Missing Data Imputation in Social Science Research:
A Simple Model Simulation
2006

Assoc. Prof. Montree Piriyakul, Ph.D.

Department of Statistics, Faculty of Science, Ramkhamheang University, Bangkok, 10240, THAILAND Tel (66)23108398 Fax (66)23108386
mpiriyakul@yahoo.com

Abstract. This study is aimed at evaluation and comparison of imputation behavior, under missing at random (MAR) pattern, among 18 different missing data techniques (MDT) where various designs were planned - i.e. combination of 4 m_c (missing case of 5 %, 10 %, 15 % and 20 %), 3 m_i (missing item of 10 %, 20 % and 30 %) and 3 sample size (n = 100, 200, 500) with one hundred replications each. In addition, 5 levels of multi-collinearity were added if Monte Carlo experiment was investigated. An empirical analysis depended its data on file of National Health and Welfare Survey conducted by NSO, Thailand in 2001.There were 18,000 samples employed in simulation study while 3,600 samples did in empirical analysis.

Experiments in Monte Carlo fashion with simulated Likert type data and in empirical analysis style with real nominal data were processed separately whereas average values of MSE, bias, Pearson product moment and Cronbach's α were extracted and plotted. From 3 of these 4 evidences, except Cronbach's α which discriminated nothing. There were 2 groups of MDT's judged to be optimal with no obligations of sample size, percentage of missing cases, percentage of missing items and multicollinearity level; they are IMS, HDD-IMS and HDD-IMS-Z for real nominal data and PMS, HDD-PMS and HDD-PMS-Z for simulated LIKERT scaled data. Among optimal MDT's in either group, they can be chosen to be used arbitrarily since the different among their average value of MSE or bias or correlation coefficient are subjectively small.

However, for ease of use and less computation arrangement, IMS and PMS are recommended, but for more accurate and general usage, HDD-IMS and HDD-IMS-Z or HDD-PMS and HDD-PMS-Z are suggested.

Keyword : Monte Carlo, Empirical analysis, MDT, Hot-Deck

1. Introduction

Researchers frequently face problems of missing data (Heeringa, 2000; DeSilvio, 1999; Wesenbaer, 1998; Samuhel, 1983). Even if surveys or experiments are carefully controlled, such problems are inevitable (Huisman, 1999). Researchers in the social sciences have increasingly paid attention to these problems (Adam, 2001). If the analysis of data is conducted using univariate statistics, i.e., percentage, mean or other techniques of descriptive statistics, consequences need not be severe. However, if the data analysis employs techniques of multivariate analysis such as when multiple regression analysis, path analysis, factor analysis, cluster analysis and discriminant analysis are utilized, problems of missing data have severe effects. If any unit of analysis has a variable with missing data, that unit of analysis has to be deleted regardless of whether other

variables have complete data or not (Heeringa, 2000; Roth, 1994). Listwise deletion is a default of multivariate analysis in the SPSS program or other statistical programs. Experimental study indicates that if each variable has a randomly selected ten percent of its data missing, the unit of analysis must be deleted at a rate of 59% (Kim & Curry, 1977 quoted in Roth, 1995). The loss is therefore at a high rate. Data analysis on the basis of the remaining data after the incomplete observation values have been deleted will yield a biased and inaccurate analytic result (Wang, 2000).

Failure to respond in the case of surveys is mainly a social factor (Ruben, 1992) and concerns personality and psychology. Although there are many statistical mechanisms available to decreases rate of failures to respond in a survey, failures to respond to questions still appear and affect results. The failure to respond may be because in such a social exchange context that the respondent may not receive any material or psychological benefit in exchange for giving information. This state of affairs also sometimes concerns symbolic interaction (Murata, 2001). Some words, phrases, sentences or other means of communication used in the question-and-answer process do not convey clear messages or do not convey the same message to all respondents in the same context. The questions asked may be too difficult or the layout of questions is inappropriate, etc. All such factors affect interaction which results in missing data which in turn leads to lost and insufficient information necessary to the derivation of conclusions or results. Imputation of missing data is necessary to replace missing data in order to make the findings of the study more reliable, clearer, complete, accurate and unbiased.

Although an estimate of missing data is a pseudo-value, it can substitute for the actual value. In fact, respondents may not give consistent information. It may not be believed that they had such an opinion or that the information given would always be the same. When a period of time has passed, respondents may change their minds or change their answers. In a study of asking the same unanswered questions or the same questions previously answered, it was found that 95% of the respondents changed their answers (Huisman, Kron, & Van Sonderen, 1998). Since actual answers of respondents could be so readily changed, it is as if such answers would of themselves supply imputed values. An argument that imputed value is only an estimate and not the actual value has to take this state of affairs into consideration. Although the estimate of observed values in terms of imputed values are not actual values, these values were still estimates based on repeated experiments which obtain results at a satisfactorily high degree of accuracy. A model of missing data imputation with appropriate missing values should be acceptable. It should therefore be taken into consideration that although imputed values are estimate values, this state of affairs is still better than deleting data because of incomplete observation and would be less damaging. Otherwise, researchers would have to conduct a panel study.

Failures to respond lead to missing items and to analyses of data with incomplete observation, a state of affairs leading to both theoretical and practical difficulties.

1. If missing data do not have values close to the actual data pertaining to the same variable or have totally different values, data analysis based on actual data will yield results that are inaccurate and will lead to erroneous findings. When there are missing data, it is found that statistical measures of central tendency will exhibit levels of bias (DeSilvio, 1999). That is to say, the findings are empirically invalid (Rearden, 1991; Kim, 2000; DeSilvio, 1999). Missing data replacement with some appropriate imputed values will reduce bias (Kwang, 2000). A biased mean is a mean that is indicative of a definite tendency, but it is not an actual value. Statistical analysis in such cases is, therefore, not as precise as would be expected and is inaccurate (Joen, 1998; Ibrahim, 1998; DeSilvio, 1999; Wang, 2000).

In addition, missing data will result in inaccurate measurements of variance and standard deviation. They may be lower or higher than usual depending on the location of the missing data (Roth, 1995).

Determinations of variance and standard deviation greatly influence research in the social sciences. Low measurements of standard deviation indicate that there is little variety in the data with few differences. For example, if the standard deviation in a statistical analysis of income is low, it means that the targeted population has no differences in their economic status and few differences regarding levels of wealth and poverty. Such a society, therefore, should be happy and warm because there are no problems regarding the status of the people. Such a conclusion could be drawn as such because some data are missing. In fact, the conclusion may well be different if the data were complete.

Mistakes regarding the mean and variance are not the only results of the problem of missing data. This problem also results in mistakes being made in applications of t-statistics and F-statistics, as well as in applying correlation (r) techniques. Moreover, the statistical results of measuring the quality of a scale may be so wrong that we are no longer confident that the measures already used could actually measure what we aim to measure.

2. Good estimates depend on sufficient data governing variables. They should encapsulate as great a variety as possible. Missing data results in the loss of necessary information. Estimates convey insufficient information which will be inappropriate to the actual state of affairs regarding what is being estimated.

3. Statistics used in an analysis presuppose using an array of complete data in accordance with predetermined sample size. If data were missing, the sample would only be partially random. Statistics appropriate for such cases must be changed.

4. Missing data will decrease the reliability of a statistical test. (Witta, 2000; Roth, 1994).

For the most part, the larger the sample, the more powerful the test being used. Experiments indicate that if a variable being analyzed is based on data that is missing randomly to the extent of 2 % , it will be necessary to delete 18.3% of cases from the whole , if listwise deletion was employed. Thus it can be seen that even a small amount of lost data resulting from a failure of respondents answering questions or a failure to obtain certain information can have severe consequences for an investigation. For example, suppose 370 sample items were designated as sufficient for a reliable estimate. However, if 2% of the required data were missing, this would result in the deletion of 68 cases (Roth, 1995).

5. Missing data results in mistakes concerning the correlation value (r) that is calculated on the basis of the group sampled. Mistakes will be made by virtue of specifying a correlation lower than it should have been or there will be a downward bias. The major cause of this is the missing data which results in differences in variance of the variable being considered. The level of correlation with other variables changes as well (Kim & Curry, 1997; Mahotra, 1987 quoted in Roth, 1994).

6. The measurement of the quality of scale, i.e. a questionnaire and other types of evaluation forms, needs to have measures that are accurate. They should be valid, and have reliability. Missing data result in bias both in regard to the mean and standard deviation which in turn affects test reliability. The consequence is lost test credibility.

There are many ways to solve missing data problems. The best way should be the simplest way in order not to create undue difficulties for researchers without expertise in statistics.

This research aims to evaluate the performance of missing data techniques (MDT)

in a situation in which data are missing at random (MAR). MAR is a missing pattern that the missing data involve demographical characteristics, i.e., occupation, income, and the respondents' level of education. Experiments are conducted for the purposes of a comparative study using eighteen methods with actual data and simulated data. The limitations of the study are only in the use of data used to provide measurements in the ordinal scale in accordance with Likert's scale and the nominal scale and the missing data occurs only in regard to dependent variable.

2. Missing data imputation

The model proposed here is partly a way of imputing the data by replacing the missing data with the mean (Downey & King, 1998, Huisman, Krol & Sonderen, 1998). The other methods are hot-deck and ICS.

The hot-deck method-also known as the neighbor-next door, the nearest neighbor, closest fit, or nearby household method is the imputation of the missing data with the data taken from the donor. This can be done by the method of categorization or Euclidian distance which is the method employed by the US Census Bureau, the Canadian Census Bureau and the UK Census Bureau in imputing missing data (Roth & Switzer, 1995; Paullin & Ferraro, 1994). The eighteen methods for the imputation of missing data are as follows:

1. Random Draw Substitution (RDS) is the method of imputing the missing data by randomly selecting data from the possible complete data list. This method is found to be the lowest in quality (Huisman, 1997).

2. Hot-deck Next Case (HNC) is the method of imputing the missing data by using the data from the first complete case next in order. This method is low in quality, but in general, its quality is higher than that of RDS (Huisman, 1997).

3. Item Mean Substitution (IMS) is the method of imputing missing data by using the mean of the existing data of the same item (variable).

4. Pearson Mean Substitution (PMS) is the methods of imputing the missing data by replacing them with mean of the other complete items of the same case (person) itself. This method is found to be higher in quality than IMS (Huisman, 1997).

5. Corrected Item Mean Substitution (CIMS) is the method that uses the effect from an individual and the effect from the question as counterweights for imputing the missing data. The effect from an individual is the total number of an individual's responses. If any respondent completes more questions, the weight of that individual may be higher than that of the individual responding to fewer questions. Any variable having less data missing bears more weight than any variable with more data missing. In such a case, CIMS is found to be the best method (Huisman, 1997).

$$CIM_{vi} = w_v \bar{x}_{.i}^{(i)} = \left(\frac{\sum_{i \varepsilon obs\,(v)} x_{vi}}{\sum_{i \varepsilon obs\,(v)} \bar{x}_{.i}^{(i)}} \right) \bar{x}_{.i}^{(i)}$$

; v = 1, 2, …, mi

$\bar{x}_{.i}^{(i)}$ is the mean of variable i calculated from the data of respondents for question i. The symbol (i) is to indicate that the interest is in respondents to question i.

obs (v) designates the responses of respondents v to the questions requiring responses.

$i \in$ obs (v) designates interest only in the responses to the questions to which there was a response. Questions for which there were no responses are not considered.

6. Item Correlation Substitution (ICS) is the method used in finding correlations

between variables using the method of pairwise deletion. On the basis of the correlation value the variable under consideration can be seen to correlate with at most any other variable. When found, take the responses of the complete case to replace the variables with which there is correlation even if there is also missing data. If the respondent does not answer both questions, use the responses based on the variable that correlates with the variables under consideration in descending order.

The consideration of any variable or any pair of variables can be undertaken in the study of opinion at a group level. If any pair of variables is correlated at a high positive level, it means that there is a positive relationship between opinions expressed by the group regarding both issues. Therefore, the data from paired variables might be used to replace one another.

7. **Hot-Deck Deterministic** (HDD) is the calculation of distance between an incomplete case and a complete case using the pairwise deletion method. In such a case, the data from the complete case (with responses for all questions) that correlates with the incomplete case at the highest level (the shortest distance) can be used as imputed data. If the distance between an incomplete case and a complete case has the same value, use the data from the closer case will be chosen as imputed data. It is believed that the cases with numbers close to one another are cases in the same area or same subgroup of the population or same geographical area. Therefore, they should have more or less similar characteristics than the case further away.

8. **The Hot-Deck Random** (HDR) method is similar to the HDD method. A minor difference is that when it is found that the distance between an incomplete case and many complete cases revealed the same value, select donor by an application of the random method, i.e. use the data from that randomly chosen donor as the imputed data source for the missing item of the incomplete case. Reasons behind this algorithm are what we believed that the cases are in the same area or have a similar context or have a similar environment it is possible to have similar features. So, the selection of any donor can be chosen randomly from any complete case.

9. **Hot-deck Deterministic-Item Mean Substitution** (HDD-IMS) is a mixed method that replaces missing data with the mean of the variable in order to have complete observation values in all cases first. Then the distance between cases is calculated and the decision is made to use the donor in accordance with the HDD method. Advantages of this method is that all cases have complete information prior to the calculation of distance which results in more distance values calculated from the information than when the HDD method is used.

10. **Hot-Deck Deterministic-Person Mean Substitution** (HDD-PMS) is a mixed method in which the missing data is replaced by the mean of that case in order to have complete observation values in all cases prior to the calculation of distance. This method is only different from the HDD-IMS method in the stage of replacing data.

11. **Hot-Deck Random-Item Mean Substitution** (HDR-IMS) is a mixed method that replaces the missing data of any case with the mean of that variable. And hence, the distance between cases is calculated. Then the donor is selected to provide data to the case specified to receive such data. The donor is selected from the case at the shortest distance from the case receiving data. If any case has shortest distance with many donors, choose one at random.

12. **Hot-Deck Random-Person Mean Substitution** (HDR-PMS) is a mixed method that replaces the missing data of any case with the mean of the data from that case. Then the distance between cases is calculated. The donor is then selected from the distance calculated. In case that any case receiving data are at the shortest distance from many donors, randomly chosen the donor is recommended.

13. **Standardized HDD** (HDD-Z) is the HDD method that works when the data of all variables are transformed into a standard score (z). The transformation of data into a standard score aims to adjust the value of variables measured with different scales to be data using the same standard $\left(Z = \frac{x - \bar{x}}{SD}\right)$ and after the use of the HDD method, the data is transformed back to previous scale by $Z = \bar{x} + SD*Z$.

14. **Standardized HDR** (HDR-Z) is the HDR method that works when the data of all variables are transformed into a standard score. After the use of the HDR method, the data are transformed back to the original scale.

15. **Standardized HDD-IMS** (HDD-IMS-Z) is a mixed method that selects the case that provides data and replaces the missing data using the HDD-IMS method. All data in all cases must be transformed into a standard score first. After the replacement of the data, all data are transformed back to the original scale.

16. **Standardized HDD-PMS** (HDD-PMS-Z) is a mixed method used to select the case that provides data and replaces the missing data using the HDD-PMS method. The data in all lists must be first transformed into a standard score. After the replacement of the data, the data must be transformed back to the original scale.

17. **Standardized HDR-IMS** (HDR-IMS-Z) is a mixed method that selects the donor and replaces the missing data in accordance with the HDR-IMS method. The data in every case must be first transformed into a standard score. Then after the replacement of the data, convert the data back to the original scale.

18. **Standardized HDR-PMS** (HDR-PMS-Z) is a mixed method that selects the donor and replaces the missing data by using the HDD-IMS method. The data in all cases must be converted into a standard score and converted back to their original scale after they have been replaced.

3. Methods of analysis

This research divided the experiments and accompanying analysis into two types.

1. **An empirical experiment** is an experiment with actual data using information and existing data as a tool.

1) Findings indicate that gender, age, education, income, and occupation affect the rate of responses which in turn affects the phenomenon of missing data. It was found that females, those educated at a low level, the unhealthy, and the aged had a high rate of non-response (DeLeeuw, 2001; Huisman, & van de Zouwen, 1998; Huisman, et al., 1998). Females living alone had a high rate of non-response (Rucker, 1990 quoted in Couper & Groves, 1996). It was, however, also found that age may or may not affect the rate of responses or non-responses. In other words, the aged may be willing or unwilling to cooperate depending on the subject on which they are questioned and on the health of the respondents (Couper & Groves, 1998). At the same time, it was also found that females were more cooperative (Couper & Groves, 1998). And, as expected, people with low incomes and unstable occupations had a high rate of non-responses (Demaio, 1980 quoted in Rylander et al., 1995).

Because of all of these factors, this research designated gender, age, education, income and occupation as independent variables both in the experiment using actual data and in the experiment using simulated data. In the case of using simulated data, dependent variables were designated to be variables in the ordinal scale. In the case of the actual data, dependent variables were designated to be variables in the nominal scale.

2) Sample size was designated at three levels: 100, 200, and 500. The sample size cannot be smaller than this because a large sample size would make the replacement

of data require more donors in hot deck method. Therefore, the findings would be more accurate (Roth & Switzer, 1995).

3) From the data file with all units of analysis are complete cases, the missing case was designated to be incorporated at m_c %. Each case was designated to have the missing item at m_i %.

The decision to determine which case is the missing case was designated by using logistic regression.

$$\Pr(\text{missing case}) = p = \frac{1}{1+e^{-z}}$$

$$\text{While } Z = 2.5 + 1 * SEX - 1 * AGE - 1 * EDU - 1 * INC + 1 * OCC$$

The above equation governs results in which there is a higher level of failure to respond by females, the aged, those with low levels of education and income or unstable occupations as compared with other groups under examination (Huisman, 1997). The findings from the model indicate that any case yielding $p \geq 0.5$ is the missing case. Then it is stipulated that the data of variables covering such a contingently missing case is ascertained by the random method in regard to the sum of m_i % by replacing with 0 or any symbol such as "-" with $m_c = 5, 10, 15, 20$, and $m_i = 10, 20, 30$, respectively.

4) Impute the data for variables with missing value temporarily using eighteen methods of data replacement. The data file used was "A Survey of Health and Welfare A.D. 2001" conducted by the Office of the National Statistics involving 118,285 cases of which only 83,641 complete 15-item questionnaire cases were actually selected for the experiment.

5) Experiments were conducted in 100 replications for each combination of m_c = 5%, 10%, 15%, 20%, m_i = 10%, 20%, 30% and n = 100, 200, 500. Then the statistics from 100 samples was averaged in order to show the quality of each 18 data imputation method. There was 4 x 3 x 3 combinations with 100 repetitions each, totaled to 3,600 groups, applied into our imputation algorithm.

2. **A Monte Carlo experiment** is an experiment with simulated data using random numbers as a tool. Random numbers were generated in two phases such that the random number indicating the independent variables of gender, age, level of education, income, and occupation is designated to have a value in accordance with the probability value as found in table of frequency. The second phase, random number indicating dependent variables as measured in the Likert scale is postulated to exist through an application of the regression equation.

1) The construction of the value of independent variables. Independent variables indicating demographic characteristics are construed as a group variable consisting of five variables: gender, age, education, income, and occupation. The structure of these independent variables depends on the result of the designation of code and percentage taken from Thailand's 2000 census.

2) The construction of dependent variables. Dependent variables were designated to have values in accordance with the Likert scale as follows: V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, V_9, and V_{10}. Each variable had five codes.

Dependent variables depend there values on independent variables. In this research, dependent variables were designated to have values in accordance with the Likert scale, therefore designated to have values which varied in conformity to the variance of the independent variables. This means that they varied in accordance with internal relationships between independent variables, a state of affairs allowing the researcher to determine the level of correlation desired.

The process of constructing dependent variables is as follows.

Step 1. Designate five independent variables: gender, (1, 2), age (1, 2, 3, 4, 5), education (1, 2, 3), income (1, 2, 3, 4, 5), occupation (1, 2, 3, 4, 5).

Step 2. Independent variables in Step 1 were constructed using the following equations (Mc Donald, 1975; Gibbons, 1981; Wichem & Churchill, 1978).

$$X_{ij} = (1-\alpha^2)^{\frac{1}{2}} Z_{ij} + \alpha Z_{i6} \; ; \; j = 1, 2, 3 \quad ; i = 1, 2, ..., n$$

$$\text{and } X_{ij} = (1-\alpha_*^2)^{\frac{1}{2}} Z_{ij} + \alpha_* Z_{i6} \; ; \; j = 4, 5 \quad ; i = 1, 2, ..., n$$

Z_{ij} is a standard normal variable constructed through the use of Marsarglia and Bray algorithm (Wichern & Churchill, 1978). Additionally, $\alpha^2 = .99^2, .90^2, .70^2$ and $\alpha_*^2 = .99^2, .90^2, .30^2, .10^2$ were designated to be the correlation bond among independent variables, and recoded to values as follows:

X_1 = gender, code 1, 2
X_2 = age, code 1, 2, 3, 4, 5
X_3 = education, code 1, 2, 3
X_4 = income, code 1, 2, 3, 4, 5
X_5 = occupation, code 1, 2, 3, 4, 5

The recode mechanism that transformed values of independent variables to new code in accordance with what is required is designated by the calculation of the area under standard normal distribution or N (0, 1) as $\int_{-\infty}^{X_{ij}} f(z)dz$ and changed to code using probability values taken from the frequency table in accordance with the results of Thailand's 2000 census.

The correlation value of α^2 among independent variables X_1, X_2, and X_3, the correlation value of α_*^2 between X_4 and X_5 and the combined correlation value of $\alpha\alpha_*$ among X_1, X_2, X_3 with X_4 and X_5 indicated that the correlation matrix was $X'X$, size 5x5. The correlation value α^2 equaled $.99^2, .90^2, .70^2$ and α_*^2 equaled $.99^2, .90^2, .30^2$ and $.10^2$. The outcome allowed for the construction of a correlation matrix with which could be used to find eigen values and the multicollinearity index (called spectral condition number, $R_m = \lambda_{max}/\lambda_{min}$). If the multicollinearity index assumed higher value, independent variables have a high level of correlation. If the multicollinearity index has a low value, it is indicated that the independent variables have a low level of correlation. The study was divided into five cases ranging from the case at the highest level of multicollinearity to the case having the lowest level of multicollinearity. It was found that each case had five eigen values. Each eigen value always led to one eigen vector of size 5x1. The member of normalized eigen vectors agreeing with the lowest eigen value (λ_{min}) was used as the coefficient value $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ (Gibbons, 1981; Wichern & Churchill, 1978, Mc Donald & Garlarneau, 1975). This research used only normalized eigen vectors that agreeing with eigen value λ_{min} because it aims to study the case of risk in regard to the problem of the highest level of multicollinearity. If there is no damage done in this case, there will be no damage done in the case of risk in regard to the problem of multicollinearity at a lower level.

Dependent variables are attitudes toward any matter and have the value that can be construed as follows.

$$Y_i = \beta_{0s} + \beta_{1s}X_{i1} + \beta_{2s}X_{i2} + \beta_{3s}X_{i3} + \beta_{4s}X_{i4} + \beta_{5s}X_{i5} + e_i \text{; i = 1, 2, … , n}$$

It was stipulated that β_o equaled 0 and e_i was assumed N (0, 1) (Gibbons, 1981).

Dependent variables were coded as ordinal numbers 1, 2, 3, 4, 5 in accordance with the empirical rules of probability. Each code was designated to exist with equal probability. Codes 1, 2, 3, 4, 5 implicated the level of attitude expressed as most disagree, disagree, no opinion, agree, and most agree, respectively.

Step 3. Designating missing data. This uses the same method as the empirical experiment by deciding which case is the missing case using the probability value from a logistic regression equation given as follows.

$$\text{Pr (missing data)} = p = \frac{1}{1+e^{-z}}$$

$$Z = 2.5 + 1 * \text{SEX} - 1 * \text{AGE} - 1 * \text{EDU} - 1 * \text{INC} + 1 * \text{OCC}$$

Results obtained were that females, the aged, those with low levels of education and income, and those with unstable occupations had a higher rate of failure to respond than other groups. Any case with the value of $p \geq 0.5$ is considered the missing case.

The stipulation was as follows: of m_c = 5%, 10%, 15%, 20%, m_1 = 10%, 20%, 30%. The multicollinearity level, R_m of the variables was at five levels. The cases studied are critical cases of $\lambda_{\min}$ only. The sample size n equaled 100, 200, and 500 to illustrate the situation of a sample on a small, medium and large scale. In regard to replication, there were 100 replications for all conditions of m_c x m_1 x R_m x n, a total of 18,000 groups. Then the statistical results were averaged in order to illustrate the results of the experiment in each condition.

Statistics for Decision Making

The decision to see which method of data substitution is the most appropriate can be considered on the basis of the following four statistics (Roth & Stwitzer, 1995; Huisman, 1997).

1) $\text{MSE} = \frac{1}{n}\sum_{(residual)^2}$ with residual = actual value – substitution value; $\text{MSE} \geq 0$

2) $\text{Bias} = \frac{1}{n}\sqrt{\sum(residual)}$ with residual = actual value – substitution value; $\text{Bias} \geq 0$

3) $\alpha = \frac{k}{k-1}\frac{s^2 - \sum_i^k s_i^2}{s^2}$ where k = number of questions; $0 \leq \alpha \leq 1$, imputation method with higher average of α is more consistent with the actual value.

4) $r_{\text{old, new}} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$; $1 \leq r_{\text{old, new}} \leq 1$

x designates an original set of data and y is a new set of data with missing data have been imputed by specific imputation method. Any imputation of data that yields a higher r value indicates such a method yields the estimated value in a more consistent manner.

4. Conclusion

1. Statistics for making a decision of MSE, bias and r yields consistent results while α did not show any difference.

(1) In case of the empirical context, the best methods of data substitution (MDT)

are IMS, HDD-IMS, and HDD-IMS-Z.

(2) In case of simulated data, the best methods of data substitution (MDT) are PMS, HDD-PMS and HDD-PMS-Z.

The reason for the incompatible results of this study may be because the two experiments used data with different scales, especially in the number of choices. The Monte Carlo experiment used the 5-point Likert scale, whereas the empirical analysis used the nominal scale with 2 to 11 choices. The differences between scales may not have any effect because, regardless of the scale used, the values of variables were pre-supposed as code designated by the researcher and could be designated to have any values. However, the unequal possible values may have certain effects.

2. The best methods for data imputation were IMS, HDD-IMS, and HDD-IMS-Z for the case that measures in accordance with the nominal scale. The PMS, HDD-PMS, and HDD-PMS-Z methods were the best for the case that used the Likert scale, with no regard to sample size (n), percentage of missing case (m_c), percentage of missing item (m_1), and the multicollinearity level.

3. The findings of this study agreed with the findings of Huisman (1997) which found that the IMS and PMS methods were the best methods and that the PMS method was better than IMS and that the missing case should not be valued at more than 20% of the sample and the sample size should be large (Downey & King, 1998). This study also found that the IMS method was more appropriate than the PMS method when used with the data in nominal scale where as the PMS method was more appropriate them the IMS method when used with the data in Likert scale, not in accord with the percentage of the missing case and size of sample.

4. The IMS method mixed with the HDD method or the HDD-Z method and the PMS method mixed with the HDD or HDD-Z methods was the data substitution method that should be used because they were at a satisfactory level of quality. This study agreed with the findings of Huisman (1997) and Strike (2001).

Considerations regarding selections of methods of missing data imputation depend on the scale measuring variables, the cases being used and the researcher himself. If the researcher‘s interest is convenience, applicability, and simplicity, he or she should use the IMS method if the data under investigation is being measured in accordance with the nominal scale. The PMS method should be used when the data is being measured in accordance with the Likert scale. If accuracy is needed, a mixed method of IMS or PMS with HDD or HDD-Z should be used. In fact, these missing data techniques do not differ appreciably in quality.

Added material

Some experimental results are

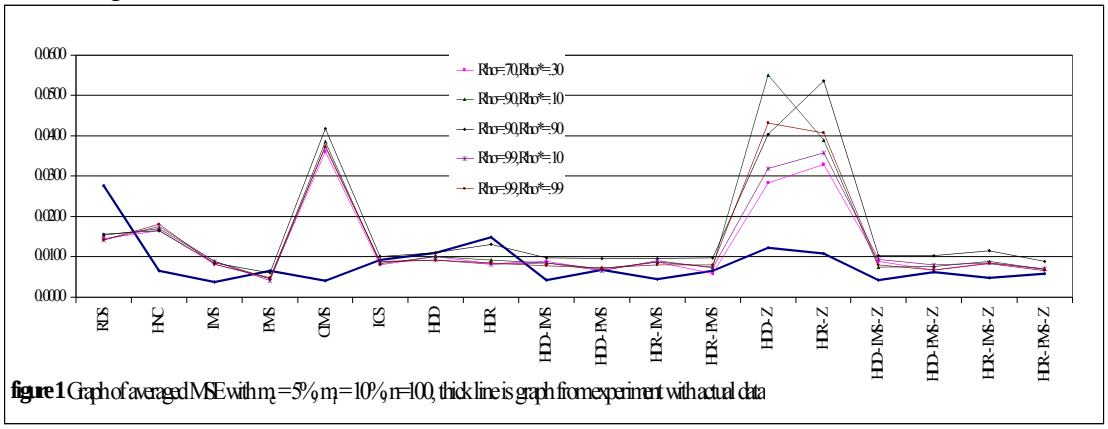

figure 1 Graph of averaged MSE with m_c=5%, m_1=10%, n=100, thick line is graph from experiment with actual data

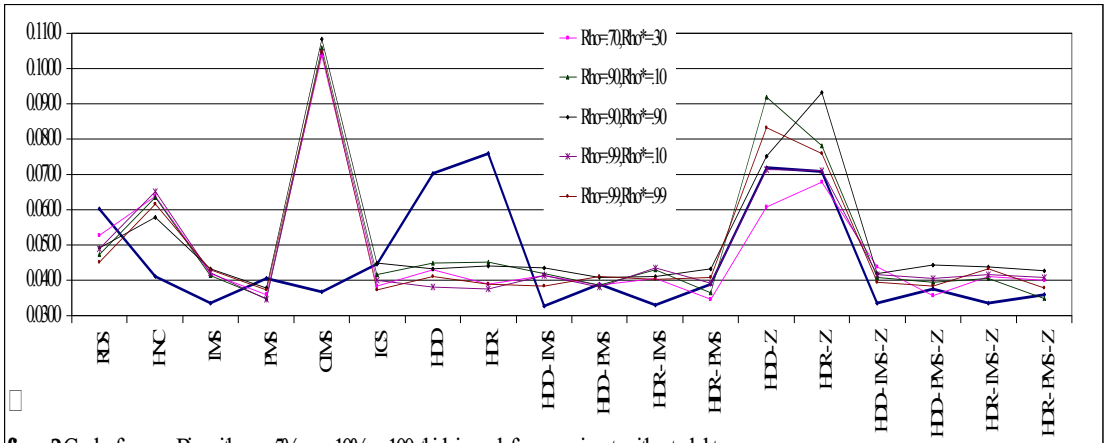

figure 2 Graph of average Bias with $m_2 = 5\%, m_1 = 10\%, n = 100$, thick is graph from experiments with actual data

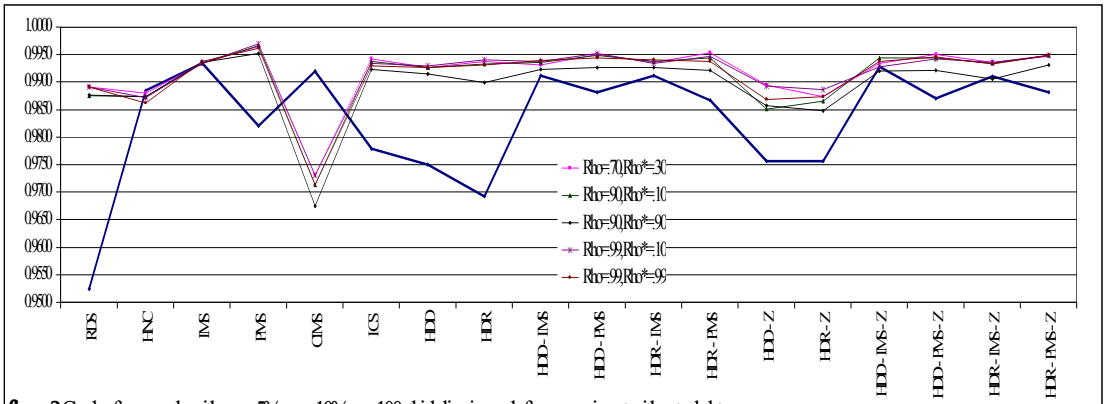

figure 3 Graph of average r with $m_2 = 5\%, m_1 = 10\%, n = 100$, thick line is graph from experiment with actual data

Reference

Churchill, G. A. and Wichern, D.W. (1978). A Comparison of Ridge Estimators. *Technometrics*, 20:301 – 311.

Couper, Mick P. and Groves, Robert M. (1996). Household-Level Determinant of Survey Nonresponse. *New Directions for Evaluation.*

Couper, Mick P. and Groves, Robert M. *Nonresponse in Household Interview Survey*. John Wiley & Sons Inc., N.Y., 1998.

DeLeeuw, E.D. (2001). Reducing Missing Data in Survey: An Overview of Methods. *Quality & Quantity*.

DeSilvio, Michelle Lee (1999). A Variance Ratio Statistics for Assessing the Missing Data Mechanism: An Empirical Study. Ph. D.-Tulane University. *Dissertation Abstracts International*.

Downey, R.G. and King, C. (1998). Missing Data in Likert Ratings: A Comparison of Replacement Methods." *The Journal of General Psychology*.

Gibbons, D.G. (1981). A Simulation Study of Some Ridge Estimators. *Journal of the American Statistical Association,* 76: 131 - 139.

Heeringa, Steven George. (2000). Multivariate Imputation of Coarsened Survey on Household Wealth. Ph.D. -- University of Michigan. *Dissertation Abstract International.*

Huisman, M. (1997). *Imputation of Missing Item Responses: Some Simple Techniques*. In Huisman,M. (edited).Nonresponse: Occurrence causes and Imputation of Missing Answers to Test Item. DSWO Press, Lieden University. The Netherlands, 1999.

Huisman, M. (1998). *Missing Data in Behavioral Science*. In Huisman, M. (edited). Item Nonresponse: Occurrence causes, and Imputation of Missing Answers to Test Item. DSWO Press, Lieden University, The Netherlands, 1999.

Huisman, M., Krol, B. and Van Sonderen, F.L.P. (1998). *Handling Missing Data by Re-approaching Nonrespondent*. Quality & Quantity. In Huisman, M. (edited). *Item Nonresponse:* Occurrence causes, and Imputation of Missing Answers to Test Item. DSWO Press, Lieden University, The Netherlands, 1999.

Ibrahim, J.G. (1988). Incomplete Data In Generalized Linear Model. Ph.D. -- University of Minnesota. *Dissertation Abstracts International*.

Jeon, Yoon Sook (1998). Inference in Structural Models With Missing Data. Ph.D. -- Iowa State University. *Dissertation Abstracts International*.

Kwang (2000). Variance Estimation after Imputation. Ph.D.-Iowa State University . *Dissertation Abstract International.*

King, C.V. and Downey, R.G. (1998). Missing Data in Likert Ratings: A Comparison of Replacement Methods. *The Journal of General Psychology*, 125: 175 - 191.

Ruben, Krohn, Karl (1992). Incomplete Data in Surveys of Human Populations: A Review of Sources and Solution (volume I and II). Ph. D.-University of Minnesota. *Dissertation Abstract International*.

McDonald, G.C. and Galarneau, D.I. (1975). A Monte Carlo Evaluation of Some Ridge-Type Estimators. *Journal of the American Statistical Association.* 70: 407 - 416.

Murata, Toshinoko. (2001). Item Nonresponse in Telephone Surveys*. Dissertation Abstract International.*

Paulin, Geoffrey D. and Ferraro, David L. (1994). Imputing Income in the Consumer Survey. *Monthly Labor Review.*

Rearden, David Thomas (1991). Missing Data in Linear Models (Parameter Estimation). Ph.D. - University of Kansas. *Dissertation Abstracts International*.

Roth, P.L. (1994). Missing Data: A Conceptual Review for Applied Psychology. *Journal of Personal Psychology*. 47:537-560.

Roth, P.L. And Switzer III, F.S. (1995). A Monte Carlo Analysis of Missing Data Techniques in HRM Settings. *Journal of Management.* 21: 1003-1023.

Roth, P.L. and Switzer III, F.S. (1999). Missing Data: Instrument-Level Haffalumps and item-Level woozles. *Research Method Forum.* Retrieved 25 February, 2001 from *www.aom.pace.edu/rmd /1999_ RMD-Forum_Missing_Data.htm1.*

Rylander, Roy G., Propst, Dennis B. and McMurtry, Terri R. (1995). Nonresponse and the Recall Biases in a Survey of Traveler Spending. *Journal of Travel Research.*

Samuhel, M.E. (1983). A General Approach to the Missing Data Problem. Ph. D. - The American University*. Dissertation Abstracts International*.

Strike, Devin et al. (2001). Software Cost Estimation with Incomplete Data. *IEE Transaction on Software Engineering.*

Wang, Betty Lu-Ti (2000). Imputation Methods for missing Data in Growth Curve Models. Ph. D. -University of Southern California*. Dissertation Abstract International*

Wisenbaer, Joseph M.(1998) . "A Compassion of Missing Data Treatments In the Content of Structural Equation Modeling ." Ph. D. -University of Georgia. *Dissertation Abstracts International*.

Witta, L.E. (2000*). Comparison of Missing Data Treatment in Producing Factor Scores.* Paper Present at the Annual Meeting of the American Evaluation Association, Honolulu. Nov. 1-5, 2000