The Report Committee for Xiaoyin Li
Certifies that this is the approved version of the following report:

A Comparison of Techniques for Handling Missing Data in
Scenarios with Different Missing Data Mechanisms

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Susan Natasha Beretvas

Carole K. Holahan
A Comparison of Techniques for Handling Missing Data in Scenarios with Different Missing Data Mechanisms

by

Xiaoyin Li, L.L.B.; M.A.; M.S.

Report
Presented to the Faculty of the Graduate School of The University of Texas at Austin in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Statistics

The University of Texas at Austin
December 2015
Abstract

A Comparison of Techniques for Handling Missing Data in Scenarios with Different Missing Data Mechanisms

Xiaoyin Li, M.S.Stat.
The University of Texas at Austin, 2015

Supervisor: Susan Natasha Beretvas

The purpose of this study was to illustrate the influence of missing data mechanisms on results of a multiple regression analysis and to demonstrate the influence of the use of traditional techniques (including listwise deletion, pairwise deletion and mean substitution) versus use of multiple imputation (MI) for handling missing data. A methodological approach involving a real dataset was adopted. Results from descriptive analyses and multiple regression models indicated that traditional missing data handling methods and MI yield similar regression coefficients and standard error estimates. Although the means and correlations are almost always biased regardless of missing mechanism and missing data techniques, the bias was less severe when using MI.
# Table of Contents

List of Tables ........................................................................................................... vi

Introduction ............................................................................................................... 1

Literature Review ..................................................................................................... 3
  The Problem ............................................................................................................. 3
  Missing Data Mechanism ......................................................................................... 4
  Types and Patterns of Missing Data ......................................................................... 5
  Old Techniques that Assume MCAR ....................................................................... 6
    Deletion Methods .................................................................................................. 6
    Single Imputation .................................................................................................. 7
  Preferred MAR Analysis Methods ............................................................................ 8
    Maximum Likelihood Estimation (ML) ................................................................. 9
    Multiple Imputation (MI) ....................................................................................... 9
    Comparison of Methods ....................................................................................... 11

Methods .................................................................................................................... 13
  Datasets ................................................................................................................... 13
  Statistical Analysis .................................................................................................. 14
  Traditional Techniques for Handling Missing Data ................................................. 14
  Multiple Imputation ................................................................................................. 14

Results ....................................................................................................................... 16

Discussion ................................................................................................................ 21

References ............................................................................................................... 23
List of Tables

Table 1: Influence of missing mechanism and procedures for handling missing data on the mean, standard deviation and sample size of the outcome variable.................................................................17

Table 2: Influence of missing mechanism and procedures for handling missing data on the correlation between independent (X) and dependent variable (Y). .........................................................................................................................18

Table 3. Influence of missing mechanism and procedures for handling missing data on parameter and standard error of the regression intercept (β0).....19

Table 4. Influence of missing mechanism and procedures for handling missing data on parameter and standard error of the regression coefficient (β1)..20
Introduction

Missing data is a ubiquitous issue and a significant challenge for social scientists. The goal of any analysis is to obtain unbiased estimates of the parameters of interest and to provide an estimate of the uncertainty about those estimates (standard errors or confidence intervals). Once the data are missing, the ability to either infer or deduce the nature of phenomenon is reduced, the extent to which is often unknown.

For decades, excluding the cases with missing data has been a widely used strategy and a default method for dealing with missing data for statistical software (Baraldi & Enders, 2010; Peugh & Enders, 2004). Since the 1970s, methodologists have developed a theoretical framework for missing data problems, and the methods for dealing with missing data have improved dramatically. There are many available methods for handling missing data and each may have different effects on resulting estimates. For this reason it is important to select the correct technique that is suited to the nature of the missing data and the statistical model for a particular analysis.

The purpose of this report was to illustrate the influence of missing data on the results of a statistical analysis (multiple regression) and to demonstrate the influence of traditional techniques for handling missing data and of multiple imputation. Throughout the analysis, the author used an artificially constructed, complete dataset abstracted from the first two waves of a longitudinal study on tobacco use among college students. Three incomplete datasets were generated to represent the three different missing data mechanism: missing completely at random (MCAR), missing at random (MAR) and
missing not at random (MNAR). To mimic a realistic dropout mechanism, 26% of the cases were sampled to have missing values on the dependent variable.

Three commonly used traditional methods (listwise deletion, pairwise deletion and mean substitution) and multiple imputation were used to handle the missing data in each incomplete dataset before data analysis. Results were compared against the baseline results calculated using the complete dataset. In order to demonstrate the performance of listwise deletion, pairwise deletion, and mean substitution in multiple regression, parameters and standard error estimates were compared across these techniques for handling missing data.
Literature Review

THE PROBLEM

Missing data can adversely impact the validity of statistical inferences. When observations are missing for any reason, interpretation (internal validity) and generalization of study results will be threatened (McKnight et al., 2007; Schafer & Graham, 2002). It has been recommended by the American Psychological Association (APA) that social scientists should handle and report missing data when they present empirical research results (Wilkinson & Task Force on Statistical Inference, Science Directorate, 1999).

Despite evidence for biased estimates due to missing data, many social and health science researchers fail to report missing data as part of their data analysis process. A group of researchers assessed prevalence and treatment of missing data in the literature by collecting 300 articles across a 3-year period from a prominent psychological journal. It was reported that about 90% of the published articles had data with missingness, and the average proportion of missing data was over 30% (McKnight, 2007; Sterner, 2011). However, very few of the researchers performed statistical procedures or made explicit disclaimers regarding remediation of missing data.

Some researchers with standard training in the social and health sciences might argue that advanced missing data handling techniques are difficult to implement and statistical software applications for handling missing data are limited. However, some solutions for dealing with missing data have improved substantially in the last two decades. In addition, techniques for handling missing data are available in most statistical software packages that are commonly used by social and behavioral scientists (Enders, 2011).
MISSING DATA MECHANISM

Before discussing different techniques for handling missing data, it is important to understand the theoretical framework for missing data problems. In the statistical literature, the classification system used refers to the “missing data mechanism”. Based on their relationships with measured variables, missing data are typically characterized as: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) (Little & Rubin, 2002; Rubin, 1976, 1987). In general, these mechanisms explain how the data are missing and provide assumptions for different techniques for handling missing data (Allison, 2002; Enders, 2010; Schafer & Graham, 2002).

MCAR is a condition when the probability of missing data on a variable is completely unrelated to any observed or unobserved data including the missing value itself. In other words, both observed data and the missing data can be thought of as random subsamples of the hypothetically complete sample. Statistically speaking, MCAR is ideal because we can get unbiased estimates based on observed cases. However, MCAR data are very unlikely to be found in real social or behavioral science research (Raghunathan, 2004).

MAR data occur when the probability of missingness depends on available information. In other words, data are MAR if missingness is related to other measured variables in the analysis model, but not to unobserved data (i.e., the hypothetical values that would have resulted had the data been complete). The “randomness” of the MAR mechanism is conditional on study-related variables in an analysis (Baraldi & Enders, 2010).

The MNAR mechanism describes a condition in which the probability of missing data is related to unobserved predictors or to the missing value itself. In other words, the
values of the missing scores cannot be predicted by available data. MNAR data that are analyzed typically result in biased estimates (Allison, 2002).

Missing data mechanisms are important assumptions for particular missing data techniques. For example, most traditional techniques for handling missing data can produce unbiased estimates only when missing data are MCAR. More advanced mainstream techniques such as maximum likelihood and multiple imputation can perform properly when MCAR or MAR assumptions are met. Although methodologists have developed analysis models to handle MNAR data (e.g., selection model, shared parameter model and pattern mixture model), limited evidence supports that these models perform better than other techniques (Baraldi & Enders, 2010). One big challenge is that no statistical test determines the mechanism underlying missing data. The only exception is that Little (1988) developed a chi-squared test with the null hypothesis that missing data is MCAR. However, it is impossible to distinguish or verify MAR from MNAR using a statistical model. Researchers need to examine their study design and procedure, weigh factors that may cause missing data and judge whether missing data are MAR or MNAR.

Methodologists have also pointed out that missing data classification applies to specific analyses instead of an entire data set (Baraldi & Enders, 2010). Consequently, the same data set may produce subsets of data that are MCAR, MAR, or MNAR depending on which variables are included in the analysis.

**TYPES AND PATTERNS OF MISSING DATA**

Survey research is commonly used in social and behavioral science, and missing data are often referred to as nonresponse. MCAR and MAR is also called *ignorable*
nonresponse, whereas MNAR is often used synonymously with nonignorable nonresponse (Graham, 2009).

*Item nonresponse* occurs when respondents fail to complete some parts of the survey and only partial data are available. In longitudinal studies, that is, research in which the same individuals are measured in two or more waves, *wave nonresponse* occurs when respondents present for some waves of data collection and are missing for others. In some cases, the participant is missing entirely from one wave but reappears to complete the survey in later waves; in other cases, respondents might leave the study and not return, which is a special case of wave nonresponse, referred to as *attrition* or *dropout* (Schafer & Graham, 2002). Researchers should pay attention to all the available data for each participant. Because measurements on the same individual over time tend to be correlated, missing information can then be partially recovered from earlier or later waves. Even with dropout or attrition, when information about participants is from one prior wave, the researcher might be able to get unbiased estimates from available data using modern missing data analysis procedures (Graham, 2012).

**OLD TECHNIQUES THAT ASSUME MCAR**

For a long time, researchers used traditional strategies like eliminating incomplete cases or imputing the missing scores with a single set of replacement values. These strategies require strict assumptions stated in the previous section and may cause bias in most situations. The goal of this section is to introduce these commonly used traditional procedures for handling missing data.

**Deletion Methods**

Deletion methods are default approaches used in several commonly used statistical packages such as SAS and SPSS and widely used in social and behavioral
science. It is easy to implement these methods because missing data are simply ignored and deleted from the analysis.

Listwise deletion, also known as case deletion and complete-case analysis, discards any case that has a missing value on any of the variables. The major advantage of listwise deletion is that it will produce a complete data set with a fixed sample size (Baraldi & Enders, 2010) and the complete dataset can be used with any analysis technique (Allison, 2002). The primary problem with listwise deletion is that it requires MCAR data. In real research settings, the MCAR assumption is often violated and listwise deletion will produce biased estimates. In addition, it can dramatically decrease the total sample size and statistical power, particularly for data sets with a large amount of variables or a large proportion of missing data (Baraldi & Enders, 2010).

With pairwise deletion, also known as available-case analysis incomplete cases are deleted on “an analysis-by-analysis basis” (Enders, 2010). This approach minimizes the number of cases omitted in given analyses, and results in different sample sizes for each bivariate analysis. Like listwise deletion, pairwise deletion can produce biased estimates when the data are inconsistent with an MCAR mechanism. Using different subsamples of cases also poses problems with measures of association (Enders, 2010).

**Single Imputation**

Single imputation is a collection of techniques whereby the researchers fill each piece of missing data with a single replacement score. There are dozens of single imputation techniques, and two of the more common approaches are: mean substitution which replaces missing scores with the mean of complete cases and regression imputation which replaces missing scores with predicted values from a regression estimated using complete variables (Enders, 2010). In longitudinal datasets, researchers often use
observed scores from the prior waves to replace missing values in later waves. These methods are convenient because they produce a complete dataset, but they also have limitations. Replacing missing values with a constant or group mean decreases standard errors and increases the risk of Type I errors, that is rejecting the null hypothesis when it should not be rejected.

If variable Y has \( n \) observations, \( k \) of which are missing but replaced by the mean of non-missing observations, the squared standard error of the mean can be expressed as follows:

\[
SE_M^2 = \frac{\sum_{i=1}^{n-k}(X_i - M)^2 + \sum_{i=n-k+1}^{n}(X_i - M)^2}{n(n-1)}
\]  

(1)

If a case is missing, \( X_i \) is replaced with a value that exactly equals \( M \), and deviations of imputed values from the mean, i.e., \( \sum_{i=n-k+1}^{n}(X_i - M)^2 \), are zero. Thus, \( SE_M^2 \) is always smaller and biased because in reality it is very unlikely that all missing values are equals to the mean.

**Preferred MAR Analysis Methods**

Maximum likelihood estimation (ML) and multiple imputation (MI) are two widely recommended “state of the art” methods for handling missing data. These two methods were designed specifically to achieve unbiased estimation with missing data when the MAR assumption holds. MI and ML methods can yield parameter estimates at least as good, usually better and often very much better than traditional methods (e.g., listwise deletion, pairwise deletion, mean substitution).
Maximum Likelihood Estimation (ML)

Instead of deleting or imputing values for missing data, ML involves analyzing the full, incomplete data set using the available data to compute maximum likelihood estimates of each parameter, which is the value of the parameter that is most likely to have resulted given the observed data. When data are missing, we can factor the likelihood function. The likelihood is computed separately for those cases with complete data on some variables and those with complete data on all variables. These two likelihoods are then maximized together to calculate the parameter estimates.

Multiple Imputation (MI)

Multiple imputation fills in estimates for the missing data. Unlike other imputation methods and ML, MI imputes multiple values – each including randomly sampled error – to recognize potential sampling error that goes unrecognized in conventional imputation procedures. A multiple imputation analysis consists of three distinct steps: the imputation phase, the analysis phase, and the pooling phase. The automated MI process is available in several software packages used in social and behavioral science research.

In the imputation phase, several copies of the data set are generated, and each copy contains a unique set of plausible replacement (imputed) scores. The primary goal of an imputation model is not to answer the research question but to impute missing values using observed variables. The variable selections used to do the imputation are very important in the imputation phase. It is recommended that in addition to variables that are of interest in the analysis phase, the imputation phase should include any other variables that might be correlates of the missing data, i.e., auxiliary variables (Enders, 2010). Inadequate variables in the imputation model can bias the subsequent analysis results.
The number of imputations needed in MI is critical to achieve efficient estimates. A formula was provided to show the efficiency of an estimate (E) based on number of imputations (m) and the proportion of missing data (γ) (Schafer & Olsen, 1998).

\[ E = (1 + \gamma/m)^{-1} \]  

(2)

Some researchers have argued that the rate of change in imputation efficiency is slower than the rate of change in the proportion of missing data and the number of imputations needed in MI was relatively small (Schafer & Olsen, 1998). For example, when \( m = 5 \), if the fraction of missing information increases from 10% to 30%, that is \( \gamma \) increases from .1 to .3, then \( E \) falls from .98 to .94. Increasing the number of imputations raises the efficiency slightly. Other researchers showed that although the relative efficiency may seems acceptable, MI with small \( m \) could lead to a critical dropoff in statistical power compared to the ML estimates given the same data (Graham, Olchowski, & Gilreath, 2007). Graham (2012, p.67) provided a table for the number of imputations required to maintain an acceptable power dropoff. According to the table, when the proportion of missing cases is 10% to 30%, at least 20 imputations are needed for a dropoff of less than 1% compared to ML (Graham, 2012).

In the analysis phase, the researcher performs the desired analysis on each imputed data set and gets multiple sets of parameter estimates and associated standard errors. In the final pooling phase, parameter estimates and standard errors from each imputed set are aggregated and pooled into a single set of results. Rubin (1987) derived rules to pool point estimates and standard errors to provide a set of results. The arithmetic average of the \( m \) estimates was defined as the multiple imputation point estimate \( \tilde{\theta} \) for parameter \( \theta \).

\[ \tilde{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i \]  

(3)

where \( \hat{\theta}_i \) is the parameter estimate from data set \( i \).
The aggregated of standard error (denoted by $SE$) formula is more complex because it requires aggregation of two components: the within-imputation variance (denoted by $W$) and the between-imputation variance (denoted by $B$) (Enders, 2010).

These can be calculated using the following expressions:

$$W = \frac{1}{m} \sum_{i=1}^{m} SE_i^2$$  \hspace{1cm} (4)

where $m$ is the number of imputed datasets, and $SE_i^2$ is the squared standard error from data set $i$.

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i - \bar{\theta})^2$$  \hspace{1cm} (5)

and then

$$SE = \sqrt{W + B + B/m}$$  \hspace{1cm} (6)

**COMPARISON OF METHODS**

A number of past studies have compared the performance of methods for handling missing data including listwise deletion, pairwise deletion, mean imputation, mode imputation, regression imputation, hot deck imputation, expectation maximization imputation, maximum likelihood and multiple imputation (Enders, 2011; Engels & Diehr, P., 2003; Peugh & Enders, 2004; Twisk & de Vente, 2002; Young, Weckman, & Holland, 2011). Based on findings reported in past research, the performance of a given method will be influenced by missing mechanism, analysis models and factors such as sample size and proportion of missing data in the sample (Twisk & de Vente, 2002; Young et al., 2011). No single technique was best for handling all missing data situations. However, in general, ML and MI methods are superior to traditional techniques in providing unbiased estimates.

If the MAR assumption holds, MI and ML can be asymptotically equivalent in their resulting in unbiased estimates without a loss of power (Graham, 2012). ML and MI
are different in many ways. Multiple imputation (MI) fills in each missing value with a random sample of plausible imputations, while maximum likelihood (ML) integrates the missing values according to the likelihood. MI requires careful selection of variables used to impute values and the number of imputations is critical to achieve the same accuracy as ML. Some researchers reported that ML is more efficient in both large and small samples (von Hippel, 2015), however, others have argued that the assumptions underlying use of ML are crucial and when sample size is small, ML estimates can be heavily biased (Schafer & Graham, 2002). MI has been recommended as a universal imputation method given its reliable performance even when the proportion of missing data is larger than 25% (Young et al., 2011).

A very limited number of studies has demonstrated the performances of techniques for handling missing data using SPSS, the most commonly used statistical software in social and behavioral science. Lack of guidelines for missing data analysis in SPSS may be one of the reasons why social and behavioral researchers avoided sophisticated methods for handling missing data. The purpose of this study was to illustrate the influence of the missing data mechanisms on a multiple regression analysis and to demonstrate the influence of the use of traditional techniques (including listwise deletion, pairwise deletion and mean substitution) and of multiple imputation for handling missing data in SPSS.
Methods

The purpose of this study is to demonstrate the performance of different techniques for handling missing data and the influence of missing data mechanisms on a multiple regression analysis. The four techniques compared in this study are listwise deletion, pairwise deletion, mean substitution and multiple imputation. All imputation methods and data analysis were performed with SPSS 21.

Datasets

All data were obtained from a longitudinal survey on tobacco use among college students in Texas. The author generated artificial datasets to compare the performance of the different techniques when handling data missing due to different missing data mechanisms. Besides a complete dataset, three incomplete datasets were generated. The sample of complete data entailed a group of current cigarette smokers in wave I who provided valid answers to the “days of cigarette use during the past 30 days” question in wave II (n = 820). All incomplete datasets were derived from this complete dataset and only the outcome variable, i.e., the number of days smoked during the past 30 days in the second wave, was generated to have missing data. In each dataset with missing values, about 26% of cases (n = 211) in the outcome variable were missing to mimic the real dropout rate.

Missing values in the first incomplete dataset were generated to be missing completely at random, MCAR. In the second incomplete dataset, data were missing dependent on an observed variable, which is the number of days of cigarette use in wave I (missing at random, MAR). Current smokers who used cigarettes more than 20 days
during the past 30 days in wave I were set to be missing for outcome variable. In the third incomplete dataset, 26% of the data were set to be missing when the score of the outcome variable is at the higher end (missing not at random, MNAR); i.e., the subjects who smoked more days during the past 30 days in wave II were assumed to be missing.

**Statistical Analysis**

Each of the four missing data handling methods was applied to both complete and incomplete samples containing missing data. In order to examine whether the number of days of cigarette smoking during the past 30 days in wave I is associated with the number of days of cigarette smoking during the past 30 days in wave II, simple regression models were estimated. The dependent variable measures the number of days smoked during the past 30 days in Wave II; and the independent variable measures the number of days smoked during the past 30 days in Wave I. The model is:

\[ Y = \beta_0 + \beta_1 X + e \]  

(7)

**Traditional Techniques for Handling Missing Data**

In SPSS 21, listwise deletion, pairwise deletion and mean substitution were built-in methods of handling missing value for multiple regression analysis with listwise deletion as the default method.

**Multiple Imputation**

The number of imputations used was 20 (i.e., \( m = 20 \)). When 10% - 30% of cases are missing from the data, use of at least 20 imputations is recommended to ensure
minimal power reduction (Graham, 2012). With 20 imputations, 20 complete data sets are generated and statistical analyses based on each set are performed.

Variables used for the imputation models included: sex, age, number of cigarettes smoked per day, intention of using cigarettes during the next 12 months, dating preference, perceived peer use of cigarettes, perceived social acceptances of cigarette smoking, perceived harmfullness of cigarette smoking, perceived addictiveness of cigarette smoking and living with a smoker. After imputing the data, the same simple regression model was estimated and the resulting parameter and standard errors pooled as described earlier.
Results

Table 1 shows descriptive information for the outcome variable (i.e., number of days smoked during past 30 days in wave II). Table 2 shows the correlation between the independent variable (X) and the dependent variable (Y) under the different missing data handling techniques for each dataset. In Table 3, the estimates of the intercept ($\beta_0$), standard errors ($SE$) and $p$ values were reported. In Table 4, the estimates of the regression coefficient ($\beta_1$), standard errors ($SE$) and $p$ values were reported. When compared with results from the analysis of the complete dataset, the means and correlations were almost always biased regardless of missing mechanisms and missing data techniques, however, the bias was less severe when using MI. Under MCAR and MAR, MI performed better than traditional techniques in recovering the correlation coefficients.

Results from the simple regression models indicated that although the significance of the effect of the independent variable on the dependent variable ($p < .001$) did not differ between different techniques and the datasets, the point estimates and standard error differed substantially. With mean substitution, standard errors were always smaller and the estimates were always more biased than other methods. Taking both estimate and standard errors into consideration, MI performed slightly better under MAR, however the advantage of MI was not evident for the MCAR data that were analyzed. The results of this analysis also demonstrated that all four methods examined in this study performed poorly under MNAR and the estimates are drastically biased.
Table 1: Influence of missing mechanism and procedures for handling missing data on the mean, standard deviation and sample size of the outcome variable.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCAR</td>
<td>8.25</td>
<td>10.65</td>
<td>820</td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>8.44</td>
<td>10.75</td>
<td>609</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>8.44</td>
<td>10.75</td>
<td>609</td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>8.44</td>
<td>9.26</td>
<td>820</td>
</tr>
<tr>
<td>MI pooled</td>
<td>9.03</td>
<td></td>
<td>820</td>
</tr>
<tr>
<td>MAR (n = 609)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>4.20</td>
<td>6.81</td>
<td>609</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>4.20</td>
<td>6.81</td>
<td>609</td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>4.20</td>
<td>5.87</td>
<td>820</td>
</tr>
<tr>
<td>MI pooled</td>
<td>7.48</td>
<td></td>
<td>820</td>
</tr>
<tr>
<td>MNAR (n = 609)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>2.49</td>
<td>3.58</td>
<td>609</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>2.49</td>
<td>3.58</td>
<td>609</td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>2.49</td>
<td>3.09</td>
<td>820</td>
</tr>
<tr>
<td>MI pooled</td>
<td>3.14</td>
<td></td>
<td>820</td>
</tr>
</tbody>
</table>
Table 2: Influence of missing mechanism and procedures for handling missing data on the correlation between independent (X) and dependent variable (Y).

<table>
<thead>
<tr>
<th>Method</th>
<th>MCAR</th>
<th>MAR</th>
<th>MNAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listwise Deletion</td>
<td>0.728</td>
<td>0.449</td>
<td>0.265</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>0.728</td>
<td>0.449</td>
<td>0.265</td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>0.633</td>
<td>0.185</td>
<td>0.164</td>
</tr>
<tr>
<td>MI pooled</td>
<td>0.709</td>
<td>0.721</td>
<td>0.385</td>
</tr>
</tbody>
</table>

Note. The correlation between X and Y is 0.715 in complete data set.
Table 3. Influence of missing mechanism and procedures for handling missing data on parameter and standard error of the regression intercept ($\beta_0$).

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCAR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>0.662</td>
<td>0.421</td>
<td>0.117</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>0.598</td>
<td>0.423</td>
<td>0.158</td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>2.567</td>
<td>0.355</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MI pooled</td>
<td>1.880</td>
<td>0.371</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MAR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>1.154</td>
<td>0.349</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>1.139</td>
<td>0.349</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>3.114</td>
<td>0.285</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MI pooled</td>
<td>1.219</td>
<td>0.325</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MNAR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>1.692</td>
<td>0.183</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>1.538</td>
<td>0.198</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>1.981</td>
<td>0.151</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MI pooled</td>
<td>1.758</td>
<td>0.175</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
Table 4. Influence of missing mechanism and procedures for handling missing data on parameter and standard error of the regression coefficient ($\beta_1$).

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>0.712</td>
<td>0.024</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MCAR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>0.725</td>
<td>0.028</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>0.731</td>
<td>0.028</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>0.548</td>
<td>0.023</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MI pooled</td>
<td>0.667</td>
<td>0.025</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MAR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>0.597</td>
<td>0.048</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>0.286</td>
<td>0.023</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>0.102</td>
<td>0.019</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MI pooled</td>
<td>0.584</td>
<td>0.040</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MNAR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise Deletion</td>
<td>0.123</td>
<td>0.018</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Pairwise Deletion</td>
<td>0.089</td>
<td>0.013</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Mean Substitution</td>
<td>0.047</td>
<td>0.010</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MI pooled</td>
<td>0.129</td>
<td>0.015</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>


Discussion

In this study, the author examined the influence of missing data mechanism on a simple regression analysis and demonstrated the influence of the use of traditional techniques (including listwise deletion, pairwise deletion and mean substitution) and of multiple imputation for handling missing data using SPSS. This was done by comparing results across techniques for complete data, datasets with 26% missing cases (MCAR, MAR and MNAR) and datasets in which the missing data were imputed by MI.

Consistent with previous studies, results of analyses using SPSS indicated that although MI did not always perform the best in providing unbiased estimates, its results were more reliable across different datasets (Peugh & Enders, 2004; Young et al., 2011). Mean imputation was the least mathematically sophisticated missing data imputation method and as expected it resulted in underestimated estimates of the standard error of the mean.

One limitation of this study was that it did not investigate effects of sample size, nor vary the proportion of missing data. It has been reported that the performance of a given method was dependent on factors such as proportion of missing data in the sample, sample size, distributions of variables in the sample, and the relationships between those variables. For example, Afifi and Elashoff (1966) found that regression-based imputation works best when such correlations are high, listwise deletion works best when such correlations are moderate, and mean imputation works best when correlations among regression variables are low. Young et al. (2011) showed that when missing data are less than 1%, the effect of missing data handling methods is very small; if 1-5% of data are
missing, simple methods such as listwise deletion and regression imputation work well; for 5-15% missing data, sophisticated methods should be selected; and when the proportion of missing data are over 15%, imputation results appear to be biased regardless of the imputation method used.

Other limitations of this study include that the dataset was artificial and the analysis model is overly simple. The dataset was generated from a longitudinal survey on tobacco use among college students in Texas. In order to understand the missing data mechanism, the author first took a subsample of the original dataset to obtain a dataset without missing data. In addition, participants of this survey were clustered with a sample of 24 schools and a multi-level regression model should have been fitted to obtain unbiased standard error estimation. Therefore, the results of this study can only be used to compare the techniques for missing data, but should not be used to understand cigarette smoking among college students.

Literature has suggested that, with more complex models, the advantage of modern missing data approaches over traditional methods could be substantial. Future studies are needed to compare the performance of missing data handling methods in more complex models.
References


http://doi.org/10.1016/j.jsp.2009.10.001


http://doi.org/10.1016/S0895-4356(03)00170-7

http://doi.org/10.1146/annurev.psych.58.110405.085530


24


25