 |
| | 3 | 0.035 | 0.067 | 0.098 | 0.135 | 0.224 | 0.364 | 0.557 |
| | 4 | 0.035 | 0.060 | 0.098 | 0.138 | 0.234 | 0.378 | 0.573 |
| | 5 | 0.035 | 0.067 | 0.103 | 0.154 | 0.257 | 0.399 | 0.586 |
| | 6 | 0.035 | 0.075 | 0.119 | 0.180 | 0.297 | 0.452 | 0.647 |

### 4.2.2 Simulation results

Table 6 shows the mean and standard deviation when there are diverse missing ratio, under the various missing mechanisms. The first column represents the mean and standard deviations when the data are complete.

In common, as missing rate increases, standard deviations are decreased

seriously and means are also affected except for the MCAR assumption.

(1) The results of Iris Sepal length data

The mean, standard deviation and MSE of Iris sepal length data are represented in Table 7 and Table 8. Under MCAR, all of these methods estimate the true mean very well. Even at 50% missing, all means are within 1% of the target value. In the MAR

<Table 6> Mean and SD of Iris SL, Iris PL and generated data

| data | | | 0% | 5% | 10% | 15% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|---|---|---|
| Iris SL | Mean | MCAR | 5.843 | 5.843 | 5.843 | 5.842 | 5.841 | 5.843 | 5.843 | 5.843 |
| | | MAR | 5.843 | 5.882 | 5.950 | 6.004 | 6.069 | 6.212 | 6.367 | 6.477 |
| | | NMAR | 5.843 | 5.753 | 5.674 | 5.612 | 5.541 | 5.411 | 5.285 | 5.158 |
| | SD | MCAR | 0.828 | 0.827 | 0.828 | 0.827 | 0.826 | 0.825 | 0.826 | 0.826 |
| | | MAR | 0.828 | 0.820 | 0.791 | 0.773 | 0.750 | 0.687 | 0.605 | 0.595 |
| | | NMAR | 0.828 | 0.738 | 0.680 | 0.643 | 0.600 | 0.535 | 0.459 | 0.392 |
| Iris PL | Mean | MCAR | 3.758 | 3.757 | 3.757 | 3.756 | 3.760 | 3.757 | 3.763 | 3.768 |
| | | MAR | 3.758 | 3.874 | 4.019 | 4.155 | 4.340 | 4.751 | 5.050 | 5.221 |
| | | NMAR | 3.758 | 3.622 | 3.484 | 3.364 | 3.221 | 2.950 | 2.640 | 2.269 |
| | SD | MCAR | 1.765 | 1.724 | 1.673 | 1.629 | 1.578 | 1.472 | 1.361 | 1.243 |
| | | MAR | 1.765 | 1.725 | 1.666 | 1.601 | 1.478 | 1.064 | 0.732 | 0.670 |
| | | NMAR | 1.765 | 1.694 | 1.642 | 1.602 | 1.552 | 1.470 | 1.357 | 1.174 |
| generated | Mean | MCAR | 11.044 | 11.043 | 11.043 | 11.048 | 11.044 | 11.044 | 11.041 | 11.050 |
| | | MAR | 11.044 | 11.355 | 11.609 | 11.827 | 12.080 | 12.528 | 13.009 | 13.412 |
| | | NMAR | 11.044 | 10.601 | 10.257 | 9.954 | 9.667 | 9.109 | 8.533 | 7.938 |
| | SD | MCAR | 3.916 | 3.916 | 3.917 | 3.914 | 3.917 | 3.917 | 3.911 | 3.917 |
| | | MAR | 3.916 | 3.732 | 3.616 | 3.549 | 3.438 | 3.313 | 3.190 | 3.212 |
| | | NMAR | 3.916 | 3.473 | 3.234 | 3.068 | 2.932 | 2.706 | 2.490 | 2.302 |

<Table 7> Mean and SD of Iris SL data

| | | Mean | | | | | | | | SD | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0% | 5% | 10% | 15% | 20% | 30% | 40% | 50% | 0% | 5% | 10% | 15% | 20% | 30% | 40% | 50% |
| MCAR | MCMC | 5.843 | 5.843 | 5.842 | 5.840 | 5.844 | 5.838 | 5.843 | 5.844 | 0.828 | 0.828 | 0.831 | 0.831 | 0.831 | 0.836 | 0.837 | 0.845 |
| | Reg | 5.843 | 5.843 | 5.843 | 5.842 | 5.842 | 5.842 | 5.842 | 5.840 | 0.828 | 0.825 | 0.822 | 0.819 | 0.816 | 0.810 | 0.805 | 0.800 |
| | Mean | 5.843 | 5.843 | 5.843 | 5.842 | 5.841 | 5.843 | 5.843 | 5.843 | 0.828 | 0.808 | 0.785 | 0.763 | 0.738 | 0.690 | 0.638 | 0.582 |
| | Cohen | 5.843 | 5.835 | 5.835 | 5.842 | 5.841 | 5.834 | 5.843 | 5.833 | 0.828 | 0.847 | 0.872 | 0.895 | 0.923 | 0.987 | 1.066 | 1.168 |
| | A-Cohen6 | 5.843 | 5.842 | 5.843 | 5.842 | 5.841 | 5.843 | 5.842 | 5.841 | 0.828 | 0.819 | 0.814 | 0.809 | 0.805 | 0.804 | 0.810 | 0.831 |
| MAR | MCMC | 5.843 | 5.831 | 5.829 | 5.817 | 5.815 | 5.728 | 5.594 | 5.617 | 0.828 | 0.837 | 0.845 | 0.861 | 0.864 | 0.968 | 1.135 | 1.111 |
| | Reg | 5.843 | 5.832 | 5.832 | 5.827 | 5.823 | 5.788 | 5.682 | 5.659 | 0.828 | 0.833 | 0.834 | 0.839 | 0.841 | 0.878 | 0.991 | 1.007 |
| | Mean | 5.843 | 5.882 | 5.950 | 6.004 | 6.069 | 6.212 | 6.367 | 6.477 | 0.828 | 0.801 | 0.750 | 0.713 | 0.670 | 0.574 | 0.467 | 0.419 |
| | Cohen | 5.843 | 5.874 | 5.942 | 6.004 | 6.069 | 6.205 | 6.367 | 6.470 | 0.828 | 0.840 | 0.834 | 0.837 | 0.838 | 0.821 | 0.781 | 0.842 |
| | A-Cohen1 | 5.843 | 5.827 | 5.835 | 5.837 | 5.844 | 5.891 | 5.973 | 5.962 | 0.828 | 0.838 | 0.826 | 0.820 | 0.808 | 0.755 | 0.673 | 0.665 |
| NMAR | MCMC | 5.843 | 5.823 | 5.800 | 5.767 | 5.739 | 5.686 | 5.643 | 5.621 | 0.828 | 0.793 | 0.765 | 0.724 | 0.696 | 0.653 | 0.626 | 0.628 |
| | Reg | 5.843 | 5.821 | 5.795 | 5.764 | 5.731 | 5.676 | 5.624 | 5.562 | 0.828 | 0.787 | 0.748 | 0.713 | 0.676 | 0.624 | 0.585 | 0.546 |
| | Mean | 5.843 | 5.753 | 5.674 | 5.612 | 5.541 | 5.417 | 5.285 | 5.158 | 0.828 | 0.721 | 0.644 | 0.594 | 0.536 | 0.447 | 0.355 | 0.276 |
| | Cohen | 5.843 | 5.746 | 5.667 | 5.612 | 5.541 | 5.411 | 5.285 | 5.154 | 0.828 | 0.756 | 0.716 | 0.696 | 0.671 | 0.640 | 0.593 | 0.555 |
| | A-Cohen1 | 5.843 | 5.803 | 5.772 | 5.751 | 5.721 | 5.655 | 5.555 | 5.416 | 0.828 | 0.754 | 0.710 | 0.682 | 0.646 | 0.593 | 0.528 | 0.491 |

<Table 8> MSE of Iris SL data

| | | 5% | 10% | 15% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|---|
| MCAR | REG | 0.004 | 0.010 | 0.014 | 0.020 | 0.031 | 0.041 | 0.052 |
| | MEAN | 0.032 | 0.069 | 0.102 | 0.141 | 0.210 | 0.280 | 0.350 |
| | Cohen | 0.048 | 0.098 | 0.142 | 0.198 | 0.320 | 0.469 | 0.668 |
| | A-Cohen6 | 0.007 | 0.016 | 0.024 | 0.033 | 0.050 | 0.068 | 0.091 |
| MAR | REG | 0.009 | 0.011 | 0.013 | 0.017 | 0.023 | 0.915 | 0.996 |
| | MEAN | 0.045 | 0.133 | 0.202 | 0.287 | 0.493 | 0.743 | 0.914 |
| | Cohen | 0.081 | 0.270 | 0.382 | 0.514 | 0.760 | 1.042 | 1.188 |
| | A-Cohen1 | 0.017 | 0.018 | 0.023 | 0.030 | 0.043 | 0.093 | 0.133 |
| NMAR | REG | 0.013 | 0.030 | 0.052 | 0.071 | 0.127 | 0.197 | 0.289 |
| | MEAN | 0.173 | 0.298 | 0.386 | 0.489 | 0.668 | 0.872 | 1.080 |
| | Cohen | 0.262 | 0.473 | 0.608 | 0.766 | 1.062 | 1.321 | 1.601 |
| | A-Cohen1 | 0.035 | 0.060 | 0.078 | 0.105 | 0.185 | 0.314 | 0.511 |

case, adjusted Cohen, regression and MCMC methods hold true values up to 50% missing within a 5% level. In NMAR, regression and MCMC methods are acceptable up to 50% missing while mean and Cohen methods are valid in 15% missing. The adjusted Cohen method is good up to 40% missing.

Under MCAR, the outcome of standard deviation is fine for regression, adjusted Cohen and MCMC methods. Mean imputation underestimates standard deviation while Cohen method overestimates it. In MAR, Cohen method is considered as the best, and the adjusted Cohen method is good up to 20%missing, also MCMC and regression methods are fine up to 20% missing. In NMAR, only regression and MCMC methods preserve the variance up to 5% missing.

The feature of MSE is that it inflates as the missing rate goes up. Under all missing mechanisms, the Cohen method has the worst outcome while regression method has the best, except in the MAR case. At 40% and 50% missing, MSE of regression method inflates suddenly, but for the adjusted Cohen method, it increases gradually in MAR.

(2) The results of Iris Petal length data

Table 9 and Table 10 indicates the mean, standard deviation and MSE of Iris petal length data. Under MCAR, all methods estimate the true value within 1% level up to 50% missing like the case of Iris sepal length data. In MAR, at 5% missing, all methods are fine, but mean and Cohen method are not recommended, with over 5% missing. The adjusted Cohen method has a reasonable value up to 20% missing, with reg, up to 30% missing and with MCMC, up to 40% missing. In NMAR case, MCMC has excellent accuracy up to 40% missing, as does the regression method. The adjusted Cohen method also can be attractive up to 30% missing.

Under MCAR, there is almost no change in standard deviation with regression, MCMC method and adjusted Cohen method, while the mean method

underestimates it and the Cohen method overestimates it. In MAR, mean method has the worst result, at the 50% missing. It has a 70% loss in standard deviation. MCMC and regression are good up to 30% missing. Cohen method and adjusted Cohen method are fine up to 15% missing. Under NMAR, Cohen method is valid up to 40% missing, the MCMC up to 20% missing, regression and the adjusted Cohen up to 15% missing.

The adjusted Cohen method remarkably lowers MSE, at least 40%, in comparison with the Cohen method. Regression method has the best outcome.

<Table 9> Mean and SD of Iris SL data

| | | Mean | | | | | | | | SD | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0% | 5% | 10% | 15% | 20% | 30% | 40% | 50% | 0% | 5% | 10% | 15% | 20% | 30% | 40% | 50% |
| MCAR | MCMC | 3.758 | 3.758 | 3.756 | 3.757 | 3.758 | 3.756 | 3.761 | 3.759 | 1.765 | 1.766 | 1.769 | 1.772 | 1.773 | 1.782 | 1.791 | 1.808 |
| | Reg | 3.758 | 3.749 | 3.758 | 3.765 | 3.770 | 3.764 | 3.746 | 3.788 | 1.765 | 1.756 | 1.759 | 1.764 | 1.759 | 1.754 | 1.738 | 1.766 |
| | Mean | 3.758 | 3.757 | 3.757 | 3.756 | 3.760 | 3.757 | 3.763 | 3.768 | 1.765 | 1.724 | 1.673 | 1.629 | 1.578 | 1.472 | 1.361 | 1.243 |
| | Cohen | 3.758 | 3.741 | 3.740 | 3.756 | 3.760 | 3.738 | 3.763 | 3.747 | 1.765 | 1.808 | 1.860 | 1.910 | 1.974 | 2.106 | 2.273 | 2.494 |
| | A-Cohen6 | 3.758 | 3.757 | 3.757 | 3.757 | 3.758 | 3.756 | 3.760 | 3.762 | 1.765 | 1.753 | 1.740 | 1.732 | 1.726 | 1.721 | 1.737 | 1.783 |
| MAR | MCMC | 3.758 | 3.752 | 3.748 | 3.736 | 3.740 | 3.732 | 3.932 | 3.991 | 1.765 | 1.774 | 1.782 | 1.799 | 1.796 | 1.815 | 1.583 | 1.518 |
| | Reg | 3.758 | 3.754 | 3.752 | 3.743 | 3.740 | 3.765 | 4.601 | 4.624 | 1.765 | 1.769 | 1.772 | 1.784 | 1.789 | 1.758 | 0.803 | 0.784 |
| | Mean | 3.758 | 3.874 | 4.019 | 4.155 | 4.340 | 4.751 | 5.050 | 5.221 | 1.765 | 1.684 | 1.579 | 1.478 | 1.321 | 0.888 | 0.565 | 0.472 |
| | Cohen | 3.758 | 3.857 | 4.003 | 4.155 | 4.340 | 4.740 | 5.050 | 5.213 | 1.765 | 1.767 | 1.756 | 1.733 | 1.653 | 1.271 | 0.945 | 0.948 |
| | A-Cohen1 | 3.758 | 3.759 | 3.777 | 3.809 | 3.897 | 4.254 | 4.572 | 4.641 | 1.765 | 1.763 | 1.739 | 1.698 | 1.592 | 1.170 | 0.814 | 0.749 |
| NMAR | MCMC | 3.758 | 3.743 | 3.741 | 3.739 | 3.739 | 3.830 | 3.926 | 4.007 | 1.765 | 1.745 | 1.747 | 1.746 | 1.752 | 1.860 | 1.960 | 2.043 |
| | Reg | 3.758 | 3.733 | 3.710 | 3.690 | 3.663 | 3.659 | 3.611 | 3.448 | 1.765 | 1.728 | 1.701 | 1.679 | 1.652 | 1.654 | 1.612 | 1.468 |
| | Mean | 3.758 | 3.622 | 3.484 | 3.364 | 3.221 | 2.950 | 2.640 | 2.269 | 1.765 | 1.653 | 1.557 | 1.479 | 1.387 | 1.228 | 1.049 | 0.827 |
| | Cohen | 3.758 | 3.606 | 3.468 | 3.364 | 3.221 | 2.935 | 2.640 | 2.255 | 1.765 | 1.734 | 1.731 | 1.734 | 1.735 | 1.757 | 1.752 | 1.660 |
| | A-Cohen1 | 3.758 | 3.735 | 3.722 | 3.710 | 3.686 | 3.636 | 3.495 | 3.148 | 1.765 | 1.731 | 1.714 | 1.699 | 1.671 | 1.616 | 1.527 | 1.407 |

<Table 10> MSE of Iris SL data

| | | 5% | 10% | 15% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|---|
| MCAR | REG | 0.003 | 0.007 | 0.010 | 0.014 | 0.022 | 0.030 | 0.038 |
| | MEAN | 0.144 | 0.316 | 0.463 | 0.630 | 0.955 | 1.276 | 1.589 |
| | Cohen | 0.174 | 0.318 | 0.269 | 0.248 | 0.383 | 1.469 | 2.163 |
| | A-Cohen6 | 0.030 | 0.065 | 0.095 | 0.130 | 0.198 | 0.271 | 0.362 |
| MAR | REG | 0.001 | 0.005 | 0.009 | 0.014 | 0.020 | 2.083 | 2.158 |
| | MEAN | 0.291 | 0.688 | 1.089 | 1.711 | 3.319 | 4.476 | 5.048 |
| | Cohen | 0.492 | 1.175 | 1.877 | 2.677 | 4.085 | 4.675 | 4.671 |
| | A-Cohen1 | 0.001 | 0.005 | 0.023 | 0.103 | 0.836 | 1.946 | 2.309 |
| NMAR | REG | 0.015 | 0.028 | 0.041 | 0.066 | 0.083 | 0.115 | 0.261 |
| | MEAN | 0.399 | 0.765 | 1.082 | 1.481 | 2.263 | 3.274 | 4.662 |
| | Cohen | 0.820 | 1.554 | 2.118 | 2.865 | 4.467 | 6.078 | 7.859 |
| | A-Cohen1 | 0.014 | 0.024 | 0.037 | 0.059 | 0.125 | 0.387 | 1.265 |

(3) The results of generated data

Table 11 and Table 12 presents the mean of generated data. Under MCAR, the result is similar as the previous one. All methods are fine up to 50% missing. In

MAR, adjusted Cohen, regression and MCMC methods estimate true mean up to 50% missing. In the NMAR case, MCMC has the best effect up to 50% missing. Regression method and adjusted Cohen methods are fine up to 40% missing respectively.

<Table 11> Mean and SD of generated data

| | | Mean | | | | | | | | SD | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0% | 5% | 10% | 15% | 20% | 30% | 40% | 50% | 0% | 5% | 10% | 15% | 20% | 30% | 40% | 50% |
| MCAR | MCMC | 11.044 | 11.043 | 11.040 | 11.047 | 11.049 | 11.038 | 11.045 | 11.046 | 3.916 | 3.918 | 3.918 | 3.919 | 3.920 | 3.918 | 3.918 | 3.927 |
| | Reg | 11.044 | 11.044 | 11.044 | 11.044 | 11.044 | 11.044 | 11.043 | 11.045 | 3.916 | 3.910 | 3.904 | 3.896 | 3.890 | 3.875 | 3.861 | 3.850 |
| | Mean | 11.044 | 11.043 | 11.043 | 11.048 | 11.044 | 11.044 | 11.041 | 11.050 | 3.916 | 3.816 | 3.716 | 3.608 | 3.503 | 3.272 | 3.209 | 2.768 |
| | Cohen | 11.044 | 11.043 | 11.043 | 11.048 | 11.048 | 11.044 | 11.041 | 11.050 | 3.916 | 4.017 | 4.129 | 4.246 | 4.379 | 4.675 | 5.050 | 5.539 |
| | A-Cohen6 | 11.044 | 11.043 | 11.044 | 11.045 | 11.045 | 11.043 | 11.042 | 11.047 | 3.916 | 3.888 | 3.865 | 3.843 | 3.832 | 3.823 | 3.858 | 3.963 |
| MAR | MCMC | 11.044 | 11.046 | 11.044 | 11.019 | 11.034 | 11.038 | 11.058 | 11.031 | 3.916 | 3.920 | 3.919 | 3.935 | 3.908 | 3.909 | 3.903 | 3.940 |
| | Reg | 11.044 | 11.050 | 11.050 | 11.020 | 11.038 | 11.050 | 11.072 | 11.041 | 3.916 | 3.907 | 3.899 | 3.912 | 3.879 | 3.853 | 3.830 | 3.838 |
| | Mean | 11.044 | 11.355 | 11.609 | 11.827 | 12.080 | 12.528 | 13.009 | 13.412 | 3.916 | 3.638 | 3.431 | 3.272 | 3.075 | 2.771 | 2.470 | 2.270 |
| | Cohen | 11.044 | 11.355 | 11.609 | 11.827 | 12.080 | 12.528 | 13.009 | 13.412 | 3.916 | 3.829 | 3.812 | 3.850 | 3.844 | 3.960 | 4.119 | 4.543 |
| | A-Cohen1 | 11.044 | 11.087 | 11.084 | 11.042 | 11.049 | 10.980 | 10.926 | 10.631 | 3.916 | 3.820 | 3.776 | 3.769 | 3.703 | 3.644 | 3.552 | 3.591 |
| NMAR | MCMC | 11.044 | 11.013 | 10.978 | 10.951 | 10.916 | 10.812 | 10.688 | 10.493 | 3.916 | 3.861 | 3.802 | 3.798 | 3.769 | 3.691 | 3.610 | 3.491 |
| | Reg | 11.044 | 11.012 | 10.977 | 10.939 | 10.902 | 10.801 | 10.668 | 10.453 | 3.916 | 3.851 | 3.802 | 3.757 | 3.722 | 3.633 | 3.533 | 3.383 |
| | Mean | 11.044 | 10.601 | 10.257 | 9.954 | 9.667 | 9.109 | 8.533 | 7.938 | 3.916 | 3.385 | 3.068 | 2.829 | 2.622 | 2.263 | 1.928 | 1.627 |
| | Cohen | 11.044 | 10.601 | 10.257 | 9.954 | 9.667 | 9.109 | 8.533 | 7.938 | 3.916 | 3.563 | 3.409 | 3.328 | 3.279 | 3.234 | 3.215 | 3.256 |
| | A-Cohen1 | 11.044 | 10.850 | 10.708 | 10.597 | 10.415 | 10.062 | 9.614 | 8.982 | 3.916 | 3.555 | 3.379 | 3.265 | 3.192 | 3.091 | 3.027 | 3.084 |

<Table 12> MSE of generated data

| | | 5% | 10% | 15% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|---|
| MCAR | REG | 0.051 | 0.104 | 0.155 | 0.207 | 0.309 | 0.416 | 0.522 |
| | MEAN | 0.773 | 1.532 | 2.318 | 3.077 | 4.638 | 6.174 | 7.687 |
| | Cohen | 2.387 | 4.843 | 7.448 | 10.101 | 15.985 | 22.777 | 31.064 |
| | A-Cohen6 | 0.374 | 0.748 | 1.134 | 1.517 | 2.344 | 3.247 | 4.298 |
| MAR | REG | 0.038 | 0.086 | 0.151 | 0.212 | 0.298 | 0.384 | 0.460 |
| | MEAN | 2.201 | 3.889 | 5.246 | 6.959 | 9.865 | 13.104 | 15.800 |
| | Cohen | 3.471 | 5.795 | 7.919 | 10.339 | 14.842 | 18.975 | 23.551 |
| | A-Cohen1 | 0.308 | 0.708 | 1.152 | 1.586 | 2.529 | 3.473 | 4.913 |
| NMAR | REG | 0.071 | 0.132 | 0.203 | 0.283 | 0.465 | 0.710 | 1.135 |
| | MEAN | 4.076 | 6.543 | 8.527 | 10.360 | 13.965 | 17.936 | 22.353 |
| | Cohen | 5.901 | 10.005 | 13.675 | 17.001 | 23.514 | 30.226 | 37.720 |
| | A-Cohen1 | 0.905 | 1.577 | 2.153 | 3.478 | 6.087 | 9.426 | 14.292 |

Under MCAR, regression and MCMC method and adjusted Cohen method offer available values up to 50% missing. In MAR, MCMC and regression are fine up to 50% missing Cohen is valid up to 30% missing and the adjusted Cohen is good up to 15% missing. Under NMAR, only MCMC and regression methods offer suitable values up to 20% missing.

Regression method has the best result while Cohen has the worst. Adjusted Cohen method offers better results than outcome of the mean method.

We can verify that the adjusted Cohen methods are effective to strengthen the drawbacks of Cohen method.

## 5. Conclusions and Future Directions

The best way to deal with missing data is to prevent it, however it happens anyway. After missing data has occurred, it is possible to impute the missing data with predicted or simulated values.

In this study, the adjusted Cohen methods have been suggested and simulation has been done to compare the efficiency with existing imputation methods include Cohen method.

The simulation was performed using real data and generated data with 7 types of missing rate comparing mean, standard deviation of the variable after imputation and MSE, under MCAR, MAR and NMAR assumptions.

The result showed multiple imputation using MCMC offered the best outcome, especially because it had valid estimates under MAR and NMAR. However this method took a long time to tabulate the results, since it creates numbers of complete data set and analyze them. Among single imputation methods, regression imputation showed the best results and gave reasonable values in some of NMAR cases, but it needs a model and becomes difficult in multivariate data when more than one variable has missing values. Mean imputation gave the worst effect to impute missing values, it could be used to estimate means under only MCAR assumption, but these are affected in other cases, besides the variance was underestimated in all cases. The Cohen methods presented good estimator with variances estimation in some cases of MAR and NMAR, but it tended to overestimate variance under MCAR as two values that dropped from mean to opposite sides were imputed. That reason brought about inflation of MSE. The adjusted Cohen method produced valid mean and variance under MCAR and lowered MSE compared with the Cohen method and mean imputation. Also it provided fine values in mean estimation under some of MAR and NMAR, but the variance estimation was not good under NMAR.

The following some suggestions are given through a simulation study.

1. Do not use mean imputation unless the data is MCAR and variance estimation does not matter, namely, only means or totals are required.
2. The adjusted Cohen method6 for MCAR and the adjusted Cohen method1 for MAR and NMAR are recommend.

In this study, we have done a simulation study with three kinds of data, but more simulation is needed to compare the efficiency of the adjusted Cohen

methods and the other methods. In addition, the adjusted Cohen method is not valid in variance estimation under NMAR, so a study under NMAR is required.

## References

1. Allison, P. D.(2000), Multiple Imputation for Missing Data : A Cautionary Tale, *Sociological Methods and Research*, 28, 301-309.
2. Cohen, M. P.(1996), A new approach to imputation, *American Statistical Association Proceedings of the Section on Survey Research Methods*, 293-298.
3. Horton, N. J. and Lipsitz, S. R.(2001), Statistical Computing Software reviews, *The American Statistician,* 55, 244-254.
4. Little, R. J. A. and Rubin, D .B.(2002), *Statistical Analysis with Missing Data*, Second Edition, New York : Wiley.
5. Rubin, D. B.(1976), Inference and Missing Data, *Biometrika*, 63, 581-592.
6. Rubin, D. B.(1987), *Multiple imputation for nonresponse in surveys,* New York : Wiley.
7. Scheffer, J.(2002), Dealing with missing data, *Research Letters Information Mathematical Sciences,* 3, 153-160.

[ received date : May. 2006, accepted date : Jul. 2006 ]