DEALING WITH MISSING DATA IN THE WORK ENVIRONMENT AT KING KHALID UNIVERSITY

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE IN MATHEMATICS

IN THE GRADUATE SCHOOL OF THE TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

COLLEGE OF ARTS AND SCIENCE

BY

ZAHRA ASIRI, M.S.

DENTON, TEXAS

DECEMBER 2020

DEDICATION

First and foremost, praises and thanks to my family for all the support and encouragement. This project would have been impossible without the support of King Khaled University. Finally, my appreciation goes to everyone who has supported me to complete this work directly or indirectly.

ABSTRACT

ZAHRA ASIRI

DEALING WITH MISSING DATA IN THE WORK ENVIRONMENT AT KING KHALID UNIVERSITY

DECEMBER 2020

This study is about missing data mechanisms developed by Rubin, including missing data completely at random, missing data at random, and missing data not at random. This study utilizes a scenario at King Khaled University where potential employees complete a Post-Graduate General Aptitude Test (PGGA) to represent techniques for handling missing data. There are both traditional methods of handling missing data and modern methods that are more sophisticated for subsequent analyses and offer specific advantages. This study will go through the process of imputing data to understand how to deal with missing data depending on the missing data mechanism.

This study concludes by providing recommendations for handling missing data primarily through regression imputation and multiple imputation, which are exemplified through the researcher's simulated data related to the PGGA and job performance. Strengths and limitations of different techniques are discussed.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

INTRODUCTION

Missing data influences researchers across multiple disciplines (Baraldi & Enders, 2010). Specifically, missing data is defined as "the data value that is not stored for a variable in the observation of interest" (Kang, 2013, p. 1). There are different explanations for missing data. If not handled appropriately, missing data can negatively affect the conclusions researchers draw from the data, leading to inaccurate conclusions (Missing Values in Data, n.d.). According to Kang (2013), some of the problems posed by missing data include reduction in statistical power, bias in the estimation of parameters, reduction in the representativeness of the samples, and complications for the analysis of the study. Furthermore, these problems then lead to the overarching issue of threatening the validity of the study. Therefore, it is of the utmost importance for researchers to understand how to handle missing data appropriately.

First, one must understand why the data is missing because this influences the missing data mechanism, which then determines how to address the missingness. Collecting data are the responses of individuals to question of typically validated and reliable instruments or surveys. Regarding survey responses, there are several reasons why data can be missing, including a respondent's stress, fatigue, or lack of knowledge regarding a specific item/question (Missing Values in Data, n.d.). Some respondents may skip a few questions (Lin et al., 2012). For example, some people refuse to share age or

income in surveys even if the survey states that it maintains privacy. Understanding the reasons for missing data and their associated patterns or lack thereof can allow us to determine the types of missing data, which can be classified based on a system established by Rubin (as cited in Baraldi & Enders, 2010). According to this system, missing data can be classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR; Baraldi & Enders, 2010). These terms will be explored in Chapter 2.

Assuming good research practice, if less than 5% of the survey data is missing, a researcher can confidently drop the data with a low risk of adverse effects on the analyses. However, if more than 5% of the data is missing, it is important to find a way to appropriately handle the missing data, which is the focus of this paper (Missing Values in Data, n.d.). However, the methods of handling missing data differ depending on the analysis utilized on the collected data. For instance, multivariate analyses may favor dropping the records of those individuals generating missing data, while univariate analyses may be better served by addressing the missing data through statistical techniques (Missing Values in Data, n.d.). Replacing missing data utilizing a statistical technique is a process commonly referred to as imputation. For this paper, different imputation techniques will be discussed and examined thoroughly.

CHAPTER II

OVERVIEW

This chapter will cover some of the main concepts that appear repeatedly throughout this study. This chapter, initially, is devoted to the definition of missing values and the typical classification system for missing data, which was established by Rubin (1976). The second part of this chapter will cover the causes of missing data and its negative impact. However, not all missing values are unplanned which, according to Rubin (1976), has a positive impact. In fact, researchers use this idea of planned missing value designs to solve individual problems. For example, researchers design a survey that allows participants to randomly leave items blank with no response to make them feel comfortable, without sacrificing the validity of the survey (Little & Rhemtulla, 2013).

Missing Data Mechanism

According to Rubin (1976), missing data can be classified based on a missing data mechanism. The data can be MCAR, MAR, or MNAR (Baraldi & Enders, 2010).

Missing Completely at Random

The definition of MCAR is "when the probability that the data [are] missing is not related to either the specific value which is supposed to be obtained or the set of observed responses" (Kang, 2013, p. 402). This represents the ideal situation when it comes to missing data. An example of this would be if a piece of equipment fails to collect a measurement. For instance, if the blood pressure monitor stopped working in the middle

of a medical experiment and therefore there were three missing values for blood pressure. Another example would be if someone misses a survey due to a scheduling conflict (Baraldi & Enders, 2010). When data are termed MCAR, this offers a distinct statistical advantage for dealing with missing data because the analysis can remain free from biases. The downside is that if a substantial proportion of the data is MCAR and the researcher chooses to drop the data, there will be a loss of statistical power, which can hinder determination of significant outcomes of the study.

Missing at Random

This term is defined as "if missingness is related to other measured variables in the analysis model, but not to the underlying values of the incomplete variable" (Baraldi & Enders, 2010, p. 7). This type of missing data is systematic, and therefore, unlike its name implies, is not random at all (Baraldi & Enders, 2010). An example of MAR would be if a researcher was studying the relationship between depression and academic achievement. Because individuals with higher levels of depression tend to miss school, missing academic achievement measures are more probable for those with high levels of depression. In other words, a probabilistic pattern or difference exists on the "missingness," conditional on the level of depression.

Missing not at Random

The simplified explanation of MNAR is data that are "missing based on the would-be values of the missing scores" (Baraldi & Enders, 2010, p. 8). These cases are considered problematic because they lead to biases in the analyses (Kang, 2013).

Therefore, according to Kang (2013), a researcher must model the missing data to obtain an unbiased estimate of the parameters.

A Simulated Example

By using a small table of data, this study will help the readers understand the concept behind each method used to handle missing data values. The data in this thesis simulates a Saudi employee's selection scenario at King Khaled University where potential employees complete a Post-Graduate General Aptitude Test (PGGA), which according to the National Center for Assessment in Saudi Arabia measures the analytical and inferential capabilities among applicants to be researchers in the future.

The university hires applicants who receive high scores in the aptitude distribution, and the supervisor evaluates their job performance (JP) after a 1-year trial period. Individuals who applied for the job but did not take the PGGA exam or received a lower score (under 60) were never hired, so have no performance ranking. Thus, a function of JP scores systematically is missing. Moreover, two psychological well-being scores were deleted randomly to mimic a situation where the applicant's psychological well-being scores were accidentally lost. Table 2.1 displays an example of the missing data pattern occurring in the PGGA. The displayed table clearly presents that the JP variable is incomplete (contains missing data), and therefore must be handled appropriately for subsequent analyses. Without knowing about the data, the researcher may think that the missing data is MCAR, but upon closer examination, would see that the high and low JP scores are missing. Therefore, this data is determined to be MNAR.

In addition, the missing PGGA scores are related to JP, where low job performance values are associated with missing PGGA values, making this MAR data.

Table 2.1

The Missing Values Pattern

PGGA	Job Performance (incomplete)	Status (1 single, 0 married)	Job Performance (complete)	Status (1 single, 0 married) complete
63	-	1	18	1
64	15	1	15	1
65	9	0	9	0
-	4	0	4	0
67	12	0	12	0
69	-	1	2	1
70	13	-	13	1
73	12	0	12	0
76	19	0	19	0
77	-	-	18	0
78	9	1	9	1
80	10	1	10	1
80	15	0	15	0
83	13	-	13	1
-	3	1	3	1
-	8	1	8	1
88	-	-	16	0
90	5	0	5	0
92	12	1	12	1
96	-	1	6	1

Now, assume that King Khalid University hired all 20 applicants but terminated a few due to poor JP prior to their year evaluation. Note that after a year, the JP column is missing for the employees with the lowest performance scores. Thus, the probability of a missing JP score is dependent on the employee's performance, not PGGA or previous experience.

The conventional wisdom is that one cannot know as shown in Table 2.2 whether the missingness is MAR or MNAR. For example, Schafer and Graham (2002, p. 152) say, "[w]hen missingness is beyond the researcher's control, its distribution is unknown, and MAR is only an assumption. In general, there is no way to test whether MAR holds in a data set, except by obtaining follow up data from non-respondents . . ."

Table 2.2

Describe the Missing Values Pattern MAR, MCAR, MNAR

PGGA	Complete	MAR	MCAR	MNAR
63	18	-	-	18
64	15	-	15	15
65	9	-	9	-
65	4	4	4	-
67	12	-	12	12
69	2	2	-	2
70	13	13	13	-
73	12	12	12	12
76	19	19	19	19
77	18	18	-	-
78	9	9	9	9
80	10	10	10	-
80	15	15	15	15
83	13	-	13	13
84	3	3	3	3
85	8	8	8	8
88	10	-	-	10
90	5	5	5	5
92	12	12	12	12
96	6	6	-	6

Why is the Missing Data Mechanism Important?

The main factor of solving missing values is to understand the reasons that generate missing data, which is a serious challenge for researchers. The missing data mechanism clarifies the situations that lead to missing values. For this reason, once determining the condition of the missing values mechanism, researchers can use different methods to handle missing values.

Consequently, under MAR and MNAR, researchers may use multiple imputations since those conditions do not require information about the relationship between the distribution and the data (Enders, 2010). These mechanisms help researchers find missing values by observing the relationships between missing and observed values.

CHAPTER III

TRADITIONAL METHODS FOR DEALING WITH MISSING DATA

Missing data is not a new concept. Researchers have studied missing values for the past 10 years, and there have been many methods developed to deal with missing data. However, this chapter will focus on the traditional methods of handling missing data. The most common method that most researchers in a variety of fields, such as psychology and education, still use is deletion.

Deletion Methods

Listwise and pairwise deletions are the most popular methods to handle missing data. However, it only can be used when the missing data are completely at random. It is recommended that when less than 5% of the data are missing, the researcher drops the data (Missing Values in Data, n.d.).

Listwise Deletion

By using listwise deletion, a researcher will remove the entire data set for subjects 1, 6, 10, 17, and 20 from Table 2.1 before doing any further analysis. Therefore, listwise deletion leads to a reduced sample size. This factor alone concludes that listwise deletion might only be used when there are fewer missing values and a larger sample size (Marsh, 1998). Listwise deletion can be biased when the missing value is related to its observation value (MNAR; Pepinsky, 2018).

Pairwise Deletion

While listwise deletion minimizes the sample size, pairwise deletion helps to keep the loss of subjects that have missing values in one variable, but not all variables. This method is used by deleting only the missing values in the variable not the entire subjects. Pairwise deletion can be used when missingness is MCAR or MAR. This technique attempts to increase the analysis power. It helps to understand pairwise better by using a correlation matrix. A correlation matrix is used to measure the strength of the relationship between an independent and dependent variable. The standard errors are calculated by using the average sample size in most software packages, which causes a misestimate of the standard errors (Missing Data: Listwise vs. Pairwise, n.d).With listwise deletion, the size of the population will be the same, but with pairwise deletion, population size will vary for every correlation.

Using independent T-test, listwise minimize the sample data from 20 to 13 samples where there are seven missing values. Table 3.1 shows that the subjects were removed from both variables PGGA and JP when analyzing the data using listwise deletion. The output on the right of Table 3.1 shows pairwise deletion, where the missing values were removed in the variable.

Table 3.1

Listwise Deletion (Top) and Pairwise (Bottom) from Table 2.1

	Status	N	Mean	Std. deviation	Std. error mean
PGGA	1.00	6	80.5000	9.41807	3.84491
	.00	7	73.7143	9.19627	3.47586
	1.00	6	9.5000	4.03733	1.64823
	.00	7	10.8571	5.33631	2.01694
	Status	N	Mean	Std. deviation	Std. error mean
JP	1.00	9	79.0000	11.73669	3.91223
	.00	7	73.7143	9.19627	3.47586
	1.00	6	9.5000	4.03733	1.64823
	.00	7	10.8571	5.33631	2.01694

Single Imputation Methods

Replacing a missing value with a plausible value, such as the mean, median, or mode of a variable, is the idea behind imputation. Therefore, using an imputation method transforms incomplete data to imputed data (Schafer & Graham, 2002).

Imputation seems extremely attractive to researchers because of its simplicity when dealing with a mean imputation — a researcher simply finds the mean of a variable and uses that value for missing data for the variable. Imputation is also more popular than deletion methods due to its way of keeping all subjects, which are sacrificed when using listwise deletion (Schafer & Graham, 2002). The term *single imputation* comes from the way that this method seeks to generate a single value replacement for each missing data value (Enders, 2010).

Mean – Median – Mode Imputation

Mean imputation is when the mean is substituted for every missing value for a specific variable. For example, a survey has 1,000 participants and 15 of them did not answer a question related to income. By using the mean of the income, analysts will substitute the mean in every missing value.

To use this imputation method, the researcher must assume that the missing values were missing completely at random. The disadvantage of inserting the mean for the missing value can at times be misleading, specifically if data are MNAR. Note that researchers cannot use this strategy if there is an extreme point because it will not give an accurate result. Therefore, if there is an extreme point or more points, it should be reduced by the median imputation. Another way to use imputation, if the incomplete data has repeated values, is by using mode imputation. This works through imputing each missing value by the number that is repeated most. Using the observed MAR values in Table 2.2, the mean is calculated, and each missing value is replaced by the mean to complete the data:

$$\mu_Y = \frac{4 + 2 + 13 + 12 + 19 + 18 + 9 + 10 + 15 + 3 + 8 + 5 + 12 + 6}{14} = 9.71$$

Once all missing values are replaced with the mean, we can use a scatterplot to see how the new values act. Six of the employers, whose JP is missing, will receive 9.7.

Hot Deck Imputation

In hot deck imputation, or "similar response pattern imputation SRPI," the researcher has the recipient (missing value) and the donor (observed value), where the

missing values are replaced with observed value. However, the condition to do that is finding similarities of characteristics between observed and missing values. Random hot deck imputation, where the donor should be selected randomly from a set of possible donors, is another way to handle missing data.

Deterministic hot deck methods are specified by a single donor by usually using the nearest neighbor based on the similarities of characteristics (Adridge & Little, 2010). Hot deck imputation is used only when the value is missing completely at random or missing at random.

CHAPTER IV

MODERN METHODS FOR DEALING WITH MISSING DATA

This chapter will focus on the modern methods of dealing with missing data, including regression imputation and multiple imputation. The process of conducting each method is described in detail and examples using the simulated data set for this thesis are provided throughout the chapter.

Regression Imputation

Regression imputation or conditional mean imputation is a strategy where missing values are predicted from a regression model. Finding the regression model and then the fit line model, an investigator can predicate the missing values (Zhang, 2016). This method can be used only when the missing values are missing completely at random or missing at random. Using regression imputation when the missing value is missing not at random is not valid. According to Kang (2013), the regression imputation method has some advantages over deletion and mean imputation. Essentially, the existing variables and given data are used to make a prediction about the missing data. The predicted value is then replaced as the actual obtained value.

To clarify the linear regression imputation strategy, an analyst should know what other information should be related to the missing values. For example, it is not fair to predict missing values about smoking questions with other variables like income. On the other hand, this value may be used to predict information relating to health issues. This

method is better than others as the mean and variance will not change significantly before and after missing values have been found. To illustrate this idea, use the Y_{obs} values to be analyzed to find the predicted regression model.

Table 4.1

Coefficients Summary

		Unstandardized Coefficients		Standardized Coefficients		
	Model	B	Std. Error	Beta	t	Sig.
1	(Constant)	13.716	13.794		.994	.340
	PGGA	-.050	.172	-.084	-.292	.775

$$\hat{Y} = 13.716 + (-.050)\, X \qquad (2.1)$$

In Table 4.1 where $\hat{Y}$ is the predicted JP, −.050 is the slope, 13.716 is the y-intercept, and X (PGGA) is an independent variable. This equation (2.1) will help the supervisor who wants to find all missing values of employer's JP. Also, letting SPSS do linear regression imputation is an easy way to predict missing values when the sample size is large.

Incomplete values in Table 4.2 were predicted by using regression imputation. In fact, an employer who scored 63 on the PGGA will get 10.566 and so on with the other missing values.

Table 4.2

Regression Imputation

PGGA	Incomplete	Predicted value (JP)
63	-	10.566
64	-	10.516
65	-	10.466
65	4	4
67	12	10.366
69	2	2
70	13	13
73	12	12
76	19	19
77	18	18
78	9	9
80	10	10
80	15	15
83	-	9.566
84	3	3
85	8	8
88	-	9.136
90	5	5
92	12	12
96	6	6

In Table 4.3 there are 20 predicted samples, and by looking to the coefficient of determination (R^2), 1.3% of the variability in JP can be explained by the regression line given PGGA scores. When interpreting the correlation coefficient R, the linear relationship between the two variables positive and weak.

Table 4.3

Model Summary

Model	R	R square	Adjusted R square	Std. error of the estimate
1	0.115[a]	.013	-.041	10.27431

[a]:predictors:(Constant),JP

Multiple Imputation

In 1970, multiple imputation was developed as a method for handling missing data, and in recent years it has become increasingly popular (Jakobsen et al. 2017). Imputation is used by replacing a missing value with an estimated value based on other available information. According to Kang (2013), what sets multiple imputation apart from other techniques is, "instead of substituting a single value for each missing data, the missing values are replaced with a set of plausible values which contain the natural variability and uncertainty of the right values" (p. 405). The central idea is that the researcher predicts missing data using the existing data, like a regression imputation, but then a full data set of ascribed data is created and then the process iterates making multiple imputed data sets (Kang, 2013). Then, the assigned data is analyzed, just as a complete data set would be, and it produces multiple analysis results. The researcher then combines the multiple analysis results into a single comprehensive analysis result. This method incorporates uncertainty into the data and also restores natural variability (Kang, 2013). In addition, multiple imputation is robust and valid.

Multiple imputation follows three steps:

1. Use the observed values (plausible values) to be replaced with missing values randomly — this is called imputation.
2. Estimation step — after the imputation step, analyze every data set separately.
3. Pooling step — when an analyst collects the results from each completed dataset to be combined in one result.

Having more regressions means having less bias, so 5 to 10 regressions might be enough for 50 to 100 samples. Investigators will end up having different regression equations, and those equations produce different estimates of parameters, therefore, will lead to a large variance (Padgett et al., 2014). Note that it looks like the single imputation method, but there will be more than one regression. The data in Table 4.4 is used to find the regression of PGGA which will illustrate the idea behind multiple imputation.

Table 4.4

Multiple Imputation

PGGA	JP	Status
63	-	Single
64	15	Single
65	9	Married
65	4	Married
67	12	Married
69	-	Single
70	13	-
73	12	Married
76	19	Married
77	-	Single
78	9	Single
80	10	Single
80	15	Married
83	13	-
84	3	Single
85	8	Single
88	-	-
90	5	Married
92	12	Married
96	-	Married

Table 4.5 shows the new model after replacing missing values after using multiple imputation. The employer who scored 63 in the PGGA exam, will get five for JP while the employer who scored 96 will receive a JP score of eight. In addition, the employer who received 13 in the job performance will be a single person which might be predicted from the other employer who received the same grade.

Table 4.5

Multiple Imputation Five Times

IMPUTATION	PGGA	JP	Status
5	63	5	1
5	64	15	1
5	65	9	0
5	65	4	0
5	67	12	0
5	69	5	1
5	70	13	1
5	73	12	0
5	76	19	0
5	77	5	1
5	78	9	1
5	80	10	1
5	80	15	0
5	83	13	1
5	84	3	1
5	85	8	1
5	88	8	0
5	90	5	0
5	92	12	1
5	96	8	0

Figure 4.1 explains the number of missing values as percentages. The first pie chart illustrates how many missing values are in the variables, so the two variables have some missing values each. The center pie chart shows how many values in each variable are missing. Since there are 20 subjects, 60% are observed and 40% are missing. In the right pie chart, 20% of the values are missing from the entire sample

Figure 4.1

The Number of Missing Values in Percentage

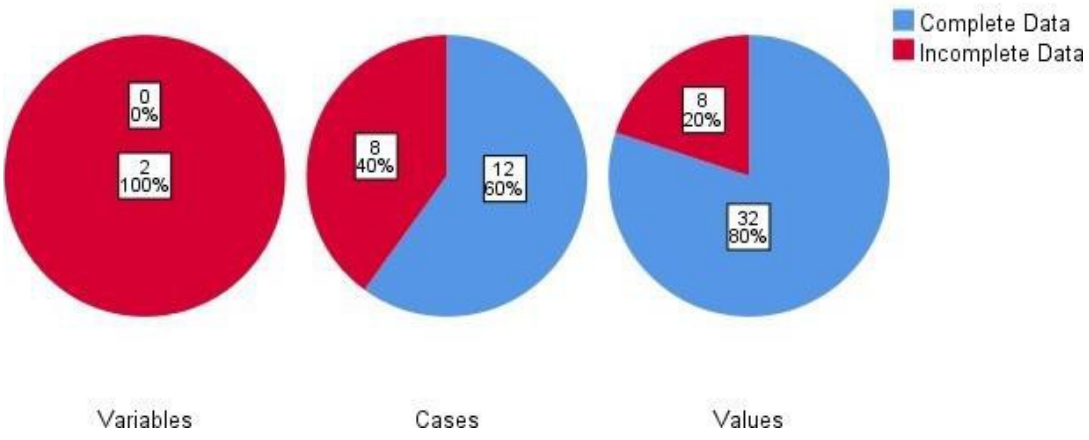

Table 4.6 shows that JP has five missing values, and the employer's status has three missing values. It is extremely helpful for researchers to know how many values are missing for each variable and which one of those variables contain many missing values. The researcher can use this information to try to determine the reason for missingness. Table 4.5 starts with variable JP because it has the most missing values. The next variable has the next highest number of missing values, and so on.

Table 4.6

Variable Summary

	Missing		
	N	Percent	Valid N
JP	5	25.0%	15
Status	3	15.0%	17

Table 4.7 shows how many imputations have been done. Here, the number of imputations completed is five by using linear regression.

Table 4.7

Imputation Specifications

	Imputation Specifications
Imputation Method	Automatic
Number of Imputation	5
Model for Scale Variables	Linear Regression
Intersections Included in Models	(None)
Maximum Percentage of Missing Values	100.0%
Maximum Number of Parameters in Imputation Model	100

Table 4.8 displays many versions of multiple imputation, and it uses the fully conditional specification approach. Conditional distributions or regression models are used to clarify each variable with missing values under conditions that applied on all the other variables in the imputation model. As a result, the imputations were generated by estimating each regression model by using only the observed values for the variables and

imputed values for the other variables that have missing values (Lee & Carlin, 2010). The fully conditional specification method was completed 10 times.

Table 4.8

Imputation Results

	Imputation Specifications
Imputation Method	Fully Conditional Specification
Fully Conditional Specification Method Iterations	10
Dependent Variables	Imputed Not Imputed (Too Many Missing Values) Not Imputed (No Missing Values)
Imputation Sequence	Status, JP

Table 4.9 shows the number of missing values in each variable and how many times the imputed values have been created. JP variable has three missing values and 15 imputed values while there are five missing values in the status variable with 25 imputed values.

Table 4.9

Imputation Models

Imputation Models				
	Type	Effective	Missing Values	Imputed Values
Status	Logistic Regression	JP	3	15
JP	Logistic Regression	Status	5	25

Table 4.10 shows the new observed value of each imputation and the proportion for single and married employers in the complete data after imputation. The same work has been done with status variable in Table 4.11.

Table 4.10

The JP Variable Imputation

Data	Category	N	Percent
Original Data	3	1	6.7
	4	1	6.7
	5	1	6.7
	5	1	6.7
	8	2	13.3
	9	1	6.7
	10	3	20
	12	2	13.3
	13	2	13.3
	15	1	6.7
Imputed Values	3	2	40
	5	3	60
	3	2	40
	5	3	60
	8	2	40

	19	3	60
	4	3	60
	10	2	60
	3	2	40
	4	3	40
Complete Data	3	3	15
	4	1	5
	5	4	20
	8	1	5
	9	2	10
	10	1	5
	12	3	15
	13	2	10
	15	2	10
	19	1	5

Table 4.11

The Status Variable Imputation

Data	Category	N	Percent
	0,1	9,8	52.9, 47.1
	0,1	2,1	66.7,33.3
	0,1	2,1	66.7,33.3
Original Data	0,1	1,2	33.3,66.7
	0,1	1,2	33.3,66.7
Imputed Values	0,1	2,1	66.7,33.3
	0,1	2,1	66.7,33.3
	0,1	2,1	66.7,33.3
	0,1	1,2	33.3,66.7
	0,1	1,2	33.3,66.7
Complete Data	0,1	11,9	55,44
	0,1	11,9	55,44

0,1	11,9	55,44
0,1	10,10	50
0,1	10,10	50

After using multiple imputations, the variable ‘status’ was added to the dataset to form a group of single and married for the JP variable. For the original data, the single’s mean is nine and it is seven for the married group. In the fifth imputation, the average of the single group is 8.2 and it is 9.5 for the married group. Table 4.10 shows these imputations.

CHAPTER V

CONCLUSION

Discussion

While missing data can sometimes be prevented by designing a simple and clear survey when conducting survey research, there are times when even well- designed studies result in missing data. Rubin (1976) developed the missing data mechanism to identify different types of missing data. In this paper, it has been shown that depending on the missing data mechanism at hand, missing data will be dealt with differently. For instance, this paper discussed traditional methods of handling missing data such as deletions and single value imputations. However, in more recent times, additional modern techniques such as regression imputation and multiple imputation were developed for handling missing data in a more sophisticated manner.

Recommendation

The supervisor should be aware that exposure to missing data is common and potential, so this will help to solve missing data rapidly. At the same time, the fear of having missing data should not be a reason to forgo the study. Using useful programs such as SPSS for analyzing data will help handle missing data. Gaining experience with different types of missing data, such as MAR, MNAR, and MCAR, will improve the way one handles missing data. Missing data is a common challenge in research and influences its integrity, so it is worth studying to potentially minimize its affects.

The results from the methods addressed in this study, such as single and multiple

imputation, are particularly important methods in missing data. All the imputation methods discussed explain how to deal with missing data, but some methods are better than others depending on the data.

REFERENCES

Adridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1), 40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*(1), 5–37. https://doi.org/10.1016/j.jsp.2009.10.001

Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press.

Jakobsen, J. C., Gluud, C., Wetterslev, J. & Winkle, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials: A practical guide with flowcharts. *BMC Medical Research Methodology 17*, 162. https://doi.org/10.1186/s12874-017-0442-1

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology, 64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology, 171*(5), 624–632. https://doi.org/10.1093/aje/kwp425

Lin, J. Y., Lu, Y., & Tu, X. (2012). How to avoid missing data and the problems they pose: Design considerations. *Shanghai Archives of Psychiatry*, *24*(3), 181–184. https://doi.org/10.3969/j.issn.1002-0829.2012.03.010

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives, 7*(4), 199– 204.

https://doi.org/10.1111/cdep.12043

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 22–36. https://doi.org/10.1080/10705519809540087

Missing Data: Listwise vs. Pairwise. (n.d.). *Statistics Solutions*. Retrieved May 12, 2020 from https://www.statisticssolutions.com/missingdata-listwise-vs-pairwise/

Missing Values in Data. (n.d.). *Statistics Solutions.* Retrieved from https://www.statisticssolutions.com/missing-values-in-data/

Padgett, C., Skilbeck, C., & Summers, M. J. (2014). Missing data: The importance and impact of missing data from clinical research. *Brain Impairment*, *15*(1), 1–9. https://doi.org/10.1017/BrImp.2014.2

Pepinsky, T. (2018). A note on listwise deletion versus multiple imputation. *Political Analysis*, *26*(4), 480–488. https://doi.org/10.1017/pan.2018.18

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989x.7.2.147

Zhang, Z. (2016). Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine*, *4*(1), 9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38