MISSING VALUES - A CLOSER LOOK

by

KERRI THORPE

A project submitted in partial fulfilment of the requirements for the degree of

Master of Science (Cwk/Thesis)

in

Mathematical Statistics

Department of Statistics

Rhodes University

December 2016

Supervisor: Dr. I. Garisch

Abstract

Problem: In today's world, missing values are more present than ever. Due to the ever-changing and fast paced global society in which we live, most business and research data produced around the world contain missing data. This means that locating data which is meticulously precise can be a hard task in itself, but at times may prove essential as the consequences of making use of incomplete data could be disastrous. The reasons for missing data cropping up in almost all forms of work are numerous and shall be discussed in this dissertation. For example, those being interviewed or polled may choose to simply ignore questions which are posed to them, recording equipment may malfunction or be misplaced, or organisers may not be able to locate the respondent in order to rectify the missing data. Whatever the reasons for data being incomplete, it is necessary to avoid having to use inefficient and incomplete data as a result from the above problems. Therefore, various strategies or methods have been developed in order to handle these missing values. It is important, however, that these strategies or methods are utilised effectively as missing data treatment can introduce bias into the analysis. This dissertation shall look at these and other problems in more detail by using a data set which consists of records for 581 children who were interviewed in 1990 as part of the National Longitudinal Survey of Youth (NLSY).

Approach: As mentioned above, many strategies or methods have been developed in order to deal with missing values. More specifically, traditional methods such as complete case analysis, available case analysis or single imputation are widely used by researchers and shall be discussed herein. Although these methods are simple and easy to implement, they require assumptions about the data that are not often satisfied in practice. Over the years, more up to date and relevant methods, such as multiple imputation and maximum likelihood have been developed. These methods rely on weaker assumptions and contain superior statistical properties when compared to the traditional techniques. In this dissertation, these traditional methods shall be reviewed and assessed in SAS and shall be compared to the more modern techniques.

Results: The ad hoc techniques for handling missing data such as complete case and available case methods produce biased parameter estimates when the data is not missing completely at random (MCAR). Single imputation techniques likewise produce biased estimates as well as result in the underestimation of standard errors. Although the expectation maximisation (EM) algorithm yields un-

biased parameter estimates, the lack of convenient standard errors suggests that using this algorithm for hypothesis testing is not a good idea. Multiple imputation, however, yields unbiased parameter estimates and correctly estimates standard errors.

Conclusion: Ignoring missing data in any analysis produces biased parameter estimates. Using single imputation to handle missing values is not recommended, as using a single value to replace missing values does not account for the variation that would have been present if the variables were observed. As a result, the variance will be greatly underestimated. The more modern missing data methods such as the EM algorithm and multiple imputation are preferred over the traditional techniques as they require less stringent assumptions and they also mitigate the downsides of the older methods.

Keywords: Missing Data; Missing Data Mechanisms; Complete Case; Available Case; Single Imputation; Maximum Likelihood; EM Algorithm; Multiple Imputation

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In today's world, information and knowledge play a crucial role in the lives of all individuals and, in many cases, the success of many businesses and institutions. Researchers in particular believe that it is becoming more and more common for organisations to base strategic decisions on information inferred from data. Thus data quality plays a critical role in decision making.

One hallmark of good data quality is the absence of any missing values within a data set. However, in practice, one is bound to face the familiar and unavoidable problem of missing values. According to Osborne (2013), data is considered to be 'missing' or 'incomplete' if any data on any variable does not exist. These missing values may either be legitimate or illegitimate. Legitimate missing data may be described as an absence of data, however, the data is absent for a reason. An example is a survey which first asks individuals whether they are married, and second, if so, for how long. It makes sense for those individuals who are not married to skip the second part of the question.

On the other hand, an example where illegitimately missing values can hinder a study is where participants may, for example, refuse or forget to answer specific questions. Or, where files are lost or human error has played a role and the data has not been recorded correctly. According to Osborne (2013), legitimate and illegitimate missing values can add meaning to analyses. For instance, in the above case regarding legitimate missing values, the missing values can provide information and reinforce the status of an individual, as well as provide researchers with the opportunity to confirm an individual's response.

The main focus of this dissertation is, however, on illegitimate missing data. In order to find solutions to the problems described above, researchers spend copious amounts of time, effort and funding to minimise incomplete data or non-responses amongst participants. Given the expense of collecting data, researchers cannot afford to start over or wait until there are effective methods of collecting in-

formation in place.

It is therefore important to determine what options are available for analysing data with missing information. According to Luengo et al. (2012), missing information can be treated in three different ways:

1. Those cases that have missing values that can be discarded; or
2. Maximum likelihood procedures may be used, where available data for each subject is used to compute a likelihood and to estimate parameters where these parameters are later used for imputation; or
3. The technique of imputation may be used to treat missing values. Here, missing values are replaced with estimated values. In most cases, variables in a data set are not independent of each other and finding relationships between these variables will be important in order to determine missing values.

This dissertation shall demonstrate how missing values make data analysis very difficult and will provide an in-depth study of some of the methods available to address this problem. Deletion techniques which include pairwise and listwise deletion will be discussed. Single imputation methods such as mean/mode/median, k-nearest neighbour as well as regression imputation will be studied. Further, maximum likelihood and multiple imputation will also be discussed. It is important to note that the above list is by no means exhaustive.

The remainder of this thesis will be set out as follows: Chapter 2 shall discuss the goals and criteria of the analysis, as well as look at the various missing data patterns and missing data mechanisms. Chapter 3 shall provide a discussion of the methods that may be used to treat missing values. In Chapter 4, the different methods discussed will be applied to a data set that has missing values. The data set chosen is one used by Allison (2001) and consists of records for 581 children who were interviewed in 1990 as part of the National Longitudinal Survey of Youth (NLSY). The different missing data methods will then be compared in order to find the most appropriate technique to deal with this problem. Chapter 5 will address future scope and some concluding remarks will be made.

Chapter 2

Understanding the Basics

Before beginning an examination of the specific methods referred to in Chapter 1, Chapter 2 shall spend some time discussing the aims or goals of the analysis. Then, the differences between missing data patterns and missing data mechanisms will be examined as these terms are often used interchangeably when they are, in fact, different in meaning. Lastly, possible assumptions as to how the data came to be missing will be discussed as this impacts the performance of missing data methods.

2.1 Goals and Criteria of the Analysis

The aim or goal of every statistical analysis should be to draw valid inferences about a population of interest, whether or not missing values exist in the data (Schafer & Graham (2002)). Researchers, however, should not focus on trying to estimate or recover missing values that may be present in a data set, as trying to recover these missing values could result in invalid inferences being made. For example, if one decides to replace all missing values in the data set with the average of the observed values, the missing items might be predicted correctly but the variance and correlations may be adversely affected.

Schafer & Graham (2002) state that Neyman & Pearson (1933) and Neyman (1937) developed principles in order to assess statistical procedures. That is, let Q represent a quantity denoting the population of interest and let $\hat{Q}$ be an estimator of Q. If missing values exist in the sample, then the techniques used to treat them should play a role in the overall method used to compute $\hat{Q}$.

Further, the bias of $\hat{Q}$ is defined as the difference between the estimator's expected value ($E(\hat{Q})$) and the true value of Q (Hand et al. (2001)). Ideally, one would want the bias as well as the variance and standard deviation of $\hat{Q}$ to be as small as possible. According to Hand et al. (2001), the mean square error (MSE) is defined as the average squared difference between the estimator and the true

value of the quantity being measured. The equations below show that when the bias and the standard deviation are combined into a single measure, the MSE is obtained (Schluchter (2014)).

$$MSE\left(\hat{Q}\right) = E\left[\left(\hat{Q}-Q\right)^2\right] \tag{2.1}$$

$$= E\left[\left(\hat{Q}-E\left[\hat{Q}\right]+E\left[\hat{Q}\right]-Q\right)^2\right] \tag{2.2}$$

$$= \left(E\left[\hat{Q}\right]-Q\right)^2 + E\left[\left(\hat{Q}-E\left[\hat{Q}\right]\right)^2\right] \tag{2.3}$$

$$= \left(Bias\left(\hat{Q}\right)\right)^2 + \mathrm{Var}\left(\hat{Q}\right) \tag{2.4}$$

$$= \text{squared bias + variance} \tag{2.5}$$

As can be seen from Equation 2.5, the MSE is considered to be an important criterion incorporating both systematic (bias) and random (variance) differences between the estimated and true values (Hand et al. (2001)).

Although the bias, variance, as well as the mean square error all play a role in describing how an estimate behaves, achieving honesty in the measures of uncertainty is also important. That is, reported standard errors, $SE(\hat{Q})$ should be as close to the true standard deviation of $\hat{Q}$ as possible (Schafer & Graham (2002)). A 95% confidence interval, for example, should not only cover the true value of Q but should also have a probability that is close to the nominal rate. According to Pal et al. (2006), the coverage probability is defined as the probability that the random interval includes or covers the true value of the parameter. When the nominal coverage probability equals the coverage probability, that is, when the coverage probability is accurate, the probability of a Type I error occurring will also be accurate. According to Banerjee et al. (2009), a Type I error occurs when the null hypothesis is rejected when it is in fact true and a Type II error occurs if one fails to reject the null hypothesis when it is actually false. Narrow confidence intervals are preferred as this helps to increase power in a study as well as helps to prevent the probability of a Type II error from occurring (Schafer & Graham (2002)).

When one has no control over the cause of missing values in a data set, assumptions about the methods that produced them need to be made. One often finds, however, that these assumptions cannot be tested (Schafer & Graham (2002)). Explicit assumptions are required and any results that are unclear and require further analysis should be reported. Researchers aim to draw similar conclusions when using several realistic but different assumptions.

2.2 Missing Data Patterns

A missing data pattern describes the pattern between the observed and missing data (Enders (2010)). These patterns represent the place where the "holes" in the data occur but they do not provide any explanations as to how missing values came about in the data set. On the other hand, missing data mechanisms refer to relationships between the probability of missing values and the observed data and they form a crucial role in Rubin's missing data theory.

As can be seen from Figure 2.1 below, there are six typical missing data patterns that one might come across when handling missing values. The areas that have been shaded indicate places where there are missing values. In panel A, a univariate pattern exists. This panel has missing values associated with a single variable, in this case Y_4 (Little (1992)). Univariate patterns exist in experimental designs. For example, let variables Y_1 to Y_3 be between subject factors in an ANOVA experiment and let Y_4 be the outcome variable that consists of incomplete data.

A unit non-response pattern exists in panel B. This pattern is very common in survey research. For example, Y_1 and Y_2 represent characteristics that are complete for every individual in a survey and Y_3 and Y_4 represent questions in the survey that some individuals may have declined to answer (Enders (2010)).

In panel C, a monotone missing data pattern exists. These patterns are usually linked to longitudinal studies where a participating individual decides to leave the study and never come back (Raghunathan (2004)). In order to illustrate this type of pattern, suppose new medication is being tested in a clinical trial and those individuals who react badly to the drug decide to leave and not return. Monotone missing data patterns are popular due to the fact that they can reduce the complexity of maximum likelihood and multiple imputation procedures.

Panel D represents the most common missing data pattern: the general missing data pattern. This pattern has random missing values occurring throughout the data set (Enders (2010)). Although the pattern is random, there may actually be relationships between the values and this is why it is important to keep in mind that missing data patterns only describe the places where the missing values exist and not the reasons why these missing values occur.

The next panel is known as a planned missing data pattern. This pattern represents a three form questionnaire plan where the aim is to allocate questionnaires across various forms and administer a subset of these forms to each participating individual. In panel E, the four questionnaires are allocated

across the three forms where each form includes Y_1 but not Y_2, Y_3 or Y_4 as they are missing from the data. According to Enders (2010), these data patterns are considered to be beneficial for gathering a large quantity of information, while at the same time reducing non-response issues.

The last missing data pattern in panel F is known as the latent variable pattern. This pattern is only used in structural equation models. According to Massell (2000), a latent variable cannot be observed or measured directly. These variables can, however, be inferred from other measured variables. As can be seen in this panel, the values of the latent variables are missing for the whole sample. Latent variables do not have to be regarded as missing data problems but some researchers have developed algorithms in order to estimate these models nonetheless (Enders (2010)). For example, multilevel models may be used.

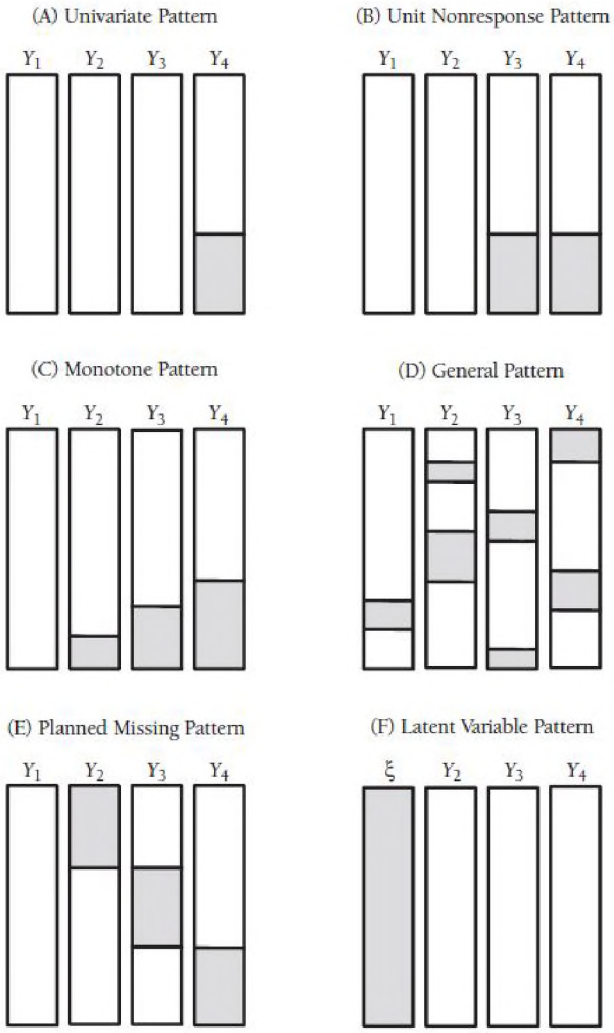

Figure 2.1: Missing data patterns where the shaded areas indicate the places where the missing values occur.

2.3 Missing Data Mechanisms

Baraldi & Enders (2010) state that Rubin (1976) and Little & Rubin (2002) developed a classification system for missing data problems that is widely cited. They discovered three missing data mechanisms that look at possible relationships that may exist between measured variables in a data set and the likelihood of missing data. These three data mechanisms include: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). When looking at these mechanisms from a practical point of view, they are assumptions that evaluate the performance of various missing data methods.

According to Luengo et al. (2012), data is considered to be MAR if the probability of an attribute having a missing value is related to other variables in the study but is independent of the missing. The term "random" may deceive some individuals since the missing values in the study are not random but depend on the observed variables in the analysis. In order to illustrate this, an example from Schafer & Graham (2002) is used, where Y_{com} represents a complete data set. This complete data set is divided into two parts, namely Y_{obs} and Y_{mis}, where Y_{obs} represents the observed data and Y_{mis} represents the missing data. Further, R is an indicator variable which represents the missingness in a data set. Schafer & Graham (2002) state that the more recent missing data methods regard missingness as a probabilistic phenomenon and they treat R as a random set of variables that have a joint probability distribution. They refer to the probability distribution for R as the "distribution of missingness" or the "probabilities of missingness". Then, using the framework proposed by Rubin (1976), they classify missing data as MAR if:

$$P(R \mid Y_{com}) = P(R \mid Y_{obs}) \tag{2.6}$$

Assumptions about the reasons for missing data for the MAR mechanism are not as strict as they are for the MCAR mechanism. For instance, MCAR data assumes that the underlying distribution of incomplete and complete data are identical, whereas for the MAR mechanism this is not the case (Luengo et al. (2012)). That is, the complete data set can be used to predict missing values when the data is MAR. According to Schafer & Graham (2002), when researchers cannot control whether missing values will occur in the data or not, MAR is only an assumption as the distribution of the missing data is not known. In other words, it is extremely difficult to test for MAR in a data set, unless one can get hold of follow-up data from those individuals who did not answer all of the questions given to them.

For an example of MAR, consider a study between men and women where they are required to record their weight. Women are less likely to report their weight and as such, the missing values depend on the variable gender. Alternatively, consider a scenario in which students may participate in an advanced maths course, provided they score above a certain percentage in an aptitude test. The marks they receive are considered to be MAR because any missing values will depend on the scores of the aptitude test (Baraldi & Enders (2010)). That is, those students that achieved scores below the required pass mark will not go on to participate in the advanced maths course and hence will not have scores recorded for this course. According to Donders et al. (2006), using techniques such as available or complete case analysis, as well as mean imputation on data that is MAR will yield biased results. Multiple imputation is preferred when missing values are classified as MAR.

According to Baraldi & Enders (2010), data is classified as MCAR when the probability of missing data on a variable X does not depend on other measured variables in the data set and is also not related to the values of X itself. That is, data that is classified as MCAR assumes that there is no relationship between the missing data and the other measured variables in the data set. Thus, according to Schafer & Graham (2002), using the framework proposed by Rubin (1976), data is considered to be MCAR when

$$P(R \mid Y_{com}) = P(R) \tag{2.7}$$

For example, consider an individual who participates in a study but halfway through the study the individual drops out because he/she had to relocate to another town. One would consider the missing values to be MCAR if the reason that the individual stopped participating in the study is not related to any of the other variables in the data set. A possible reason might be that the individual could no longer afford to stay in that particular area. Participants who flip a coin to decide whether or not to take part in a survey is another example of MCAR. According to Donders et al. (2006), since the MCAR mechanism assumes that the underlying distribution of incomplete and complete data are identical, complete and available case analyses will produce unbiased results. These methods, however, are less than ideal because some of the data is not being used in the analysis. Since the MCAR mechanism requires stringent assumptions, it is often argued that this assumption will not be satisfied in practice (Baraldi & Enders (2010)).

When observations are not classified as MCAR or MAR, they are MNAR. Data is classified as MNAR when the probability of a variable having a missing value does not depend on information that has been observed or the missing values themselves (Donders et al. (2006)). In other words, a relationship exists between the probability of missing values and the hypothetical values that are actually missing (Baraldi & Enders (2010)). This data mechanism is said to be the most challenging type since data that is MNAR cannot be ignored as it will bias the results significantly. A possible solution would be

to go back to the data source and gather more information about the attributes or alternatively, find a complete data set.

For an example of MNAR, consider a group of readers who, when completing a test, fail to answer certain questions due to a lack of understanding. In this case the likelihood of having a missing reading score is dependent on the level of understanding that the reader possess. As another example, consider high school students who have been asked to complete an alcohol assessment report. Data would be considered MNAR if those students who consume large amounts of alcohol are more likely to leave out certain questions because they may be afraid of the consequences (Baraldi & Enders (2010)). It is difficult to confirm that data is MNAR without knowing the values of the missing variables.

According to Enders (2010), only the MCAR mechanism can be tested as it is impossible to test for MAR and MNAR without knowing what the values of the missing variables are. Methodologists face difficulties with missing data analyses since techniques such as multiple imputation as well as maximum likelihood depend on the MAR mechanism. Several MCAR tests have been suggested by researchers but these tests result in a loss of statistical power and fail to detect deviations from a random mechanism. One technique that is useful for testing MCAR is to create dummy variables for the values that are missing. For example, a 1 represents a missing value and a 0 represents the observed data. Chi-squared tests or t-tests can be performed to test whether missing values are related to other variables in the study. It should be noted that these missing data mechanisms are not representative of the whole data set but are rather assumptions that are applicable to certain analyses. That is, depending on which variables were chosen to be included in the study, a data set is capable of producing analyses that are either MAR, MNAR, as well as MCAR.

Chapter 3

Approaches to Handling Missing Data

There are various methods which may be used to handle or treat missing values. According to Batista & Monard (2003), missing data techniques can be divided into three groups: (1) discarding or ignoring data (deletion methods), (2) parameter estimation methods, as well as (3) imputation techniques. Before these methods are discussed, Allison (2001) lists some criteria that can be used to evaluate missing data methods. In general, a missing data method that performs well should be able to do the following:

1. Minimise bias. Missing values may yield biased estimates and thus it is important to use a method that will minimise this bias as much as possible;

2. Avoid discarding any data. Ultimately, the aim is to maximise the use of all information and produce parameter estimates that are efficient;

3. Produce accurate estimates of standard errors, confidence intervals as well as p-values.

It would be ideal if all missing data methods could achieve the above criterion without violating any of the assumptions about the missing data mechanisms. Maximum likelihood as well as multiple imputation are capable of satisfying the above criteria, whereas the conventional methods have difficulty with accomplishing all three (Allison (2001)). What follows is a discussion on some of the more conventional missing data methods.

3.1 Deletion Methods

Complete case analysis, also commonly known as *listwise deletion*, can be used to remove cases with missing values so that the remainder of the analysis only contains cases with complete data (Zhu (2014)). The cases that are removed, however, may be highly relevant to the analysis and thus it is very important to carefully consider their relevance and influence before discarding them. One of the

advantages of using this method is that it is simple to implement and standard analysis techniques can be used as the data set is now complete.

There are, however, many disadvantages associated with this technique. According to Zhu (2014), in data sets where there are a large number of missing records, the total sample size will reduce significantly if all cases with missing values are removed. One will be left with a potentially very small and unrepresentative data set and as a result, there may be a loss of power in hypothesis testing. Listwise deletion also results in larger standard errors as well as wider confidence intervals.

Secondly, listwise deletion is usually only valid under MCAR (Baraldi & Enders (2010)). In fact, when the assumptions of MCAR are not met, the analysis may give biased results and invalid inferences because the complete cases do not represent the entire population. Only in a few situations has case deletion produced valid inferences under MAR. Also, according to Osborne (2013), in analyses that contain several variables, this method is not ideal. This is because small amounts of missing data for each variable may lead to getting rid of a large percentage of the sample, thus significantly impacting power in the analysis. Thus, if listwise deletion has to be used, the analysis should only contain a small amount of missing data and the data should be MCAR.

Alternatively, *pairwise deletion,* also commonly referred to as *available case analysis* is another technique that may be used to treat missing data. According to Osborne (2013), when using pairwise deletion, one only removes cases that have missing data when calculating a specific variable. As such, cases will contribute to some analyses but not to others. For example, in a correlation matrix, each correlation will be estimated based on those cases that have data for both variables. Pairwise deletion is considered to be a better technique than listwise deletion because unlike listwise deletion, this method keeps as many cases as possible for each analysis, thus minimising the number of cases that are removed (Baraldi & Enders (2010)).

A disadvantage of using pairwise deletion is that different analyses will be based on different subsets of data and will not necessarily be consistent with each other, thus resulting in biased parameter estimates (Graham (2009)). Also, according to Allison (2001), pairwise deletion has difficulty calculating accurate standard errors. This is because in order to estimate standard errors, the sample size needs to be identified which is not always possible. As such, when each covariance is calculated, it will be based on various sample sizes, depending of course on the missing data pattern. Moreover, this technique does not work in cases where the estimated correlation matrix is not positive definite (a symmetric matrix where all the eigenvalues are positive), as not being able to invert the matrix results in inaccurate parameters (Allison (2001)). This technique also yields biased estimates if the data is not

MCAR. In fact, just like listwise deletion, this method does not work under any other data mechanism (Russell et al. (2001)).

Since listwise and pairwise deletion require stringent assumptions, it is unlikely that they will be satisfied in practice and as a result, these methods are being used less often especially due to the development of more modern techniques.

3.2 Single Imputation Techniques

Instead of discarding all missing values in a data set, it may be tempting to rather replace all missing data with plausible values. According to Batista & Monard (2003), imputation is a technique where researchers use suitable replacement values in order to substitute missing values. These plausible values are based on information that is available in the data set. The objective is to determine relationships that can be identified in the values of the data set in order to estimate missing values.

According to Dempster & Rubin (1983), "the idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases."

Single imputation is preferred over deletion techniques as none of the data is sacrificed. According to Schafer & Graham (2002), if the entire sample is preserved, a situation resulting in a loss of power due to a reduced sample size can be avoided. Also, imputation techniques are able to make use of valuable information obtained from the observed data in order to predict missing values - thus maintaining high precision in the data set. Imputation methods also provide the analyst with a complete data set where standard methods may be used to analyse the data. Further, imputing values once is an advantage as if there are many different individuals using the data, the same sets of units will be considered by each individual, ensuring consistency when comparing results.

Single imputation works fairly well when one is dealing with data sets that have a small number of missing values but there are many disadvantages associated with this method. Firstly, the standard errors of estimates is greatly underestimated (He (2010)). This is because using a single value to replace the missing values does not account for the variation that would most probably have been present had the variables been observed. Also, according to Pigott (2001), smaller standard errors resulting from an increased sample size fail to take into account the uncertainty that is present in the data set.

Schafer & Graham (2002) state that it can also be tricky to use imputation techniques in cases with multivariate data. This is because some single imputation methods can actually alter data distributions as well as relationships. Allison (2001) does not believe using single imputation techniques is a solution to handling missing data as "all conventional imputation methods are dishonest and should be viewed with some scepticism." The various single imputation techniques will now be discussed.

3.2.1 Mean / Mode Imputation

As an alternative to using listwise and pairwise deletion, researchers can replace missing values by using either the mean or the mode. According to Acuña & Rodriguez (2004), *mode imputation* can be used for categorical data. This technique consists of using the value that occurs most frequently to replace missing data for each variable. One of the advantages of mode imputation is that one is left with a complete data set where complete data methods can be used.

However, according to Munguía & Armando (2014), mode imputation should not be used as it reduces the variability in a data set. That is, if all of the imputed values are concentrated in the mode, this will alter the shape of the distribution as spikes are often created. Further, correlation estimates may be adversely affected if any relationships that occur within the data set are ignored.

Mean imputation is a technique that involves replacing all missing values with the mean of all known values of that variable (Acuña & Rodriguez (2004)). In cases where the overall mean is used, the mean is taken from the entire distribution and used to replace missing values. Consider Table 3.1 below where the variable X_j has missing values:

Table 3.1: An example of a variable X_j, where X_j has missing values.

		X_j
case	1	X_{1j}
	2	X_{2j}
	3	missing
	...	...
	i	missing
	...	...
	n	X_{nj}

All missing values will be replaced by:

$$\frac{\sum_{r=1}^{n} x_{rj} I_{cj}\left(x_{rj}\right)}{\sum_{k=1}^{n} I_{cj}\left(x_{rj}\right)} \tag{3.1}$$

$$= \frac{\sum_{r=1}^{n} x_{rj} I_{cj}\left(x_{rj}\right)}{n_j} \tag{3.2}$$

where $C_j = \{x_{kj} | x_{kj} \text{ not missing}, k = 1, 2, \ldots, n\}$.

An advantage of using mean imputation is that it preserves the sample size. Further, standard analysis techniques may be used for the data analysis as the data set no longer contains missing values (Zhu (2014)).

According to Peng et al. (2006), there are, however, several disadvantages with mean imputation. Firstly, the sample size is often overestimated. Secondly, mean imputation underestimates the variability among the missing values because the same value is being substituted for each missing item, which in turn affects the significance of any statistical test based on it. Also, the correlation between variables is biased because relationships between variables are not taken into account. Peng et al. (2006) also state that using this technique may result in the distribution of new values not being a true representation of the population values. This is because the shape of the distribution may alter when values that are equal to the mean are added. Based on these limitations, they recommend never to use mean imputation.

3.2.2 Median Imputation

According to Acuña & Rodriguez (2004), *median imputation* is a technique that involves replacing an attribute that contains missing data items with the median of all known values of that attribute. That is, all missing data items are substituted with the $50th$ percentile. This percentile is either the middle or the average of the two middle values after sorting the data in ascending or descending order. The missing data points will thus be replaced by:

$$median\, x_j = med\left(C_j\right) \tag{3.3}$$

where $C_j = \{x_{kj} | x_{kj} \text{ not missing}, k = 1, 2, \ldots, n\}$.

Acuña & Rodriguez (2004) prefer to use this technique over mean imputation as the mean can often be plagued by outliers. There are, however, disadvantages associated with this method. Median imputation fails to take into account the correlation structure of the data which may lead to weak estimates as well as a weak covariance. In fact, median imputation has the same limitations as mean imputation and hence this method is also not recommended.

3.2.3 Overview of Linear Regression

Regression analysis is a popular tool used in Statistics. It involves using several methods for modelling and analysing variables and also explores possible relationships between dependent variables, also known as response variables, as well as independent variables which are also known as covariates or regressors (Walli (2010)). When a standard normal linear regression model is observed, an assumption is made about the mean of the response variable, that is, it is a linear function of the regressors. Hence, the term 'linear model' arises from the fact that the regressors are linear in the parameters. According to Hand et al. (2001), a simple linear model will produce predicted values $\hat{y}$, of the response variable y, and these are a linear combination of the predictor variables x_j. This can be expressed in Equation 3.4 below:

$$\hat{y} = a_0 + \sum_{j=1}^{p} a_j x_j \tag{3.4}$$

The term 'predictor variables' defines the inputs used for prediction and the term 'response variable' describes the variable that will be predicted. In most cases, it is highly unlikely that one can predict the response variable accurately and so interest lies in predicting the mean value that y takes at each vector of the prediction variables (Hand et al. (2001)). That is, $\hat{y}$ is the predicted estimate of the mean at $\mathbf{x} = (x_1, ...x_p)$.

As mentioned above, it is unlikely that the model that is chosen is perfect. In data mining scenarios, the models are usually empirical and are not based on an underlying theory. That is, in any sample, the actual y values will not be the same as the predicted values. According to Hand et al. (2001), the term residuals describe the difference between the observed and the predicted values and may be denoted by e:

$$y(i) = \hat{y}(i) + e(i) = a_0 + \sum_{j=1}^{p} a_j x_j(i) + e(i), \quad 1 \leq i \leq n \tag{3.5}$$

Using matrix notation, the observed y measurements on the n objects are represented by the vector $\mathbf{y}$ and in the same way, the p measurements of the predictor variables on the n objects are denoted by the $n \times (p+1)$ matrix $\mathbf{X}$ (Hand et al. (2001)). Here, an extra column of ones is added in order for the model to include an intercept, in this case, a_0. The relationship between the observed response and predictor measurements can be expressed in the form below:

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e} \tag{3.6}$$

where $\mathbf{y}$ represents an $n \times 1$ matrix of response values, $\mathbf{a} = (a_0, ..., a_p)$ denotes the $(p+1) \times 1$ vector of parameter values, and the $n \times 1$ vector $\mathbf{e} = (e_1, ..., e_n)$ comprises the residuals.

Ideally, one wants to choose the parameters in a model that will result in the most accurate predictions. That is, estimates for the a_j need to be explored in such a way that the error term is minimised. To accomplish this, the elements in e need to be combined so that a single numerical measure is produced and then minimised. According to Searle (1971), the method that is most commonly used today is to sum the squares. That is, we want to find the values for the parameter vector $\mathbf{a}$ that minimises:

$$\sum_{i=1}^{n} e(i)^2 = \sum_{i=1}^{n} (y(i) - \hat{y}_i)^2 \tag{3.7}$$

where $y(i)$ denotes the observed y value for the ith sample point. This approach is called the least squares method. The parameter vector that minimises the above is represented by $(a_0, ..., a_p)$. Walli (2010) argues that when using a matrix, the values of the parameters, (assuming that $\mathbf{X}^T\mathbf{X}$ is of full rank), that minimise Equation 3.7 can be expressed as:

$$\mathbf{a} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{3.8}$$

Generally, in linear regression, the $p+1$ parameters in $\mathbf{a}$ are known as the regression coefficients. The estimated parameters are used in the equation above in order to yield predictions. According to Hand et al. (2001), the predicted value of $y(k)$, $\widehat{y}(k)$, for a vector of predictor variables $\mathbf{x}_k$ can be expressed as follows:

$$\mathbf{x}_k^T = (x_0(k), x_1(k), \ldots, x_p(k)) \tag{3.9}$$

Difficulties may occur if one has a sample size that is too small, or when linear dependencies between the measured values of the predictor variables exist. Further, the matrix $\mathbf{X}^T\mathbf{X}$ must be invertible.

3.2.3.1 Regression Imputation

Regression imputation is a process that generates a predictive model in order to estimate values that will replace any missing items that exist within the data set (Batista & Monard (2003)). In cases where missing data exists on a single variable, complete case analysis can be used to estimate the regression equation. The variables that have incomplete data are used as the outcome variables and the variables with complete data are used as the predictor variables. One is now left with a regression equation that produces predicted scores for the missing items.

One of the strengths of using this approach, is that when variables are correlated with each other or are related in some way, these relationships or correlations can be used to build predictive models for regression or classification (Batista & Monard (2003)). If the predicted model does a good job at capturing relationships between variables in the data, these relationships can be preserved. Another

advantage of using regression imputation is that the sample size is preserved since variables with missing values are not being removed from the analysis (Peng et al. (2006)).

According to Peng et al. (2006), regression imputation has its disadvantages. Firstly, a regression model must be specified. Secondly, this technique may yield unstable estimates if the sample size is not large enough. Also, in analyses that contain multivariate data, regression imputation may be problematic. Peng et al. (2006) state that if the data set fails to contain relevant predictors of missing data, then these predicted values will not be any better than the mean. That is, regression imputation and mean imputation produce nearly the same results if an efficient regression model is not used.

Further, according to Baraldi & Enders (2010), regression imputation yields biased estimates. This is because regression imputation does not take into account the variability that exists between the hypothetical values and, as a result, the variance will be underestimated. A possible way to restore lost variability and eliminate biased estimates is to use stochastic regression imputation, where a stochastic component is added to each predicted score. This random error is generated from a normal distribution with a zero mean and a variance that is the same as the residual variance from the previous regression analysis.

Regression imputation also requires correlation among the attributes in the data set. If correlation does not exist and there are no relationships between the variables in the data set and the variables with missing items, then the model will not estimate these missing values accurately (Acuña & Rodriguez (2004)). Computational cost is yet another disadvantage since in order to predict the missing values, a large number of models may have to be built.

3.2.4 *K*-Nearest Neighbour Imputation (KNNI)

According to Hand et al. (2001), nearest neighbour methods are simple. That is, in order to classify a new object with input vector $\mathbf{y}$, the k-closest data set points to $\mathbf{y}$ are observed and the object is assigned to the class that has the bulk of the points among these k. In other words, the idea is to look for those objects in the data set that are very alike the new object based on the input variables and then categorise the new object into the class that has the majority of points among these most similar objects.

From a theoretical point of view, this approach takes a small volume of the space of variables that are centered around x that have a radius equal to the distance to the k-nearest neighbours. The probability that a point in this small space belongs to each class is calculated by taking the proportion of points in this volume that belongs to each class (Hand et al. (2001)). The class that has the largest esti-

mated probability is assigned a new point. Hand et al. (2001) states that nearest neighbour methods are similar to regression methods as they directly estimate the posterior probabilities of class membership.

Each time a missing value exists for a current attribute, the *k-nearest neighbour imputation* technique calculates the k-nearest neighbours and imputes a value for each missing case (Jönsson & Wohlin (2004)). When these attributes take on nominal values, the value that occurs most frequently amongst all neighbours is used. In the case of numerical values, the average value amongst all neighbours is used. A proximity measure between the attributes is required and the Euclidean distance is one of the most commonly used distance metrics. According to Jönsson & Wohlin (2004), the Euclidean distance is defined as:

$$d_E(a,b) = \left(\sum_{i \in D} (x_{ai} - x_{bi})^2 \right)^{\frac{1}{2}} \tag{3.10}$$

In the above equation, $d_E(a,b)$ represents the distance between the two cases a and b, D represents the set of attributes that do not have missing values and x_{ai} and x_{bi} represent the values of attribute i in cases a and b respectively.

According to Acuña & Rodriguez (2004), the KNNI algorithm can be summarised into the following steps:

1. The data set D must be divided into two parts say D_m and D_c. D_m represents the set of attributes where there are missing items and D_c represents the remaining attributes that have complete data.

2. Next, for every vector $\mathbf{x}$ in the data set D_m, the instance vector must be divided into two groups: namely observed and missing as follows: $\mathbf{x} = [\mathbf{x_0}; \mathbf{x_m}]$. Here, $\mathbf{x_0}$ represents the observed groups and $\mathbf{x_m}$ denotes the missing groups. Secondly, the distance between $\mathbf{x_0}$ and the rest of the instance vectors from set D_c must be computed. Only the features in the instance vectors from the set D_c which appear in the vector $\mathbf{x}$ should be used. Once the k-nearest neighbours have been determined, a replacement value needs to be estimated that will replace the missing attribute. The calculation of the replacement value depends on the type of data. As already mentioned, one can use the mode for categorical data and the mean for nominal data. The median is preferred over the mode because there may be cases where a lot of values have the same frequency. An important parameter for this method is the value of k. As the value of k increases, the mean distance to the neighbours increases. This implies that the replacement values may not be very accurate. As k approaches N, the number of neighbours, the method will simply converge to mean imputation.

According to Batista & Monard (2003), one of the advantages of using KNNI methods is that they are capable of predicting both qualitative and quantitative attributes. Further, this technique does not require a predictive model to be built for each attribute that contains missing items. In fact, no explicit models such as decision trees or a set of rules are required and the algorithm can be easily adjusted to work with any attribute. This can be achieved by adjusting the considered attributes in the distance metric.

Another advantage of using the KNNI technique is that it is able to handle attributes that have multiple missing values. This method also takes into account the correlation structure of the data (Acuña & Rodriguez (2004)). According to Jönsson & Wohlin (2004), KNNI methods do not encounter problems with the variance reducing, as is the case with mean imputation. Whereas mean imputation imputes the same value, KNNI imputes various different values depending of course on the case being imputed.

According to Acuña & Rodriguez (2004), there are many disadvantages of using the KNNI technique. Firstly, a decision needs to be made with regards to which distance metric will be used. These metrics may include Euclidean, Manhattan, Mahalanobis, Pearson as well as others. Secondly, the KNNI algorithm has to search through the entire data set in order to find the most similar cases. The data set may be large and thus trying to search through each point to find the k-nearest neighbours can take up a lot of time. Thirdly, deciding on k, the number of neighbours that will be used may be tricky. Choosing a small k results in a poor performance of the classifier after imputation. However, choosing a large k includes cases that are significantly different from the cases that have missing values, leading once again to a poor performance of the classifier. In small data sets, choosing a k that is smaller than 10 is satisfactory.

Dimensionality is another issue to consider when using the KNNI technique. Xie (2012) states that increasing dimensions may lead to difficulty when using the distance function. Locating the necessary storage space may also be tricky.

3.3 Maximum Likelihood (ML)

According to Enders (2001), researchers are becoming more aware of the theoretical advantages of using *maximum likelihood* estimation and simulation evidence suggests that maximum likelihood produces far better results over the conventional methods discussed thus far. According to Schafer & Graham (2002), maximum likelihood is dependent on the MAR assumption, however, in real life applications an analysis will still produce valid results in situations where MAR does not hold.

Myung (2003) states that estimates with attractive properties are produced by maximum likelihood only if all of the assumptions, in particular, sufficiency, consistency, asymptotic efficiency as well as asymptotic normality are satisfied.

Firstly, the sample size needs to be large enough to ensure consistency so that the parameter estimates are unbiased and normally distributed. Missing values in an analysis may reduce the sample size significantly and hence it may have to be larger than usual (Russell et al. (2001)). Asymptotic efficiency implies that the estimates that are produced have minimal standard errors and are thus close to being fully efficient (Allison (2001)). Further, asymptotic normality implies that a normal approximation can be used in order to compute confidence intervals as well as p-values.

In order to proceed with maximum likelihood estimation, a likelihood function is required where the probability of the data is expressed as a function of the unknown parameters. For example, Allison (2001) considers a situation where discrete variables X and Z have a joint probability function given by $p(x,z|\theta)$ and where θ represents a vector of parameters. In other words, $p(x,z|\theta)$ is the probability that $X = x$ and $Z = z$. If no missing values exist within the data set and the observations are independent, the likelihood function may be expressed as follows:

$$L(\theta) = \prod_{i=1}^{n} p(x_i, z_i|\theta) \tag{3.11}$$

According to Agresti (2002), in order to calculate maximum likelihood estimates, the value of θ that optimises the above function needs to be determined. Next, assume that the data is MAR on Z for the first r cases, and MAR on X for the next s cases. Let:

$$g(x|\theta) = \sum_{z} p(x,z|\theta) \tag{3.12}$$

be the marginal distribution of X (summing over Z) and let:

$$h(z|\theta) = \sum_{x} p(x,z|\theta) \tag{3.13}$$

be the marginal distribution of Z (summing over X). The likelihood function is thus given by:

$$L(\theta) = \prod_{i=1}^{r} g(x_i|\theta) \prod_{i=r+1}^{r+s} h(z_i|\theta) \prod_{i=r+s+1}^{n} p(x_i, z_i|\theta) \tag{3.14}$$

From Equation 3.14 above, it can be seen that the likelihood has been divided into parts that correspond with different missing data patterns. In order to calculate the likelihood for each pattern, the

joint distribution needs to be summed over all possible values of the variables that have missing data (Allison (2001)). In the case where variables are continuous and not discrete, integral signs will replace the summation signs.

In order to apply maximum likelihood methods to data sets that contain missing values, a model for the joint distribution of all the necessary variables is required, as well as a numerical method that maximises the likelihood (Allison (2001)). In an analysis with only categorical variables, the unrestricted multinomial model or a log-linear model that contains some limitations on the data may be suitable. A log-linear model is only required in analyses where many variables with several categories exist. If, for example, these restrictions were not in place, several parameters would need to be estimated.

In the case of continuous variables, a multivariate model can be assumed where the variables are normally distributed and are linear functions of the other variables (Allison (2001)). Each of these variables will have errors that are homoscedastic as well as a zero mean. According to Pigott (2001), when the multivariate normal model is used, the likelihood can be maximised by using the expectation-maximisation (EM) algorithm. This algorithm is discussed next.

3.3.1 The Expectation-Maximisation (EM) Algorithm

According to Dempster et al. (1977), the aim of the *expectation-maximisation* algorithm is to maximise the likelihood under several missing data models. This method involves an iterative algorithm that moves through two stages, namely the expectation stage (the E-step) and the maximisation stage (the M-step).

The E-step consists of taking all missing observations that exist within the data set and replacing them with the conditional expectation of the missing data given the observed data and an initial estimate of the covariance matrix (Enders (2001)). In other words, the predicted scores from a set of regression equations are used to replace each of the missing values occurring in the data set. In this case, every missing value is regressed on the remaining observed values for a case i. Then, in order to determine the sums of squares as well as the cross products, the observed and imputed values should be used.

In order to gain a clearer understanding, an example from Enders (2001) is used where one is interested in calculating a mean vector and covariance matrix, $\theta = (\mu, \Sigma)$ for an $n \times K$ matrix, Y. This data matrix should consist of sets of observed values (Y_{obs}) and missing values (Y_{mis}), as well as current parameter estimates, $\theta^{(t)}$. The calculations for the sufficient statistics at the ith iteration of the E-step

are given below:

$$E\left(\sum_{i=1}^{n} y_{ij}|Y_{obs},\theta^{(t)}\right) = \sum_{i=1}^{n} y_{ij}^{(t)} \qquad j = 1,...,K \tag{3.15}$$

and

$$E\left(\sum_{i=1}^{n} y_{ij}y_{ik}|Y_{obs},\theta^{(t)}\right) = \sum_{i=1}^{n}\left(y_{ij}^{(t)}y_{ik}^{(t)} + c_{jkl}^{(t)}\right) \qquad j,k = 1,...,K \tag{3.16}$$

where

$$y_{ij}^{(t)} = \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed} \\ E\left(y_{ij}|Y_{obs},\theta^{(t)}\right) & \text{if } y_{ij} \text{ is missing} \end{cases} \tag{3.17}$$

and

$$c_{jkl}^{(t)} = \begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed} \\ cov\left(y_{ij},y_{ik}|Y_{obs},\theta^{(t)}\right) & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing} \end{cases} \tag{3.18}$$

Hence, given the observed data as well as the current sets of parameter estimates, the conditional means and covariances are used to replace the missing values of y_{ij}.

In the M-step, one can calculate maximum likelihood estimates of the mean vector and covariance matrix in exactly the same way as if the data set were complete. This is done by using the sufficient statistics that were computed in the E-step mentioned above. Hence, according to Enders (2001), the M-step can also be referred to as a complete data maximum likelihood estimation problem. Next, using the covariance matrix as well as the regression coefficients obtained from the M-step, new estimates can be calculated for each of the missing observations at the following E-step. This algorithm will then start again and will continue to loop through both the E- and M-steps and will only stop once the difference between the covariance matrices in the following M-steps converge according to certain criteria that has been set (Enders (2001)).

Consider another example which looks at the following matrix where m_1 and m_2 represent missing values:

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{21} & y_{22} & m_1 & y_{24} \\ y_{31} & y_{32} & y_{33} & y_{34} \\ y_{41} & y_{42} & y_{43} & y_{44} \\ y_{51} & m_2 & y_{53} & y_{54} \\ y_{61} & y_{62} & y_{63} & y_{64} \end{bmatrix}$$

Suppose we want to find $\theta = (\mu, \Sigma)$ where

$$\mu = \text{mean vector} \tag{3.19}$$

and

$$\Sigma = \text{covariance matrix} \tag{3.20}$$

That is,

$$\mu^{'} = [E(Y_1), E(Y_2), ..., E(Y_4)] \tag{3.21}$$

and

$$\Sigma = \begin{bmatrix} Var(X_1) & ..... & ..... & Cov(X_1, X_4) \\ ..... & ..... & ..... & ..... \\ ..... & ..... & ..... & ..... \\ Cov(X_4, X_1) & ..... & ..... & Var(X_4) \end{bmatrix} \tag{3.22}$$

Next, replace m_1 with starting value Γ_1^0 where

$$\Gamma_1^0 = \frac{1}{5}(y_{13} + y_{33} + y_{43} + y_{53} + y_{63}) \tag{3.23}$$

and replace m_2 with starting value Γ_2^0 where

$$\Gamma_2^0 = \frac{1}{5}(y_{12} + y_{22} + y_{32} + y_{42} + y_{62}) \tag{3.24}$$

Then, the E-step is such that:

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{21} & y_{22} & \Gamma_1^0 & y_{24} \\ y_{31} & y_{32} & y_{33} & y_{34} \\ y_{41} & y_{42} & y_{43} & y_{44} \\ y_{51} & \Gamma_2^0 & y_{53} & y_{54} \\ y_{61} & y_{62} & y_{63} & y_{64} \end{bmatrix}$$

and the M-step contains the following equations:

$$\mu^{0'} = \left[\frac{1}{6} \sum_{k=1}^{6} y_{k1} + ... \frac{1}{6} \sum_{i=1}^{6} y_{k4} \right] \tag{3.25}$$

$$= \left[\bar{Y}_1^{(0)} ..... \bar{Y}_4^{(0)} \right] \tag{3.26}$$

and

$$Cov^0 \left(X_i, X_j \right) = \frac{1}{6} \sum_{k=1}^{6} y_{ki} y_{kj} - \bar{Y}_i^{(0)} \bar{Y}_j^{(0)} \tag{3.27}$$

Next, the "complete" matrix should be used in order to find the following regression equations:

The E-step is as follows:

$$\hat{Y}_3 = a_{03}^{(1)} + a_{13}^{(1)} Y_1 + a_{23}^{(1)} Y_2 + a_{43}^{(1)} Y_4 \tag{3.28}$$

$$\hat{Y}_2 = a_{02}^{(1)} + a_{12}^{(1)} Y_1 + a_{32}^{(1)} Y_3 + a_{42}^{(1)} Y_4 \tag{3.29}$$

Then

$$\Gamma_1^{(1)} = a_{03}^{(1)} + a_{13}^{(1)} y_{21} + a_{23}^{(1)} y_{22} + a_{43}^{(1)} y_{24} \tag{3.30}$$

and

$$\Gamma_2^{(1)} = a_{02}^{(1)} + a_{12}^{(1)} y_{51} + a_{32}^{(1)} y_{53} + a_{42}^{(1)} y_{54} \tag{3.31}$$

The M-step is as follows:

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{21} & y_{22} & \Gamma_1^1 & y_{24} \\ y_{31} & y_{32} & y_{33} & y_{34} \\ y_{41} & y_{42} & y_{43} & y_{44} \\ y_{51} & \Gamma_2^1 & y_{53} & y_{54} \\ y_{61} & y_{62} & y_{63} & y_{64} \end{bmatrix}$$

Continuing in this way one is able to determine values for μ^1, $Cov^1\left(X_i, X_j\right)$, Σ^1 etc. The algorithm will stop once convergence has been reached.

This algorithm can also be described as iterated linear regression imputation and Allison (2001) suggests summarising it into the following steps:

1. Firstly, one can make use of listwise or pairwise deletion in order to obtain starting values for the means, variances and covariances.

2. Next, for every missing data pattern, regression equations should be built in order to predict the missing variables based on the observed variables.

3. In order to calculate predicted values for each variable that contains missing values, the above regression equations should be used.

4. The means, variances and covariances need to be re-calculated using both the real and the imputed data, where the standard mean formula may be used to determine the mean. However, in order to calculate the variance, one should use a correction factor. The purpose of the correction factor is to account for any bias that may occur when the imputed values are used.

5. Lastly, return to step two and continue to cycle through these steps until convergence is reached.

According to Enders (2001), it is not appropriate to use the EM algorithm to obtain direct estimates of linear model parameters (e.g. regression). Only maximum likelihood estimates of a mean vector and covariance matrix can be obtained when using this algorithm.

It is, however, possible to use the covariance matrix in one of two ways: (1) as an input into further linear model analyses, or (2) using this matrix in order to replace or estimate missing values at the last iteration. Enders (2001) states that although the data set appears to be complete, it is not a good idea to use the covariance matrix to impute or estimate the missing values. Even though the filled-in values may be the most ideal estimates for the missing values, they fail to capture the residual variability that would have existed in the hypothetically complete data set. Since these filled-in values fall directly on a regression line, they are not imputed with a random error term and as such, standard errors obtained from any further analyses will be biased. Techniques such as bootstrapping should be used in order to obtain accurate estimates. According to Enders (2001), in order to remove the bias that results in the covariance matrix, a correction factor, $c_{jkl}^{(t)}$ may be added to the conditional expectation of the missing data at each E-step - as can be seen in Equation 3.18 above.

According to Peng et al. (2006), one of the downsides of using the EM algorithm is that it is iterative and specific to the model being applied. Further, the rate of convergence may be very slow if the percentage of missing values in the data set is high, as the convergence rate during iterations is proportional to the amount of observed data that exists within the data set. Also, the EM algorithm can only be used for linear and log-linear models (Soley-Bori (2013)). Allison (2001) actually prefers the direct maximum likelihood method over the EM algorithm as this method is capable of yielding accurate standard errors and is more suitable for "overidentified" models.

The EM algorithm is easily available in many software packages. According to Schafer & Graham (2002), an EM algorithm of an unstructured covariance matrix is available in BMDP Statistical software, as well as SPSS. The algorithm can also be performed by using other software such as EMCOV, NORM, SAS, SPLUS, LISREL and MPLUS. For normal models with structured covariance matrices, multilevel linear models can be fit using either HLM, MLWin, the PROC MIXED procedure in SAS or the lme function in SPLUS. It is important to note that these programs will assume MAR under all situations except when missing values occur by design.

According to Collins et al. (2001), because of the wide range of software available nowadays, fitting a model that contains missing values is not very difficult. These programs are often quite simple to use and as such, there is a misunderstanding that maximum likelihood techniques do not require the user to think about the missing data problem as any adjustments that need to be made for the missing values have already been performed automatically. However, it is important to keep in mind that these automatic adjustments are only satisfactory under specific assumptions, in particular when the data is MAR. If any important causes or correlates of missingness are left out of the model, one runs the risk of having biased maximum likelihood estimates. Many researchers are guilty of not considering including auxiliary variables that may be relevant to the missing data, although many of the ML methods do not make it clear on how to go about this.

3.4 Multiple Imputation (MI)

Even though maximum likelihood has been described as a very good technique for treating missing data, it is important to remember that one of the biggest flaws of this method is that it is imperative to specify a joint distribution for all the variables and these models are often difficult to identify. An alternative technique such as *multiple imputation* may be used. According to Allison (2001), multiple imputation is similar to the maximum likelihood method as the statistical properties are almost identical. That is, multiple imputation is also capable of producing consistent estimates that are asymptotically normal and asymptotically efficient. The major difference between multiple imputation and

maximum likelihood, however, is that with the multiple imputation technique, the aim is to determine estimates of the missing values, whereas with maximum likelihood, the aim is to obtain expected values of the sufficient statistics (Pigott (2001)). What follows is an analysis of multiple imputation and its use in handling missing data.

According to Russell et al. (2001), multiple imputation was first introduced by Rubin (1978) and improved upon in the context of large surveys, where data obtained in a particular study was to be used by many researchers for various analyses. At first, multiple imputation was not considered to be a very popular technique due to a shortage of computational facilities. This has of course changed and multiple imputation has now emerged as one of the more popular methods of handling missing data. This is due to the ubiquity of computers and the powerful technology that is now available.

According to Allison (2001), the Markov Chain Monte Carlo (MCMC) algorithm based on the multivariate normal model is a very prevalent methodology in the context of multiple imputation. That is, MI is based on a Markov chain consisting of independent draws from

$$Y_{mis} \sim P(Y_{mis}|Y_{obs}) \tag{3.32}$$

In practice, this distribution is not easy to draw from and hence an approximation may be used:

$$Y_{mis}^{(t)} \sim P\left(Y_{mis}|Y_{obs}, \theta^{(t)}\right) \tag{3.33}$$

Multiple imputation has three important stages, namely the 'imputation stage', the 'analysis stage' and the 'pooling/combining stage' (Horton & Lipsitz (2001)). Firstly, the imputation phase is responsible for creating a certain number of data sets. Each of these data sets should contain different estimates of the missing items. According to Raghunathan (2004), more than one set of possible replacement values should be created. That is, m complete data sets should be produced. Theoretically, in order to ensure that multiple imputation techniques are fully efficient, an infinite number of repetitions are required (Berglund & Heeringa (2014)). However, if one were to compare the loss of efficiency using a smaller, finite number of repetitions (e.g. when $m = 3, 5, 10, 20$) compared to an infinite number of repetitions, one would see that this loss is actually relatively small.

According to Schafer & Olsen (1998), a measure of relative efficiency is approximately given by:

$$RE = \left(1 + \frac{\gamma}{m}\right)^{-1} \tag{3.34}$$

where γ represents the fraction of missing data and m represents the number of multiple imputation repetitions. Table 3.2 below shows the relative efficiencies with different values for m and γ. It can be seen that when there are cases with very few missing values, only a small number of imputations will be necessary (Inc. (2008a)).

Table 3.2: Table displaying relative efficiency.

	γ				
m	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

Historically, it was common practice to use $m = 5$. Recent research, however, suggests using cases where m is larger than 5 as this results in good nominal coverage for confidence intervals as well as better nominal power levels in hypothesis testing (Berglund & Heeringa (2014)).

According to Baraldi & Enders (2010), a common approach used during the imputation stage is the Data Augmentation (DA) procedure for normally distributed data. This algorithm obtains multiple imputations from the Markov chain with draws from

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis}|Y_{obs}, \theta^{(t)}\right) \tag{3.35}$$

where

$$\theta^{(t+1)} \sim P\left(\theta|Y_{obs}, Y_{mis}^{(t+1)}\right) \tag{3.36}$$

In this procedure, data sets are imputed using a two-step iterative procedure. The first step, which is also referred to as the *I*-step, uses the same procedure as stochastic regression imputation. That is, regression equations are created using estimates of the means and covariances. These equations are then used to predict any incomplete values using the variables that contain complete data. Further, these equations are also responsible for generating predicted scores for each of the missing variables. Then, in order to preserve the variability in the data set, a normally distributed residual term is added to each value that has been predicted. The imputed data is then used in the posterior step, which is also known as the *P*-step. In order to create brand new estimates of the means and covariances, this step makes use of Bayesian estimation principles.

In the *P*-step, the means and the covariances are calculated from the filled-in data set and a random error term is added to each of the ensuing estimates (Baraldi & Enders (2010)). A completely new set of parameter values are produced. These values are not the same as those values that were used to produce the imputed values in the earlier *I*-step. According to Baraldi & Enders (2010), using these new parameter values, one can build a new set of regression equations in the following *I*-step, which will ultimately result in the production of a new set of imputations. The values of these imputations are also not the same as the values produced in the preceding *I*-step. When these two steps are continuously repeated, multiple versions of the data sets are created, where each data set contains different estimates of the missing items.

One must keep in mind that the resulting imputed values are not obtained from consecutive *I*-steps. That is, in multiple imputation the imputed values resulting from one data set should be independent of the imputed values obtained from other data sets (Baraldi & Enders (2010)). In order to achieve this, before the first data set is saved, a significant number of iterations should lapse before the next data set is saved. That is, if for example the researcher decided to let 100 data augmentation cycles lapse between each of the data sets, the procedure would then run for 100 iterations and the first data set will be saved. Then, the algorithm would run for another 100 iterations and the second data set would be saved and so on.

The next stage that follows is the analysis phase. Here, an analysis is carried out on each of the m imputed complete data sets and the parameter estimates and standard errors are stored (Horton & Lipsitz (2001)). This stage uses standard statistical procedures in order to analyse the m complete data sets.

The last phase which shall be discussed is the pooling phase, in which estimates and standard errors are combined into one set of values. There are various methods that have been created in order to carry out the pooling phase. Baraldi & Enders (2010) suggest, that in order to determine pooled parameter estimates, one may use the arithmetic mean of the estimates from each of the data sets. Pooling standard errors may become quite tricky as it involves two parts. Firstly, the standard errors from the imputed data sets are used, which is also known as the within-imputation variance. Secondly, a component is added that accounts for the variation of the estimates across the data sets. This is known as the between-imputation variance.

The within-imputation variance, denoted by W, is the average of the squared standard errors and is given in Equation 3.37 below:

$$W = \frac{1}{m}\sum_{t=1}^{m} SE_t^2 \tag{3.37}$$

Here, an imputed data set is represented by t and m refers to the total number of imputed data sets.

The between-imputation variance, denoted by B, can be expressed as follows:

$$B = \left(\frac{1}{m-1}\right)\sum_{t=1}^{m}\left(\hat{\theta}_t - \bar{\theta}\right)^2 \tag{3.38}$$

In this case, $\hat{\theta}_t$ denotes the parameter estimate from the imputed data set t and $\bar{\theta}$ represents the mean parameter estimate. Lastly, the pooled standard error which incorporates both the within- and between-imputation variance is given below:

$$SE = \sqrt{W + B + {}^{B}/_{m}} \tag{3.39}$$

Expanding Equation 3.39 results in the following:

$$SE = \sqrt{\frac{1}{m}\sum_{t=1}^{m} SE_t^2 + \left(\frac{1}{m-1}\right)\left(1+\frac{1}{m}\right)\sum_{t=1}^{m}\left(\hat{\theta}_t - \bar{\theta}\right)^2} \tag{3.40}$$

According to Baraldi & Enders (2010), by multiple imputation incorporating the between-imputation variance in the standard error, it fully accounts for the fact that the imputed values are simply guesstimates about the true data values. Hence, multiple imputation addresses the underestimation of the variance - a common issue with single imputation methods.

According to Berglund & Heeringa (2014), using multiple imputation to handle missing values has many advantages:

- Firstly, since multiple imputation is model based, statistical transparency and integrity of the imputation process is guaranteed. It is imperative that if one wants to ensure that the analysis is robust then the variables chosen for the imputation model should be much broader in scope than when compared to those variables chosen to be used in the analysis model. In other words, the imputation model should not be limited to only including variables that have missing values.

- Another attractive feature of using multiple imputation is the fact that it is stochastic. That is, missing values are imputed using draws of the model parameters and error terms that arise from the predictive distribution of the missing data Y_{mis} (Berglund & Heeringa (2014)). To illustrate this, suppose linear regression imputation was used to replace missing values for a continuous

variable. The conditional predictive distribution may be expressed as follows:

$$\hat{y}_{k,mis} = \hat{\beta}_0 + \hat{\beta}_{j\neq k}.y_{j\neq k} + e_k \tag{3.41}$$

In order to calculate imputed values for $Y_{k,mis}$, each of the individual predictions contain drawings from a multivariate distribution of the $\hat{\beta}$s, as well as independent draws of e_k each from their estimated distributions respectively.

- One may also find multiple imputation to be an attractive technique as it is multivariate. This means that the observed distributional properties of every variable are preserved, as well as the associations between all variables chosen for the imputation model. One must keep in mind that the multivariate relationships that are preserved in a data set assuming MAR, refer to those relationships that can be seen in the observed data, that is Y_{obs} (Berglund & Heeringa (2014)).

- Recall that multiple imputation produces multiple independent repetitions of the imputation process. By doing this, multiple imputation addresses the uncertainty in parameter estimates that occurs due to imputing missing values.

- As already mentioned, one of the most common issues with using single imputation methods is that they produce results which underestimate standard errors. That is, single imputation methods see the filled-in values as the true data values. A way in which this issue can be fixed is by using multiple imputation since the between imputation variance is included in the standard error. This term is representative of the noise in the data and results from using several different estimates in order to replace the missing values (Baraldi & Enders (2010)). As such, the standard errors resulting from multiple imputation are able to take into account that the filled-in values are only guesstimates of what the actual real data values are. Thus, multiple imputation addresses the uncertainty in missing data, as well as the issue of underestimating the variability in a data set.

- Multiple imputation is unaffected by small deviations from stringent theoretical assumptions. It is impossible to find an imputation model that will satisfy all of the assumptions associated with the missing data mechanisms and it is also difficult to find an imputation model that will truly match the distributional assumptions for the underlying random variables. Despite all of this, this technique is still capable of producing valid results.

- Another strength of multiple imputation is its applicability to real statistical scenarios – this makes its application universal in the modelling world. The implementation of multiple imputation is also not limited to any particular software package.

- Multiple imputation is also much easier to perform than maximum likelihood estimation (Russell et al. (2001)). ML techniques are computationally complicated as for each new kind of model, special implementation is required. In other words maximum likelihood may need to make use of various techniques in order to integrate out missing values for different models that are used on the same data set. Multiple imputation employs a different technique, where the calculation of the imputations remain independent of the analysis of the complete data set (Pigott (2001)). Thus, many different users may use the imputed data sets on several different analyses whilst using any software. Hence, users of multiple imputation need not be perturbed to attempt to find a solution to the missing data issue.

According to Berglund & Heeringa (2014), there are also some disadvantages of using multiple imputation:

- Firstly, there may be cases where the model used by the imputer is vastly different from the model used by the data analyst. For example, the imputer may forget or decide not to incorporate significant variables into the imputation model. Also, when imputing the missing values, the imputer might assume a linear relationship between two variables when actually this assumption is not correct. These issues mentioned above, as well as the amount of missing data occurring for each of the variables of interest can result in a biased analysis.

- Allison (2001) states that another disadvantage of using multiple imputation, is that it can be performed in many different ways and this may lead to possible doubt and confusion. Also, because each of the imputed values are random draws, different results will be obtained every time this technique is run.

- One of the biggest issues plaguing the MCMC algorithm is that it assumes that each variable containing missing values is normally distributed. Although this assumption does not hold true for categorical variables, according to Schafer (1997) (and the references within), the multivariate normal model produces satisfactory results even when the variables are binary or categorical, with the imputations achieved under the assumption of a normal model.

There exist various alternative models and computational methods that are available for performing multiple imputation which Schafer (1997) has developed in order to analyse missing values. According to Schafer & Olsen (1998) the packages listed below allows the user to generate multiple imputations using SPLUS:

- NORM which performs multiple imputation under a multivariate normal model.

- CAT is used for multivariate categorical data where the MCMC algorithm is used under a multinomial restricted log-linear model.

- MIX can be used for data sets that are mixed (i.e. data sets that contain both continuous and categorical variables)
- PAN, for multivariate panel data or clustered data under a multivariate linear mixed-effects model.

SPSS version 13 is also another software that may be used, however, this dissertation will use SAS version 9.4 to perform multiple imputation. SAS makes use of two procedures, namely PROC MI and PROC MIANALYSE. PROC MI can be used to create m complete data sets. These complete data sets can then be analysed using any standard SAS procedure such as linear regression. Then, in order to pool the results, the MIANALYSE procedure should be used. It is important to note that this procedure uses the multivariate normal algorithm as default. There are, however, alternative options that may be used.

Chapter 4

Data Analysis and Results

In order to demonstrate the missing data methods described in Chapter 3, the National Longitudinal Survey of Youth (NLSY) data set used by Allison (2001) will be examined. This data set consists of records for 581 children who were interviewed in 1990, 1992 as well as in 1994. For the purpose of this dissertation, only the data where the child was interviewed in 1990 will be used.

The NLSY data set contains the following variables:

- ANTI - which describes the child's antisocial behaviour in 1990 and is measured with a scale ranging from 0-6. 0 represents the lowest level of antisocial behaviour and 6 represents the highest level of antisocial behaviour.
- SELF - which describes the child's self-esteem in 1990 and is measured with a scale ranging from 6-24, where 6 represents a low self-esteem and 24 represents a high self-esteem.
- POV - which determines the poverty status of a family in 1990, coded 1 for in poverty and 0 otherwise.
- BLACK - coded 1 if the child is Black and 0 otherwise.
- HISPANIC - coded 1 if the child is Hispanic and 0 otherwise.
- CHILDAGE - child's age in 1990.
- DIVORCE - coded 1 if the mother was divorced in 1990 and 0 otherwise.
- GENDER - coded 1 if the child is female and 0 if the child is male.
- MOMAGE - mother's age at birth of child.
- MOMWORK - coded 1 if the mother was employed in 1990 and 0 otherwise.

It is important to note that BLACK and HISPANIC are two categories of a three-category variable, with the reference category being NON-HISPANIC WHITE.

The ultimate aim is to estimate a linear regression model using ANTI as the dependent variable while using the rest of the variables as predictors. The raw data set initially consisted of no missing values. Allison (2001) deliberately created missing values on some of the variables using a method that satisfies the MAR assumption. For more information on how to generate such data, one can consult Truxillo (2005).

The variables that contain missing values include SELF, POV, BLACK, HISPANIC as well as MOMWORK. The percentages of missing values for each of these variables are listed and summary graphs are given below:

- 25% missing on SELF,
- 26% missing on POV,
- 16% missing on BLACK and HISPANIC,
- 15% missing on MOMWORK.

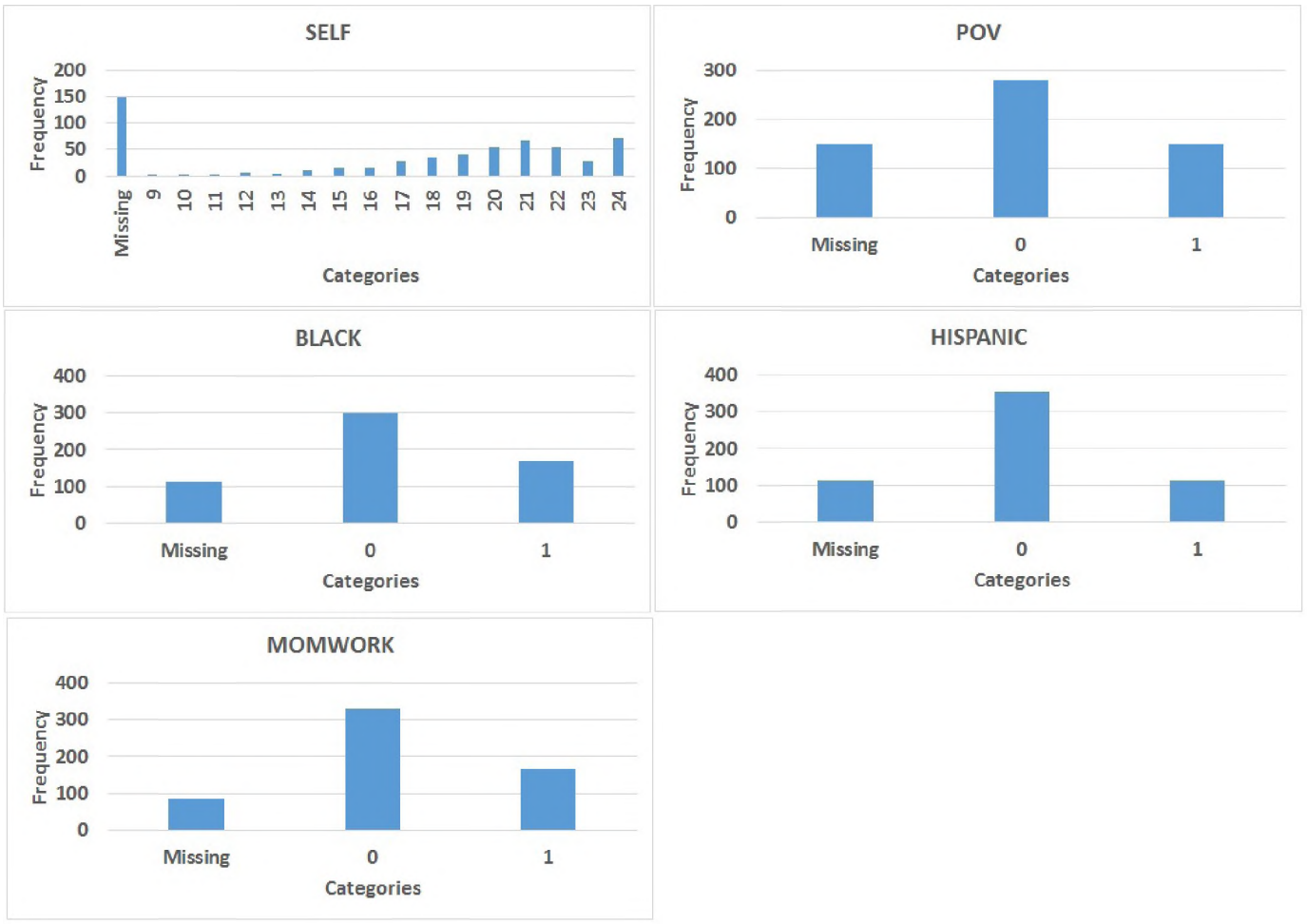

Figure 4.1: A summary of the five variables that contain missing data.

It should be noted that although this data set contains dummy variables for POV, BLACK, HISPANIC and MOMWORK, techniques such as maximum likelihood and multiple imputation will now be illustrated. These methods rely on the assumptions of multivariate normality, however, according to Allison (2001), a good deal of simulation evidence and practical experience suggests that these techniques can still be used even though these dummy variables do not have a normal distribution. In fact, according to Peng et al. (2006) and Enders (2010), empirical studies suggest that when normality assumptions are violated, the accuracy of multiple imputation estimates are not seriously affected.

The techniques discussed in Chapter 3 above will now be discussed.

4.1 Listwise Deletion

The first method that will be illustrated is listwise deletion. As already mentioned in Chapter 3, listwise deletion involves discarding all cases that contain missing values. This method is not recommended by many, as one of the problems associated with this technique is discarding a lot of potentially usable data. Listwise deletion in fact, does not satisfy any of the criteria that was proposed by Allison (2001) in Chapter 3 as discarding data leads to larger standard errors, wider confidence intervals, as well as significance tests lacking power as results are not representative of the population sampled (Peng et al. (2006)). Also, if the data is not MCAR the estimates will be biased.

Although listwise deletion is only valid under MCAR, this technique will still be illustrated on the NLSY data set. The bar graph below shows the total number of observations in the data set, the number of observations that contain missing values, as well as the remaining number of observations that will be used going forward in the analysis:

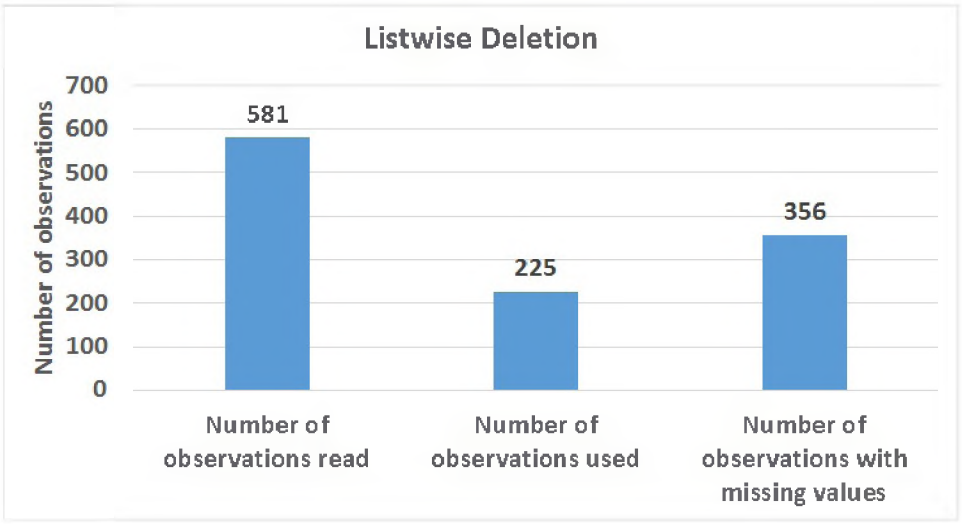

Figure 4.2: Bar graph displaying the total number of observations in the data set, the number of observations that contain missing values, as well as the remaining number of observations that will be used going forward in the analysis.

It can be seen from the graph above that out of a total of 581 observations, more than half of the observations have been discarded and hence only 225 observations will be used in the analysis.

Next, a regression equation is fit to the remaining data using ANTI as the dependent variable and the rest of the variables as predictors. The code that was used may be found in Appendix A.2. Table 4.1 displays the regression results based on the complete data set (with no missing values) and Table

4.2 shows the regression results based on listwise deletion:

Table 4.1: Table displaying regression results for the variable ANTI on the complete data set with no missing values.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	2.62768	1.22515	2.14	0.0324
SELF	**1**	**-0.05501**	**0.01863**	**-2.95**	**0.0033**
POV	**1**	**0.57477**	**0.13896**	**4.14**	**<.0001**
BLACK	1	0.09458	0.14114	0.67	0.5031
HISPANIC	**1**	**-0.34294**	**0.15335**	**-2.24**	**0.0257**
CHILDAGE	1	-0.01273	0.10049	-0.13	0.8993
DIVORCE	1	-0.06734	0.14395	-0.47	0.6401
GENDER	**1**	**-0.54043**	**0.11728**	**-4.61**	**<.0001**
MOMAGE	1	0.01172	0.02815	0.42	0.6774
MOMWORK	1	0.1824	0.12925	1.41	0.1587

Table 4.2: Table displaying regression results for the variable ANTI using listwise deletion.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	2.86533	1.99117	1.44	0.1516
SELF	1	-0.04531	0.03135	-1.45	0.1498
POV	**1**	**0.71946**	**0.23739**	**3.03**	**0.0027**
BLACK	1	0.05069	0.24918	0.20	0.8390
HISPANIC	1	-0.35696	0.25537	-1.40	0.1636
CHILDAGE	1	0.00197	0.17072	0.01	0.9908
DIVORCE	1	0.08703	0.24499	0.36	0.7228
GENDER	1	-0.33470	0.19844	-1.69	0.0931
MOMAGE	1	-0.01198	0.04611	-0.26	0.7953
MOMWORK	1	0.25440	0.21751	1.17	0.2435

From the two tables above it is clear that when listwise deletion is used, the standard errors for all of the variables are much larger than the standard errors for the complete data set. For example, the variable POV on the complete data set has a standard error of 0.13896 compared to a standard error of 0.23739 when using listwise deletion. Similarly, HISPANIC has a standard error of 0.15335 on the complete data set and a standard error of 0.25537 on the listwise deletion data set. Based on these results, listwise deletion should be avoided under all circumstances.

It can also be seen from Table 4.2 that because more than half the cases were discarded, only POV is significant with a p-value below 0.05 (highlighted in bold) and hence one can conclude that high levels

of antisocial behaviour can be associated with being in poverty. The variables that are significant at the 0.05 level on the complete data set include SELF, POV, HISPANIC and GENDER and these are highlighted in bold in Table 4.1 above.

4.2 Pairwise Deletion

As described in Chapter 3, pairwise deletion is a method that makes use of cases that contain complete data on only those variables that were chosen for the study (Osborne (2013)). For many linear models, the parameters of interest can be expressed as functions of the population means, variances and covariances.

This technique may be illustrated using the PROC CORR function in SAS. The code that was used may be found in Appendix A.3. Table 4.3 below shows the results of the correlations between each of the variables:

Table 4.3: Table displaying the correlations between each of the variables.

		ANTI	SELF	POV	BLACK	HISPANIC	CHILDAGE	DIVORCE	GENDER	MOMAGE	MOMWORK
ANTI	correlation	1	-0.14359	0.23038	0.11963	-0.08219	0.01579	0.03086	-0.172	-0.03021	0.11253
	p-value		0.0027	<.0001	0.0096	0.0757	0.704	0.4578	<.0001	0.4674	0.0122
	sample size	581	433	431	468	468	581	581	581	581	495
SELF	correlation	-0.14359	1	-0.05723	-0.05811	-0.08915	0.07902	-0.11297	-0.01844	0.08517	0.00084
	p-value	0.0027		0.3059	0.2776	0.0954	0.1006	0.0187	0.702	0.0767	0.9871
	sample size	433	433	322	351	351	433	433	433	433	372
POV	correlation	0.23038	-0.05723	1	0.26739	0.00319	0.07405	0.26932	0.05469	-0.20747	0.25512
	p-value	<.0001	0.3059		<.0001	0.9525	0.1248	<.0001	0.2573	<.0001	<.0001
	sample size	431	322	431	350	350	431	431	431	431	366
BLACK	correlation	0.11963	-0.05811	0.26739	1	-0.42466	0.01251	0.01973	0.04843	-0.12254	-0.07014
	p-value	0.0096	0.2776	<.0001		<.0001	0.7872	0.6704	0.2957	0.008	0.1631
	sample size	468	351	350	468	468	468	468	468	468	397
HISPANIC	correlation	-0.08219	-0.08915	0.00319	-0.42466	1	-0.05309	0.04062	-0.06996	-0.0054	0.07804
	p-value	0.0757	0.0954	0.9525	<.0001		0.2517	0.3807	0.1307	0.9073	0.1206
	sample size	468	351	350	468	468	468	468	468	468	397
CHILDAGE	correlation	0.01579	0.07902	0.07405	0.01251	-0.05309	1	0.01387	-0.07391	-0.23372	0.00582
	p-value	0.704	0.1006	0.1248	0.7872	0.2517		0.7386	0.075	<.0001	0.8972
	sample size	581	433	431	468	468	581	581	581	581	495
DIVORCE	correlation	0.03086	-0.11297	0.26932	0.01973	0.04062	0.01387	1	-0.00073	-0.06273	-0.04626
	p-value	0.4578	0.0187	<.0001	0.6704	0.3807	0.7386		0.9861	0.1309	0.3043
	sample size	581	433	431	468	468	581	581	581	581	495
GENDER	correlation	-0.172	-0.01844	0.05469	0.04843	-0.06996	-0.07391	-0.00073	1	0.02969	0.01281
	p-value	<.0001	0.702	0.2573	0.2957	0.1307	0.075	0.9861		0.4751	0.7762
	sample size	581	433	431	468	468	581	581	581	581	495
MOMAGE	correlation	-0.03021	0.08517	-0.20747	-0.12254	-0.0054	-0.23372	-0.06273	0.02969	1	-0.01776
	p-value	0.4674	0.0767	<.0001	0.008	0.9073	<.0001	0.1309	0.4751		0.6935
	sample size	581	433	431	468	468	581	581	581	581	495
MOMWORK	correlation	0.11253	0.00084	0.25512	-0.07014	0.07804	0.00582	-0.04626	0.01281	-0.01776	1
	p-value	0.0122	0.9871	<.0001	0.1631	0.1206	0.8972	0.3043	0.7762	0.6935	
	sample size	495	372	366	397	397	495	495	495	495	495

It is important to remember that when estimating the correlation between any two variables, the individuals that will be included are only those that have information on that pair of variables.

A linear regression model can be fitted using the pairwise correlation matrix as input where ANTI

is the dependent variable. The results are given in Table 4.4 below:

Table 4.4: Table displaying regression results for the variable ANTI using pairwise deletion.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	2.55641	1.40381	1.82	0.0693
SELF	**1**	**-0.07037**	**0.02198**	**-3.2**	**0.0015**
POV	**1**	**0.70045**	**0.16334**	**4.29**	**<.0001**
BLACK	1	0.0803	0.16454	0.49	0.6258
HISPANIC	**1**	**-0.35032**	**0.17613**	**-1.99**	**0.0473**
CHILDAGE	1	-0.00212	0.11593	-0.02	0.9854
DIVORCE	1	-0.1326	0.16659	-0.8	0.4265
GENDER	**1**	**-0.58084**	**0.13510**	**-4.3**	**<.0001**
MOMAGE	1	0.02467	0.03242	0.76	0.4471
MOMWORK	1	0.20527	0.14985	1.37	0.1715

Drawing conclusions based on the above table, it can be seen that the variables highlighted in bold, namely SELF, POV, HISPANIC and GENDER are significant at the 0.05 level. Hence, higher levels of antisocial behaviour can be linked to lower levels of self-esteem, being in poverty as well as being male. Also, antisocial behaviour is lower when the child is Hispanic.

In Table 4.5 below, a comparison of standard errors is made when the complete data set is used, as well as when listwise and pairwise deletion is used:

Table 4.5: Table comparing the standard errors obtained from (a) the complete data set, (b) listwise deletion and (c) pairwise deletion.

Standard Errors			
Variable	Complete Data	Listwise Deletion	Pairwise Deletion
SELF	0.01863	0.03135	0.02198
POV	0.13896	0.23739	0.16334
BLACK	0.14114	0.24918	0.16454
HISPANIC	0.15335	0.25537	0.17613
CHILDAGE	0.10049	0.17072	0.11593
DIVORCE	0.14395	0.24499	0.16659
GENDER	0.11728	0.19844	0.13510
MOMAGE	0.02815	0.04611	0.03242
MOMWORK	0.12925	0.21751	0.14985

It can be seen from the above that when comparing the standard errors obtained from the complete data set to the standard errors obtained from pairwise and listwise deletion, they are much smaller. For

example for the variable SELF, the complete data set has a standard error of 0.01863, listwise deletion produces a standard error of 0.03135 and pairwise deletion produces a standard error of 0.02198. Hence, making use of these methods is not ideal as they result in very large standard errors.

Although pairwise deletion does not discard any data, this technique does not satisfy all of the missing data criteria proposed by Allison (2001) in Chapter 3. It is thus submitted that this method should not be recommended, as it results in a loss of degrees of freedom as well as statistical power as standard deviations, correlations and covariances are calculated based on the available data for each variable. According to Peng et al. (2006), since the sample size constantly changes for each variable, the population to which the results are generalised becomes unclear. Estimates will also be biased if the data is not MCAR.

4.3 Mean Imputation

The next method that will be illustrated is mean imputation. In this case, the mean will be used to replace all missing values occurring for some of the variables with the mean of all the known values for that variable (Peng et al. (2006)). The means for each of the variables using the complete data set (no missing values), as well as the means for the data set containing missing values are given in Table 4.6 below:

Table 4.6: Table displaying the means for each variable for (a) the complete data set and (b) the missing data set.

Variable	Mean (Complete Data)	Mean (Missing Data)
ANTI	1.56799	1.56799
SELF	20.07057	**20.05081**
POV	0.33563	**0.34803**
BLACK	0.36317	**0.35897**
HISPANIC	0.24441	**0.24359**
CHILDAGE	8.94363	8.94363
DIVORCE	0.2358	0.2358
GENDER	0.5043	0.5043
MOMAGE	20.65577	20.65577
MOMWORK	0.33563	**0.33535**

Note that the variables highlighted in bold are the variables that have missing values and hence have different means to the complete data set.

Mean imputation will replace the variables that have missing values with the means in Table 4.6. That is, for every missing value occurring on the variable SELF, the mean of 20.05081 will be used to replace these values. The mean of 0.34803 will be used to replace each missing value that occurs for the variable POV. Similarly, the means of 0.35897 and 0.24359 will replace the missing values for the variables BLACK and HISPANIC respectively. Lastly, the mean of 0.33535 will be used to replace missing values on the variable MOMWORK. Subsequent analyses will then treat these imputed values as if they are real data.

As already mentioned, although using mean imputation preserves the data, this method is not a satisfactory solution to the missing data problem. This is because using the mean to replace missing values not only produces biased estimates under any type of missingness, but it also reduces the variance on the variable in question as well as negatively affects correlations (Peng et al. (2006)).

Table 4.7 below shows how the variance reduces for the variables SELF, POV, BLACK, HISPANIC and MOMWORK when mean imputation is used:

Table 4.7: Table displaying the variance when (a) the complete data set was used and (b) mean imputation was used.

Variable	Variance (Complete Data)	Variance (Mean Imputation)
SELF	10.18639	7.2498
POV	0.22337	0.16861
BLACK	0.23168	0.18568
HISPANIC	0.18499	0.14867
MOMWORK	0.22337	0.19023

It is clear from the above table that the variance significantly reduces when mean imputation is used. For example, the variance for SELF using the complete data set is 10.18639 compared to a variance of 7.2498 when mean imputation is used. The following box plot further illustrates the above example for the variable SELF:

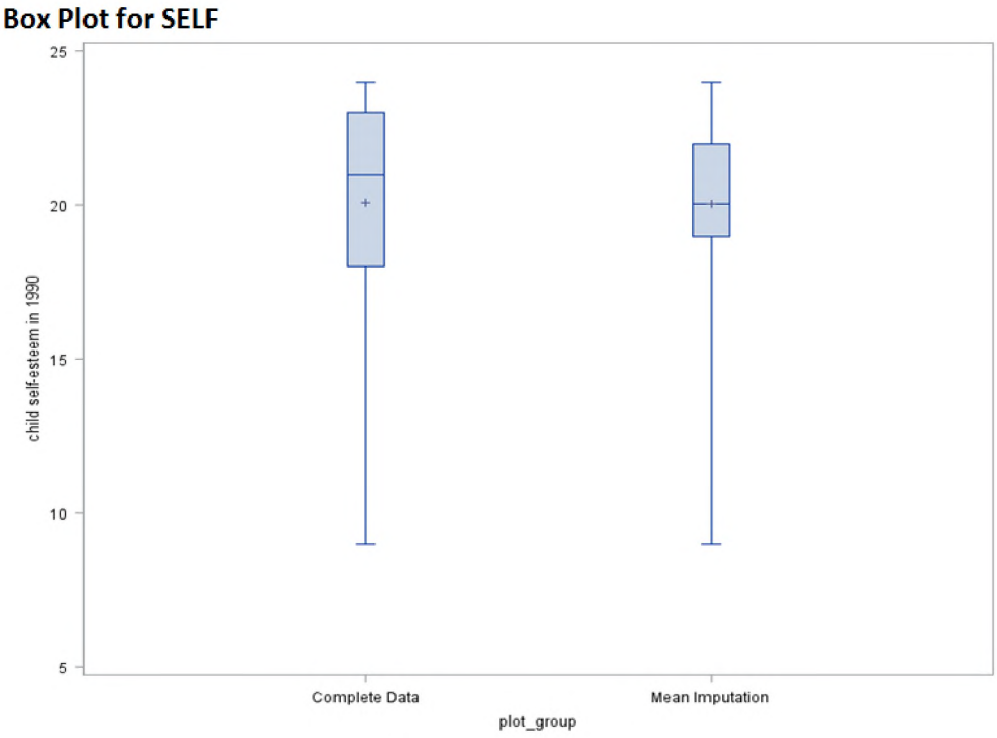

Figure 4.3: Box plot displaying how the variance reduces for the variable SELF when mean imputation is used.

From the above box plot it is clear that when mean imputation is applied to the missing data set the variability in the data set reduces. That is, the complete data set shows a wider interquartile range than the imputed data set.

Although mean imputation is not recommended, the technique will still be performed for illustration purposes only. The code that was used may be found in Appendix A.4. Once the mean has replaced the missing values, a linear regression model can be fit to the now complete data set. In this case, ANTI is used as the dependent variable and the rest of the variables are used as predictors. The regression equation is given below:

$$y = \beta_0 + \beta_1(SELF) + \beta_2(POV) + \beta_3(BLACK) + \beta_4(HISPANIC) + \beta_5(CHILDAGE) + \beta_4(DIVORCE)$$
$$+\beta_7(GENDER) + \beta_8(MOMAGE) + \beta_9(MOMWORK) \quad (4.1)$$

where y represents the variable ANTI. The results from the regression analysis are given in Table 4.8 below:

Table 4.8: Table displaying regression results for the variable ANTI using mean imputation.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	2.6846	1.23666	2.17	0.0304
SELF	**1**	**-0.07056**	**0.02198**	**-3.21**	**0.0014**
POV	**1**	**0.66452**	**0.15602**	**4.26**	**<.0001**
BLACK	1	0.1568	0.15512	1.01	0.3125
HISPANIC	1	-0.31045	0.16927	-1.83	0.0672
CHILDAGE	1	-0.00202	0.1005	-0.02	0.984
DIVORCE	1	-0.06319	0.14238	-0.44	0.6573
GENDER	**1**	**-0.56297**	**0.1172**	**-4.8**	**<.0001**
MOMAGE	1	0.01533	0.028	0.55	0.5842
MOMWORK	1	0.25741	0.13794	1.87	0.0625

From the above table it is clear that only the variables SELF, POV and GENDER are significant at the 0.05 level and are highlighted in bold above. Hence, a conclusion can be drawn that high levels of antisocial behaviour can be linked to lower levels of self-esteem, experiencing poverty as well as being male.

4.4 Mode Imputation

The next method that will be illustrated using the NLSY data set is mode imputation. Table 4.9 below shows the values that occur most frequently in the missing data set for each variable. The modes will be used to replace the missing values that occur for the variables SELF, POV, BLACK, HISPANIC as well as MOMWORK. That is, the value 24 will be used to replace all missing values that occur on the variable SELF and 0 will replace all missing values for the variables POV, BLACK, HISPANIC and MOMWORK.

Table 4.9: Table displaying the mode for each variable in the missing data set.

Variable	Mode
ANTI	0
SELF	**24**
POV	**0**
BLACK	**0**
HISPANIC	**0**
CHILDAGE	8.25
DIVORCE	0
GENDER	1
MOMAGE	20
MOMWORK	**0**

Note that the variables highlighted in bold are the variables that have missing values.

Mode imputation is usually used in cases where the data is categorical (Munguía & Armando (2014)). This technique not only reduces the variance in a data set but also produces results that contain weak covariance and correlation estimates. Table 4.10 below compares the variance on the complete data set to the variance when mode imputation was used:

Table 4.10: Table displaying the variance when (a) the complete data set was used and (b) mode imputation was used.

Variable	Variance (Complete Data)	Variance (Mode Imputation)
SELF	10.18639	10.21573
POV	0.22337	0.19185
BLACK	0.23168	0.20590
HISPANIC	0.18499	0.15799
MOMWORK	0.22337	0.20443

It is clear from the table above, that in most cases the variance reduces when mode imputation is used. Furthermore, mode imputation changes the distribution of the data. The bar graph below illustrates this by comparing the proportion of people in poverty for the complete data set to the proportion of people in poverty once mode imputation has been used to replace all missing values for the variable POV:

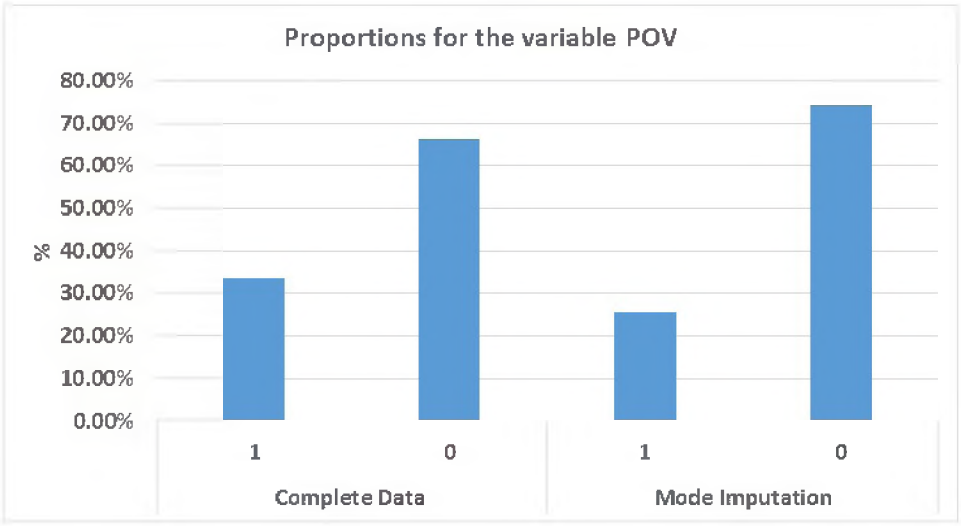

Figure 4.4: Bar graph displaying how mode imputation alters the distribution of the data.

Although mode imputation is not advised, for illustration purposes only, using the now complete data

set (as all missing values have been replaced by the mode), a linear regression will be estimated using ANTI as the dependent variable and the rest of the variables as predictors. The code that was used may be found in Appendix A.5. The results are given in Table 4.11 below:

Table 4.11: Table displaying regression results for the variable ANTI using mode imputation.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	3.22030	1.18263	2.72	0.0067
SELF	**1**	**-0.10828**	**0.01818**	**-5.96**	**<.0001**
POV	**1**	**0.68643**	**0.14060**	**4.88**	**<.0001**
BLACK	1	0.13099	0.13542	0.97	0.3338
HISPANIC	**1**	**-0.30550**	**0.15217**	**-2.01**	**0.0452**
CHILDAGE	1	0.01224	0.09777	0.13	0.9004
DIVORCE	1	-0.10130	0.13784	-0.73	0.4627
GENDER	**1**	**-0.57090**	**0.11404**	**-5.01**	**<.0001**
MOMAGE	1	0.02988	0.02732	1.09	0.2745
MOMWORK	1	0.20140	0.12926	1.56	0.1198

From the above table it is clear that the variables SELF, POV, HISPANIC and GENDER have p-values below 0.05 and are given in bold above. Once again, a similar conclusion can be made - that high levels of antisocial behaviour are associated with lower levels of self-esteem, experiencing poverty as well as being male. Antisocial behaviour is also linked to race, where a child who is HISPANIC has a lower level of antisocial behaviour.

4.5 Median Imputation

Median imputation is the next technique that will be discussed. In this case, for each variable, the missing data is replaced by the median of all known values of that attribute (Acuña & Rodriguez (2004)). Table 4.12 below displays the median for each of the variables in the missing data set:

Table 4.12: Table displaying the median for each variable in the missing data set.

Variable	Median
ANTI	1
SELF	**21**
POV	**0**
BLACK	**0**
HISPANIC	**0**
CHILDAGE	8.91667
DIVORCE	0
GENDER	1
MOMAGE	21
MOMWORK	**0**

Note that the variables highlighted in bold are the variables that have missing values. Each missing value occurring on the variable SELF will be replaced by 21 and all missing values occurring on the variables POV, BLACK, HISPANIC and MOMWORK will be replaced by 0.

Median imputation is also not recommended as the variance estimates are greatly underestimated since residual variance is not taken into account. Further, if a single value is used to replace each missing item, this can negatively affect any relationships that may occur between other variables in the data set (Baneshi & Talei (2012)). Table 4.13 below shows how the variance reduces when median imputation is used:

Table 4.13: Table displaying how the variance reduces when (a) the complete data set was used and (b) median imputation was used.

Variable	Variance (Complete Data)	Variance (Median Imputation)
SELF	10.18639	7.42113
POV	0.22337	0.19185
BLACK	0.23168	0.20590
HISPANIC	0.18499	0.15799
MOMWORK	0.22337	0.20443

The following box plot further illustrates the above for the variable SELF. It is clear that median imputation affects the variability in the data as the interquartile range for the complete data set is much wider than the interquartile range for the imputed data set.

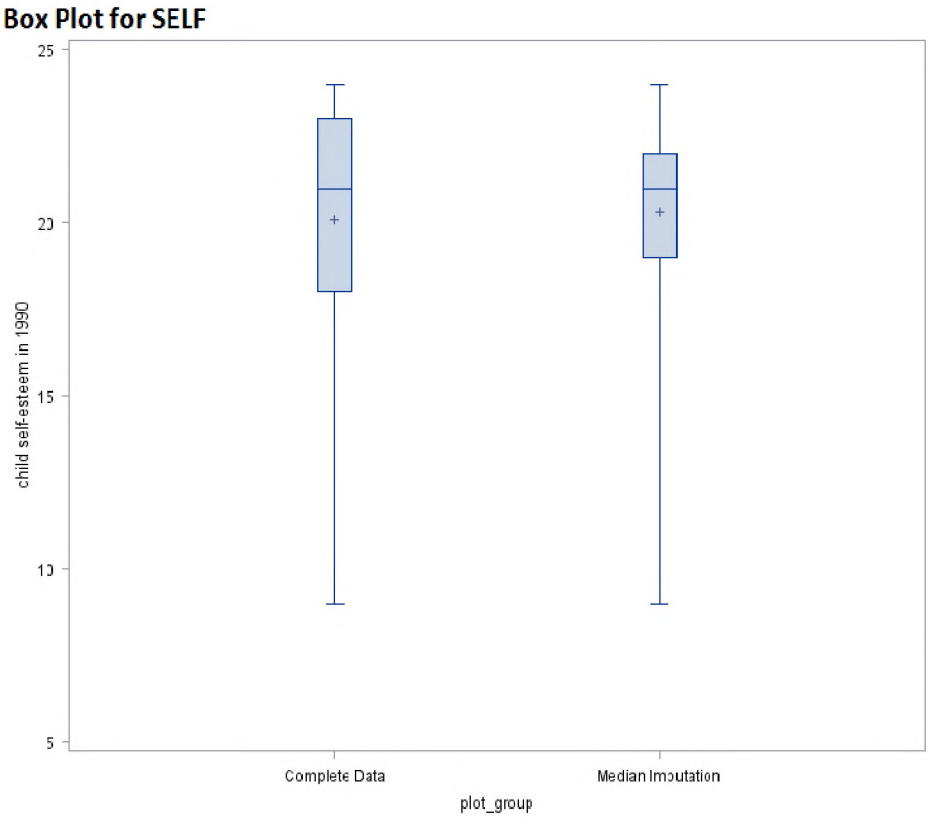

Figure 4.5: Box plot displaying how the variance reduces for the variable SELF when median imputation is used.

Although median imputation is not advised, for illustration purposes only, a linear regression is estimated using ANTI as the dependent variable and the rest of the variables as predictors. The code that was used to perform median imputation may be found in Appendix A.6. The results can be seen in Table 4.14 below:

Table 4.14: Table displaying regression results for the variable ANTI using median imputation.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	2.94934	1.21477	2.43	0.0155
SELF	**1**	**-0.08733**	**0.02153**	**-4.06**	**<.0001**
POV	**1**	**0.72695**	**0.14261**	**5.1**	**<.0001**
BLACK	1	0.16916	0.13731	1.23	0.2185
HISPANIC	1	-0.29756	0.15482	-1.92	0.0551
CHILDAGE	1	-0.00166	0.09936	-0.02	0.9867
DIVORCE	1	-0.09036	0.14021	-0.64	0.5195
GENDER	**1**	**-0.55475**	**0.11581**	**-4.79**	**<.0001**
MOMAGE	1	0.02273	0.02772	0.82	0.4125
MOMWORK	1	0.20632	0.13133	1.57	0.1167

From the above table, the variables that have p-values below 0.05 (highlighted in bold) include SELF, POV, and GENDER. That is, high levels of antisocial behaviour are associated with lower levels of self-esteem, experiencing poverty as well as being male.

4.6 Regression Imputation

As described in Chapter 3, regression imputation substitutes missing items with predicted scores resulting from a regression equation (Baraldi & Enders (2010)). In cases where missing values occur on a single variable, one can use complete case analysis in order to determine the regression equation. Variables with missing values are used as the outcome variables and variables with complete data are used as the predictor variables. The resulting regression equation can be used to predict scores as the data set is now complete. The code that was used to perform regression imputation for each of the variables may be found in Appendix A.7.

The ultimate goal is to use regression imputation to replace all missing values occurring on each of the variables and then fit a final regression model on the new complete data set.

4.6.1 SELF

Firstly, regression imputation will be performed on the variable SELF. Simple linear regression can be used for imputation using SELF as the dependent variable and the rest of the variables that have complete data as the predictor variables. That is, the predictor variables will be CHILDAGE, DIVORCE, GENDER, MOMAGE as well as ANTI. The regression equation is given below:

$$y = \beta_0 + \beta_1(CHILDAGE) + \beta_2(DIVORCE) + \beta_3(GENDER) + \beta_4(MOMAGE) + \beta_5(ANTI) \quad (4.2)$$

where y represents the dependent variable, SELF.

Note that the PROC REG function in SAS was used to fit a model to the data. According to Inc. (2008b), the MODEL statement specifies the dependent and independent variables in the regression model. The statement OUTEST produces a new data set that includes parameter estimates and other model fit statistics. Table 4.15 below displays the results from the regression:

Table 4.15: Table displaying regression results for the variable SELF when regression imputation was used.

Number of observations read	581
Number of observations used	433
Number of observations with missing values	148

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	13.2791	2.93216	4.53	<.0001
CHILDAGE	**1**	**0.5107**	**0.25232**	**2.02**	**0.0436**
DIVORCE	**1**	**-0.72123**	**0.34408**	**-2.1**	**0.0367**
GENDER	1	-0.22855	0.30009	-0.76	0.4467
MOMAGE	**1**	**0.14765**	**0.06856**	**2.15**	**0.0318**
ANTI	**1**	**-0.29842**	**0.09897**	**-3.02**	**0.0027**

From the above table, it is clear that the variables CHILDAGE, DIVORCE, MOMAGE and ANTI are significant at the 0.05 level and have been highlighted in bold.

Next, in order to replace the missing values for the variable SELF the following equation can be used:

$$\text{if SELF} = . \text{ then} \tag{4.3}$$

$$\text{SELF} = \text{INTERCEPT} + \text{CHILDAGE} * (\text{CHILDAGE\_EST}) + \text{DIVORCE} * (\text{DIVORCE\_EST}) + \text{GENDER} * (\text{GENDER\_EST}) + \text{MOMAGE} * (\text{MOMAGE\_EST}) + \text{ANTI} * (\text{ANTI\_EST}) \tag{4.4}$$

Note that CHILDAGE_EST, DIVORCE_EST, GENDER_EST, MOMAGE_EST and ANTI_EST all refer to the parameter estimates which are given in Table 4.15 above and in turn, each of these are multiplied by the original variable values (CHILDAGE, DIVORCE, GENDER, MOMAGE and ANTI) in order to replace missing values for the variable SELF.

Now that the variable SELF is complete (no longer contains missing values), a similar approach will be carried out for the rest of the variables that contain missing values.

4.6.2 POV

The next two variables, POV and MOMWORK, are categorical variables and hence a different kind of regression equation will be fit. According to Agresti (2002), logistic regression yields the best results in analyses that contain categorical response data. That is, consider a binary response variable Y as well as an explanatory variable X, and let

$$\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x) \tag{4.5}$$

The logistic regression model is given by:

$$\pi(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)} \tag{4.6}$$

Then, the log odds, also known as the logit, takes on a linear relationship in the form of:

$$logit[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x \tag{4.7}$$

Hence, the logit link function is equal to the linear predictor function (Agresti (2002)).

A logistic regression model is fitted to the data using the PROC LOGISTIC procedure available in SAS. POV is used as the dependent variable and the variables with complete data are used as the predictor variables. The regression equation is shown below:

$$log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1(CHILDAGE) + \beta_2(DIVORCE) + \beta_3(GENDER) + \beta_4(MOMAGE) + \beta_5(ANTI) \tag{4.8}$$

where Y represents the dependent variable, POV. The results from the logistic regression can be seen in Table 4.16 below:

Table 4.16: Table displaying logistic regression results for the variable POV.

Variable	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.9233	2.1686	0.1813	0.6703
CHILDAGE	1	-0.1491	0.1854	0.6466	0.4213
DIVORCE	**1**	**-1.3293**	**0.2517**	**27.893**	**<.0001**
GENDER	1	-0.4355	0.2261	3.7079	0.0542
MOMAGE	**1**	**0.202**	**0.0532**	**14.3878**	**0.0001**
ANTI	**1**	**-0.3805**	**0.0774**	**24.1707**	**<.0001**

The following three variables are significant at the 0.05 level and are highlighted in bold: DIVORCE,

MOMAGE and ANTI.

Next, a new variable called PI_POV was created using Equation 4.6 above, where PI_POV is calculated only for those cases where the variable POV is missing. Equation 4.6 calculates a probability and thus, in order to determine whether POV gets a 1 or a 0, the following rule was applied:
If

$$\text{PI\_POV} > 0.5 \tag{4.9}$$

then

$$\text{POV} = 1 \tag{4.10}$$

else

$$\text{POV} = 0 \tag{4.11}$$

A snapshot of the results when the above rule is used is given in Table 4.17 below:

Table 4.17: Table displaying a snapshot of the results from regression imputation for the variable POV.

ID	ANTI	SELF	**POV**	BLACK	HISPANIC	CHILDAGE	DIVORCE	GENDER	MOMAGE	MOMWORK	**PI_POV**
1	1	21	1	0	0	8	0	1	21	0	
2	0	20	**1**			8.41667	0	1	22	1	**0.86171**
3	5	21	0			8.08333	1	0	18	0	
4	2	23	0	0	0	8.25	0	0	24	0	
5	1	22	0	0	0	9.33333	0	1	22	0	
6	1	20.90782	0	0	0	8.58333	0	0	24	0	
7	3	24	0	0	0	9.25	1	1	23		
8	4	19	0			8.5	1	0	18	0	
9	1	21	**1**			8.08333	0	0	24	0	**0.91199**
10	4	9	0	0	0	9.16667	1	0	20		
11	3	20	1			8.83333	1	1	23	1	
12	3	15	**0**	0	1	9.16667	1	1	20	0	**0.23924**
13	3	19.5727	**1**	0	0	8.58333	0	0	19		**0.62079**
14	1	20.69763	0	0	0	8.75	0	0	22	1	
15	3	21	0			9.83333	0	1	23		
16	2	16	**1**	1	0	9	0	1	20	0	**0.64056**
17	1	18	**1**	1	0	8.66667	0	0	19		**0.77581**
18	0	21.29396	0			9.33333	0	0	22	1	
19	3	19	1	1	0	8.16667	0	1	18	0	

From the above table the following can be noted: The variable ID creates a unique ID for each observation. Then, when PI_POV has a probability that is greater than 0.5, the variable POV will equal 1. In cases where PI_POV has a probability of less than 0.5, the variable POV will equal 0. For example, consider the second observation. The variable PI_POV has a probability of 0.86171 and hence

POV will equal 1. All missing values for the variable POV will be filled-in in this way and hence this variable will no longer contain missing data.

4.6.3 MOMWORK

The regression imputation analysis on the variable MOMWORK is carried out in the same way as it was for the variable POV. The logistic regression is given below:

$$log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\beta_0+\beta_1(CHILDAGE)+\beta_2(DIVORCE)+\beta_3(GENDER)+\beta_4(MOMAGE)+\beta_5(ANTI) \tag{4.12}$$

where Y represents the dependent variable, MOMWORK. The results from the logistic regression can be seen in Table 4.18 below:

Table 4.18: Table displaying logistic regression results for the variable MOMWORK.

Variable	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6989	1.94	0.1298	0.7186
CHILDAGE	1	-0.0101	0.1648	0.0038	0.9509
DIVORCE	1	0.2933	0.2346	1.5629	0.2112
GENDER	1	-0.1506	0.1964	0.588	0.4432
MOMAGE	1	0.0177	0.0457	0.1498	0.6987
ANTI	**1**	**-0.1732**	**0.0654**	**7.0054**	**0.0081**

From the above table the only significant variable at the 0.05 level is ANTI which has been highlighted in bold.

Next, a new variable called PI_MOMWORK was calculated and only used when MOMWORK was missing. The same rules that were applied to the variable POV were also applied to the variable MOMWORK. That is, if

$$\text{PI\_MOMWORK} > 0.5 \tag{4.13}$$

then

$$\text{MOMWORK} = 1 \tag{4.14}$$

else

$$\text{MOMWORK} = 0 \tag{4.15}$$

A snapshot of the results when the above rule is used is given in Table 4.19 below:

Table 4.19: Table displaying a snapshot of the results from regression imputation for the variable MOMWORK.

ID	ANTI	SELF	POV	BLACK	HISPANIC	CHILDAGE	DIVORCE	GENDER	MOMAGE	**MOMWORK**	**PI_MOMWORK**
1	1	21	1	0	0	8	0	1	21	0	
2	0	20	1			8.41667	0	1	22	1	
3	5	21	0			8.08333	1	0	18	0	
4	2	23	0	0	0	8.25	0	0	24	0	
5	1	22	0	0	0	9.33333	0	1	22	0	
6	1	20.90782	0	0	0	8.58333	0	0	24	0	
7	3	24	0	0	0	9.25	1	1	23	**1**	**0.65369**
8	4	19	0			8.5	1	0	18	0	
9	1	21	1			8.08333	0	0	24	0	
10	4	9	0	0	0	9.16667	1	0	20	**1**	**0.63658**
11	3	20	1			8.83333	1	1	23	1	
12	3	15	0	0	1	9.16667	1	1	20	0	
13	3	19.5727	1	0	0	8.58333	0	0	19	**1**	**0.60554**
14	1	20.69763	0	0	0	8.75	0	0	22	1	
15	3	21	0			9.83333	0	1	23	**1**	**0.58324**
16	2	16	1	1	0	9	0	1	20	0	
17	1	18	1	1	0	8.66667	0	0	19	**1**	**0.6844**
18	0	21.29396	0			9.33333	0	0	22	1	
19	3	19	1	1	0	8.16667	0	1	18	0	

From the above table we note the following: When PI_MOMWORK has a probability that is greater than 0.5, the variable MOMWORK will equal 1. Similarly, when PI_MOMWORK has a probability of less than 0.5, MOMWORK will equal 0. Continuing in this way, the variable MOMWORK will no longer contain missing values and will hence be complete.

4.6.4 RACE

Recall that BLACK and HISPANIC are two categories of a three-category variable with the reference category being NON-HISPANIC WHITE. Thus, for this example, one variable called RACE will be modelled on which will take on the values BLACK (B) , HISPANIC (H) or NON-HISPANIC WHITE (W).

In order to proceed with the analysis, multinomial logistic regression will be used. According to Chan (2005), multinomial logistic regression is an extension of logistic regression which allows for the analysis of categorical dependent variables with more than two outcomes.

Consider a random variable Y_i that may take one of several discrete values, which we index $1, 2, ..., K$. In this case, the response variable is RACE and it takes on the values BLACK, HISPANIC or NON-

HISPANIC WHITE. Let

$$P(Y_i = K) \tag{4.16}$$

denote the probability that the ith response falls in the Kth category of the variable RACE.

In order to determine the multinomial model, one can look at K possible outcomes, where $K-1$ independent binary logistic regression models are fit, in which one outcome is selected to be the "pivot". Then, the rest of the $K-1$ outcomes can be separately regressed against the "pivot" outcome.

To illustrate this, suppose outcome K (which is also the last outcome) is selected to be the "pivot". Then

$$ln\left(\frac{P(Y_i = 1)}{P(Y_i = K)}\right) = \beta_1 X_i \tag{4.17}$$

$$ln\left(\frac{P(Y_i = 2)}{P(Y_i = K)}\right) = \beta_2 X_i \tag{4.18}$$

$$\cdots\cdots\cdots\cdots$$

$$ln\left(\frac{P(Y_i = K-1)}{P(Y_i = K)}\right) = \beta_{K-1} X_i \tag{4.19}$$

Next, taking the exponent of both sides and solving for the probabilities we get the following:

$$P(Y_i = 1) = P(Y_i = K)\exp\beta_1 X_i \tag{4.20}$$

$$P(Y_i = 2) = P(Y_i = K)\exp\beta_2 X_i \tag{4.21}$$

$$\cdots\cdots\cdots\cdots$$

$$P(Y_i = K-1) = P(Y_i = K)\exp\beta_{K-1} X_i \tag{4.22}$$

The K probabilities must sum to 1. Thus:

$$P(Y_i = 1) + P(Y_i = 2) + \ldots\ldots + P(Y_i = K-1) + P(Y_i = K) = 1 \tag{4.23}$$

It follows that:

$$P(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp\beta_k X_i} \tag{4.24}$$

Once we have Equation 4.24 above, it can be used to obtain the rest of the probabilities:

$$P(Y_i = 1) = \frac{\exp \beta_1 X_i}{1 + \sum_{k=1}^{K-1} \exp \beta_k X_i} \tag{4.25}$$

$$P(Y_i = 2) = \frac{\exp \beta_2 X_i}{1 + \sum_{k=1}^{K-1} \exp \beta_k X_i} \tag{4.26}$$

.....

$$P(Y_i = K-1) = \frac{\exp \beta_{K-1} X_i}{1 + \sum_{k=1}^{K-1} \exp \beta_k X_i} \tag{4.27}$$

That is, for the variable RACE, which has 3 categories, $P(Y = 1)$ refers to the probability of being BLACK, $P(Y = 2)$ refers to the probability of being HISPANIC and $P(Y = 3)$ refers to the probability of being NON-HISPANIC WHITE.

Forward selection was used as an option in the logistic regression model. According to Hand et al. (2001), forward selection allows variables to be added one at a time to the model. Every variable is tested at each step in order to determine whether or not it will be included in the model.

For this regression imputation example, only the variables that were significant at the 0.10 level were chosen for the model. That is, only the variables ANTI and MOMAGE were included. The results from the logistic regression can be seen in Table 4.20 below. Note that Intercept_B refers to the intercept for BLACK and Intercept_H refers to the intercept for HISPANIC.

Table 4.20: Table displaying the significant parameter estimates for logistic regression using forward selection.

Variable	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept_B	1	1.5936	0.8261	3.7217	0.0537
Intercept _H	1	2.6119	0.8318	9.86	0.0017
ANTI	1	0.1121	0.0577	3.7731	0.0521
MOMAGE	1	-0.1144	0.0397	8.3068	0.0039

Next, Equations 4.24 - 4.27 were used to calculate the values for three new variables, namely PI_WHITE, PI_BLACK and PI_HISP. These variables will only be calculated when the variable RACE is missing. The variable called FILLED_IN_RACE determines which variable has the highest probability and hence determines the individual's race. For example, consider Table 4.21 below:

Table 4.21: Table displaying a snapshot of the results from regression imputation for the variable RACE.

ID	RACE	PI_WHITE	PI_BLACK	PI_HISP	FILLED_IN_RACE
1	W				
2	.	0.40039	0.15911	**0.4405**	**H**
3	.	0.19438	0.21378	**0.59185**	**H**
4	W				

Note that a "." indicates a missing value and the values highlighted in bold indicate where the highest probabilities occur. The aim here is to replace all missing values for the variable RACE. It can be seen that the first and fourth individuals have complete data for RACE and are classified as NON-HISPANIC WHITE. The variables PI_WHITE, PI_BLACK and PI_HISP all have values for the second and third individuals who do not have a value for RACE. The FILLED_IN_RACE determines the race of an individual based on which variable has the highest probability. That is, based on Table 4.21 above, for the individual that has a unique ID of 2, the highest probability is 0.4405 and hence this individual's race is HISPANIC. The final table will look similar to Table 4.22 below:

Table 4.22: Table displaying a snapshot of the final results with filled-in values for the variables BLACK and HISPANIC.

ID	BLACK	HISPANIC	RACE	PI_WHITE	PI_BLACK	PI_HISP	FILLED_IN_RACE
1	0	0	W				
2	0	**1**	**H**	0.40039	0.15911	0.4405	**H**
3	0	**1**	**H**	0.19438	0.21378	0.59185	**H**
4	0	0	W				

The values highlighted in bold indicate where the variable RACE has been used in order to determine whether an individual is BLACK or HISPANIC. Consider again the individual with a unique ID of 2. Here, the variable RACE = H and hence the variable HISPANIC will equal 1 and the variable BLACK will equal 0.

Now that all the variables, in particular, SELF, POV, MOMWORK and RACE no longer contain missing values, a simple linear regression equation can be fit on the now complete data set using ANTI as the dependent variable and the rest of the variables as predictors. The results of the linear regression can be seen in Table 4.23 below:

Table 4.23: Table displaying regression results for the variable ANTI on the now complete data set.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	2.84561	1.36568	2.08	0.0377
SELF	**1**	**-0.07552**	**0.02413**	**-3.13**	**0.0019**
POV	1	0.07876	0.13554	0.58	0.5615
BLACK	1	0.36251	0.1701	2.13	0.0336
HISPANIC	1	0.03124	0.15644	0.2	0.8418
CHILDAGE	1	0.02324	0.11152	0.21	0.835
DIVORCE	1	0.18235	0.15408	1.18	0.2372
GENDER	**1**	**-0.52789**	**0.12925**	**-4.08**	**<.0001**
MOMAGE	1	-0.00126	0.03112	-0.04	0.9678
MOMWORK	**1**	**0.36425**	**0.13894**	**2.62**	**0.009**

From the above table it is clear that the variables SELF, GENDER and MOMWORK (highlighted in bold) are significant at the 0.05 level. That is, high levels of antisocial behaviour are associated with low levels of self-esteem, being male as well as if the child's mother is employed.

As already mentioned in Chapter 3, in order for regression imputation to be effective, correlation must exist among the variables in the data set. In the NLSY data set, however, this is not always the case. According to Acuña & Rodriguez (2004), if correlation does not exist and there are no relationships between the variables in the data set and the variables with missing values, then the model will not estimate these missing values accurately. Further, this single imputation technique does not reflect the uncertainty in the missing data estimates and hence the estimates will be biased if a random error term has not been incorporated into the model.

Importantly, not one of the techniques described thus far have satisfied all of the criteria listed by Allison (2001) in Chapter 3 for evaluating missing data methods. Although some of the techniques do not discard any data, they produce either biased or inaccurate estimates that have large standard errors. These methods should hence be avoided in practice.

4.7 KNNI

According to Inc. (2008a), when a data set consists of quantitative variables, as well as a variable that classifies groups of observations, the DISCRIM function in SAS can be used to derive a discriminant criterion. This discriminant criterion is used to classify each observation into specific groups and can then be used on a second data set during the same iteration of the DISCRIM procedure. The discriminant criterion is calculated on the training or calibration data set.

In order to estimate the group-specific densities, one has a choice of using either parametric or non-parametric methods. Parametric methods assume that the distribution within each group is multivariate normal, whereas nonparametric methods either assume that no assumptions can be made about the distribution or that the distribution is not multivariate normal. Nonparametric methods use kernels as well as k-nearest neighbour techniques. In order to estimate density, the DISCRIM procedure makes use of a number of kernels, in particular the uniform, normal, Epanechnikov, biweight or triweight kernels (Inc. (2008a)).

In order to determine proximity, one has the choice of using either Mahalanobis or Euclidean distance. Mathematically speaking, according to Farber & Kadmon (2002), the Mahalanobis distance between a vector $\mathbf{x}$ and a set S of vectors can be defined as follows:

$$D^2 = (\mathbf{x}-\mathbf{m})^T C^{-1} (\mathbf{x}-\mathbf{m}) \tag{4.28}$$

where $\mathbf{m}$ is the mean vector and C is the covariance matrix of S. In order to denote the transpose the superscript "T" is used. When $C = I$, the Euclidean distance is obtained:

$$D = \sqrt{(\mathbf{x}-\mathbf{m})^T (\mathbf{x}-\mathbf{m})} \tag{4.29}$$

If Mahalanobis distance is the chosen metric, either the full covariance matrix or the diagonal matrix of variances can be used to calculate the distance. When the k-nearest neighbour technique is used, in order to calculate the Mahalanobis distance, the pooled covariance matrix should be used. However, when the kernel method is used, one has a choice of using either the pooled covariance matrix or the individual within-group covariance matrices (Inc. (2008a)). For this analysis, the KNNI technique will only be performed on the variables that contain missing values.

4.7.1 SELF

For the first variable SELF, the data set is split into two parts - a data set for the missing values as well as a complete data set where there are no missing values. For example, when the variable SELF is missing, the data set "nlsy_miss_DM" is outputted and when the variable SELF is complete, the data set "nlsy_miss_DC" is outputted. The code used to perform KNNI may be found in Appendix A.8.

According to Inc. (2008a), the function TESTOUT in the DISCRIM procedure creates a new output data set. This new data set includes data from the input data set as well as information about which class each of the observations were grouped into. The METHOD statement specifies either

parametric or nonparametric methods - in this case nonparametric methods were used. Further, k specifies the number of k-nearest neighbours. According to Acuña & Rodriguez (2004), if the data set is small, choosing a k that is smaller than 10 will be satisfactory. Hence, for this dissertation, $k = 5$ was used. The CLASS statement specifies the number of classes in the classification variable and the VAR statement lists all the quantitative variables that are to be included in the study (Inc. (2008a)). Further, the THRESHOLD statement determines the minimum threshold that will be used in order to classify the objects and once again the default threshold was used, that is THRESHOLD = 0. Table 4.24 below shows the first 10 rows of output from the DISCRIM procedure for the variable SELF:

Table 4.24: Table displaying the first 10 rows of output from the DISCRIM procedure for the variable SELF.

ID	SELF	9	10	11	12	13	14	15	16	17	**18**	19	20	21	**22**	23	**24**	_INTO_
6	.	0	0	0	0	0	0	0	0	0	0.32191	0	0	0	0.20864	0	0.46945	**24**
13	.	0	0	0	0	0	0	0	0	0.31535	0	0.22075	0.1666	0.13379	0.16352	0	0	17
14	.	0	0	0	0	0.61348	0	0	0	0.10955	0	0	0.05788	0	0	0.2191	0	13
18	.	0	0	0	0	0	0	0.44742	0	0	0	0.17897	0	0.10847	0.26514	0	0	15
21	.	0	0	0	0	0	0	0	0	0.32297	0.25837	0	0	0	0.16746	0	0.2512	17
24	.	0	0	0	0	0	0	0	0	0.3123	0	0	0	0.13249	0	0.3123	0.2429	.
31	.	0	0	0	0.6761	0	0	0	0	0	0.1159	0	0.07654	0	0.07512	0	0.05634	12
33	.	0	0	0	0	0	0	0	0.44108	0	0.18904	0	0.12483	0	0.24505	0	0	16
34	.	0	0	0	0	0	0	0	0	0	0	0.72973	0	0	0	0	0.27027	19
50	.	0	0	0	0	0	0	0	0	0	0	0	0.22776	0.5487	0.22354	0	0	21

Note that the values highlighted in bold illustrate two examples (ID's 6 and 24) of how the _INTO_ field is used to calculate the nearest neighbours.

The _INTO_ field calculates the final neighbour based on the maximum probability in that row. For example, for the observation that has a unique ID of 6, probabilities are given for columns 18, 22 and 24. The highest probability is 0.46945, given in column 24 and hence _INTO_ outputs 24. Next, for the observation with a unique ID of 24, the field _INTO_ is missing since the highest probability of 0.3123 occurs for both neighbours 17 and 23. According to Khattree & Naik (2000), in the case of a tie, one can allocate x to any of the tied populations. In this case, 17 was chosen. Now that _INTO_ has values for each entry, these values will be used to replace the missing values occurring for the variable SELF.

4.7.2 POV and MOMWORK

For the next two variables, POV and MOMWORK, a similar procedure to the above was performed. Table 4.25 below shows the first 10 rows of output from the DISCRIM procedure for the variable POV:

Table 4.25: Table displaying the first 10 rows of output from the DISCRIM procedure for the variable POV.

ID	POV	0	1	_INTO_
2	.	0.68104	0.31896	**0**
9	.	1	0	0
12	.	0.26247	0.73753	**1**
16	.	0.26247	0.73753	1
17	.	0.26247	0.73753	1
23	.	1	0	0
30	.	0.72745	0.27255	0
37	.	0.11774	0.88226	1
43	.	1	0	0
44	.	0.68104	0.31896	0

Note that the values highlighted in bold illustrate two examples (ID's 2 and 12) of how the _INTO_ field is used to calculate the nearest neighbours.

Again, _INTO_ calculates the final neighbour based on the maximum probability in that row. In this case POV can either take on values 0 or 1. For the observation that has a unique ID of 2, it is clear that the highest probability of 0.68104 is given when POV $= 0$ and this is why _INTO_ results in a 0. As another example, consider the observation that has a unique ID of 12. In this case, the highest probability of 0.73753 is given when POV $= 1$ and hence _INTO_ results in a 1. Continuing in this way, the variable POV will no longer contain missing values.

In the next table, the first 10 rows of output from the DISCRIM procedure for the variable MOMWORK are given:

Table 4.26: Table displaying the first 10 rows of output from the DISCRIM procedure for the variable MOMWORK.

ID	MOMWORK	0	1	_INTO_
7	.	0.25171	0.74829	**1**
10	.	0.66868	0.33132	**0**
22	.	0.4308	0.5692	1
40	.	0.66868	0.33132	0
57	.	1	0	0
70	.	0.66868	0.33132	0
77	.	0.66868	0.33132	0
83	.	0	1	1
88	.	0.4308	0.5692	1
94	.	1	0	0

Note that the values highlighted in bold illustrate two examples (ID's 7 and 10) of how the _INTO_ field is used to calculate the nearest neighbours.

Similarly, the variable MOMWORK also only takes on values 0 or 1 and hence for the observation that has a unique ID of 7, it is clear that the highest probability of 0.74829 is given when MOMWORK $= 1$ and this is why _INTO_ results in a 1. The observation with a unique ID of 10 has its highest probability when MOMWORK $= 0$ and hence _INTO_ results in a 0. Continuing in this way, the variable MOMWORK will no longer contain missing values.

4.7.3 RACE

The last two variables, BLACK and HISPANIC, are grouped into one category called RACE (similar to regression imputation above). The following rules apply:

If

$$\text{RACE} = \text{B then BLACK} = 1 \tag{4.30}$$

else

$$\text{BLACK} = 0 \tag{4.31}$$

If

$$\text{RACE} = \text{H then HISPANIC} = 1 \tag{4.32}$$

else

$$\text{HISPANIC} = 0 \tag{4.33}$$

If

$$\text{RACE} = \text{W} \tag{4.34}$$

then

$$\text{BLACK} = 0 \text{ and } \text{HISPANIC} = 0 \tag{4.35}$$

The DISCRIM function is performed in a similar way as it was for the other variables but in this case it is performed on the variable RACE. The results for the first 10 observations are given in Table 4.27 below:

Table 4.27: Table displaying the first 10 rows of output from the DISCRIM procedure for the variable RACE.

ID	RACE	B	H	W	_INTO_
3	.	0.37878	0.2791	0.34212	**B**
8	.	0.37878	0.2791	0.34212	B
11	.	0.37878	0.2791	0.34212	B
15	.	0.19292	0.28431	0.52276	**W**
49	.	0.55794	0.27408	0.16798	B
62	.	0.37878	0.2791	0.34212	B
74	.	0.57576	0.42424	0	B
75	.	0.42466	0	0.57534	W
91	.	0.34185	0.50377	0.15438	H
105	.	0.73077	0.26923	0	B

Note that the values highlighted in bold illustrate two examples (ID's 3 and 15) of how the _INTO_ field is used to calculate the nearest neighbours.

If the 1*st* entry with a unique ID of 3 is considered, the highest probability is 0.37878 and hence the individual's race is BLACK. For the fourth entry (unique ID of 15), the highest probability is 0.52276 and hence the individual's race is NON-HISPANIC WHITE. Then, using the results obtained in the _INTO_ field, the variables BLACK and HISPANIC will no longer contain missing values. That is, if _INTO_ results in a B, then RACE will equal B and thus the variables BLACK and HISPANIC will equal 1 and 0 respectively. Table 4.28 below provides a summary of what the final results will look like:

Table 4.28: Table displaying the final filled-in results for the variable RACE.

ID	RACE	BLACK	HISPANIC	B	H	W	_INTO_
3	BLACK	1	0	0.37878	0.2791	0.34212	B
8	BLACK	1	0	0.37878	0.2791	0.34212	B
11	BLACK	1	0	0.37878	0.2791	0.34212	B
15	NON-HISPANIC WHITE	0	0	0.19292	0.28431	0.52276	W
49	BLACK	1	0	0.55794	0.27408	0.16798	B
62	BLACK	1	0	0.37878	0.2791	0.34212	B
74	BLACK	1	0	0.57576	0.42424	0	B
75	NON-HISPANIC WHITE	0	0	0.42466	0	0.57534	W
91	HISPANIC	0	1	0.34185	0.50377	0.15438	H
105	BLACK	1	0	0.73077	0.26923	0	B

Now that the variables SELF, POV, BLACK, HISPANIC and MOMWORK no longer contain missing values, a linear regression can be fit to the complete data set. The results of the linear regression using ANTI as the dependent variable are given in Table 4.29 below:

Table 4.29: Table displaying regression results for the variable ANTI on the now complete data set.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	1.65458	1.22186	1.35	0.1762
SELF	1	-0.02557	0.01661	-1.54	0.1244
POV	**1**	**0.67776**	**0.13454**	**5.04**	**<.0001**
BLACK	1	0.17595	0.13875	1.27	0.2053
HISPANIC	1	-0.2175	0.15278	-1.42	0.1551
CHILDAGE	1	-0.01311	0.10043	-0.13	0.8962
DIVORCE	1	-0.12394	0.14587	-0.85	0.3959
GENDER	**1**	**-0.56024**	**0.11727**	**-4.78**	**<.0001**
MOMAGE	1	0.02264	0.02821	0.8	0.4226
MOMWORK	1	0.27689	0.12258	2.26	0.0243

From the above, it is clear that the variables highlighted in bold, POV and GENDER, are significant at the 0.05 level. That is, high levels of antisocial behaviour are associated with experiencing poverty as well as being male. Next, Table 4.30 below compares the standard errors on the complete data set to the standard errors when listwise deletion as well as KNNI was used:

Table 4.30: Table comparing the standard errors obtained from (a) the complete data set, (b) listwise deletion and (c) KNNI.

Standard Errors			
Variable	Complete Data	Listwise Deletion	KNNI
SELF	0.01863	0.03135	0.01661
POV	0.13896	0.23739	0.13454
BLACK	0.14114	0.24918	0.13875
HISPANIC	0.15335	0.25537	0.15278
CHILDAGE	0.10049	0.17072	0.10043
DIVORCE	0.14395	0.24499	0.14587
GENDER	0.11728	0.19844	0.11727
MOMAGE	0.02815	0.04611	0.02821
MOMWORK	0.12925	0.21751	0.12258

It can be seen from the above table that the standard errors from the KNNI data set are reasonably close to the standard errors obtained from using the complete data set. Again, listwise deletion produces the largest standard errors out of all three methods. Using the KNNI method when the data set is large is not a good idea, as trying to search through each point to find the k-nearest neighbour can be extremely time consuming. This method may also not produce accurate results if enough time has not been spent on choosing the correct k.

4.8 EM Algorithm

In this section the EM algorithm is applied to the NLSY data set. According to Allison (2001), although the categorical variables in the NLSY data are treated as dummy variables and hence do not follow a normal distribution, the EM algorithm will still produce satisfactory results. Schafer & Graham (2002) state that "when missingness is not controlled by the researcher, it is unlikely that MAR is precisely satisfied. In many realistic applications, however, we believe that departures from MAR are not large enough to effectively invalidate the results of a MAR-based analysis." The analysis below assumes the data is MAR and multivariate normal.

The EM algorithm was performed using SAS and the code may be found in Appendix A.9. Using the EM statement allows the EM algorithm to determine maximum likelihood estimates (MLE) from the data set that contains missing values. That is, the EM algorithm calculates the mean and covariance matrix, (μ, Σ) from a multivariate normal distribution (Inc. (2013)). A choice of using the means and covariances from either complete cases or available cases can be made and these will then be used as the initial estimates for the EM algorithm. PROC MI uses the means and standard deviations from

available cases as the initial estimates. The initial parameter estimates are given in Table 4.31 below:

Table 4.31: Table displaying initial parameter estimates for EM.

TYPE	_NAME_	ANTI	SELF	POV	BLACK	HISPANIC	CHILDAGE	DIVORCE	GENDER	MOMAGE	MOMWORK
MEAN		1.56799	20.05081	0.34803	0.35897	0.24359	8.94363	0.23580	0.50430	20.66577	0.33535
COV	ANTI	2.16304									
COV	SELF		9.73352								
COV	POV			0.22743							
COV	BLACK				0.23061						
COV	HISPANIC					0.18465					
COV	CHILDAGE						0.36163				
COV	DIVORCE							0.18051			
COV	GENDER								0.25041		
COV	MOMAGE									4.79164	
COV	MOMWORK										0.22334

The NIMPUTE = 0 statement provides the user with the option of calculating EM estimates without using multiple imputation. The ITPRINT statement provides a print out of the iteration history (Inc. (2013)). In Table 4.32 below, the iteration history is given:

Table 4.32: Table displaying EM (MLE) iteration history.

Iteration	-2 Log L	SELF	POV	BLACK	HISPANIC	MOMWORK
0	2 286.36351	20.05081	0.34803	0.35897	0.24359	0.33535
1	1 970.74783	20.05081	0.34803	0.35897	0.24359	0.33535
2	1 947.33590	20.10496	0.34257	0.35868	0.24281	0.33504
3	1 944.21167	20.13193	0.34004	0.35867	0.24239	0.33505
4	1 943.67347	20.14328	0.33916	0.35877	0.24221	0.33516
5	1 943.56960	20.14790	0.33889	0.35885	0.24214	0.33525
6	1 943.54855	20.14976	0.33881	0.35889	0.24211	0.33530
7	1 943.54418	20.15051	0.33879	0.35891	0.24209	0.33533
8	1 943.54326	20.15081	0.33879	0.35892	0.24209	0.33534
9	1 943.54306	20.15093	0.33879	0.35893	0.24209	0.33535
10	1 943.54302	20.15098	0.33879	0.35893	0.24209	0.33535
11	1 943.54301	20.15099	0.33880	0.35893	0.24208	0.33535
12	1 943.54301	20.15100	0.33880	0.35893	0.24208	0.33535
13	1 943.54301	20.15100	0.33880	0.35893	0.24208	0.33535

The CONVERGE statement determines the convergence criteria and this value will always be between 0 and 1. In this dissertation, the default was used which is CONVERGE = 1E-4. Using a SEED option will allow the user to duplicate results under identical situations. Again, the default value was used

where an initial seed is generated using the time of day from the computer's clock.

According to Inc. (2013), the OUTEM statement creates an output SAS data set containing the parameter estimates from the EM algorithm. The data set that is produced is a _TYPE_ = COV data set which contains the estimated covariances. Observations with a _TYPE_ = MEAN data set will contain the estimated mean. The results can be seen in Table 4.33 below:

Table 4.33: Table displaying EM (MLE) parameter estimates.

OBS	_TYPE_	_NAME_	ANTI	SELF	POV	BLACK	HISPANIC	CHILDAGE	DIVORCE	GENDER	MOMAGE	MOMWORK
1	MEAN		1.56799	20.151	0.3388	0.35893	0.24208	8.94363	0.2358	0.5043	20.6558	0.33535
2	COV	ANTI	2.15932	-0.6485	0.15598	0.08251	-0.0478	0.01394	0.01925	-0.12637	-0.0971	0.07442
3	COV	SELF	-0.64846	9.7155	-0.11809	-0.09561	-0.12983	0.13871	-0.14585	-0.03042	0.6191	0.00795
4	COV	POV	0.15598	-0.1181	0.22408	0.06163	-0.00091	0.02293	0.05243	0.00726	-0.2043	0.05322
5	COV	BLACK	0.08251	-0.0956	0.06163	0.2301	-0.08719	0.005	0.00385	0.00837	-0.1348	-0.01675
6	COV	HISPANIC	-0.0478	-0.1298	-0.00091	-0.08719	0.18401	-0.01201	0.00741	-0.01486	-0.0059	0.01687
7	COV	CHILDAGE	0.01394	0.1387	0.02293	0.005	-0.01201	0.36101	0.00354	-0.0222	-0.3071	-0.00087
8	COV	DIVORCE	0.01925	-0.1458	0.05243	0.00385	0.00741	0.00354	0.1802	-0.00015	-0.0582	-0.00966
9	COV	GENDER	-0.12637	-0.0304	0.00726	0.00837	-0.01486	-0.0222	-0.00015	0.24998	0.0325	0.00397
10	COV	MOMAGE	-0.09708	0.6191	-0.2043	-0.13477	-0.00589	-0.30713	-0.05824	0.03246	4.7834	-0.01082
11	COV	MOMWORK	0.07442	0.0079	0.05322	-0.01675	0.01687	-0.00087	-0.00966	0.00397	-0.0108	0.22324

Often, it can be difficult to interpret covariances but one way to solve this issue is to convert the covariance matrix into a correlation matrix. According to Allison (2001), an attractive property of maximum likelihood estimates is that any function of those estimates will also be a maximum likelihood estimate of the corresponding function in the population. That is, if s_i is the MLE of the standard deviation of x_i and s_{ij} is the MLE of the covariance between x_i and x_j, then

$$r = \frac{s_{ij}}{\left(s_i s_j\right)} \tag{4.36}$$

is the maximum likelihood estimate of their correlation. Table 4.34 shows the maximum likelihood estimates of the correlations:

Table 4.34: Table displaying EM estimates of the correlation matrix.

	ANTI	SELF	POV	BLACK	HISPANIC	CHILDAGE	DIVORCE	GENDER	MOMAGE	MOMWORK
ANTI	1	-0.14158	0.22423	0.11705	-0.07583	0.01579	0.03086	-0.17200	-0.03021	0.10719
SELF	-0.14158	1	-0.08004	-0.06395	-0.09710	0.07407	-0.11023	-0.01952	0.09081	0.00540
POV	0.22423	-0.08004	1	0.27140	-0.00448	0.08062	0.26090	0.03067	-0.19733	0.23795
BLACK	0.11705	-0.06395	0.27140	1	-0.42373	0.01734	0.01889	0.03489	-0.12846	-0.07392
HISPANIC	-0.07583	-0.09710	-0.00448	-0.42373	1	-0.04659	0.04071	-0.06928	-0.00628	0.08322
CHILDAGE	0.01579	0.07407	0.08062	0.01734	-0.04659	1	0.01387	-0.07391	-0.23372	-0.00305
DIVORCE	0.03086	-0.11023	0.26090	0.01889	0.04071	0.01387	1	-0.00073	-0.06274	0.04814
GENDER	-0.17200	-0.01952	0.03067	0.03489	-0.06928	-0.07391	-0.00073	1	0.02969	- 0.01681
MOMAGE	-0.03021	0.09081	-0.19733	-0.12846	-0.00628	-0.23372	-0.06274	0.02969	1	-0.01047
MOMWORK	0.10719	0.00540	0.23795	-0.07392	0.08322	-0.00305	-0.04814	0.01681	-0.01047	1

Next, in order to estimate the regression of ANTI on other variables, the EM estimates from the above table can be used. Usually, regression programs allow the user to either use the covariance or correlation matrix as inputs. Table 4.35 below shows the results when the EM estimates are used as inputs:

Table 4.35: Table displaying regression results for the variable ANTI using the EM estimates as inputs.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	2.59110	0.28964	8.95	<.0001
SELF	1	-0.06718	0.00454	-14.79	<.0001
POV	1	0.64626	0.03357	19.25	<.0001
BLACK	1	0.08501	0.03395	2.50	0.0123
HISPANIC	1	-0.32439	0.03636	-8.92	<.0001
CHILDAGE	1	-0.00386	0.02391	-0.16	0.8716
DIVORCE	1	-0.10598	0.03422	-3.10	0.0020
GENDER	1	-0.56116	0.02783	-20.16	<.0001
MOMAGE	1	0.02076	0.00668	3.11	0.0019
MOMWORK	1	0.21896	0.03074	7.12	<.0001

From the table above, the parameter estimates are the true maximum likelihood estimates of the regression coefficients, however, the p-values are not accurate. One of the issues with this two-step approach is that it has difficulty obtaining accurate standard errors. Traditional regression software programs fail to produce accurate standard errors and as such the EM algorithm requires one to specify a sample size. In fact, the EM algorithm experiences similar issues as pairwise deletion. According to Allison (2012), the output data set resulting from the PROC MI statement can be modified in such a way in order to stipulate a sample size. However, deciding on this sample size may be tricky as one cannot choose the original sample size of 581 as then the sample would no longer contain missing

values. Also, this would yield significantly small standard errors and p-values. Since there is no sample size that will yield accurate standard errors for all coefficients, Allison (2012) suggests using the bootstrapping technique which can be summarised into the following four steps:

1. Draw several samples of size N (with replacement) from the original sample.
2. Determine the EM estimates of the mean vector and covariance matrix for every sample.
3. Use the covariance matrix in order to estimate the regression model.
4. Determine the standard deviation of every regression coefficient across each of the samples.

The results from the bootstrapping method can be seen in Table 4.36 below:

Table 4.36: Table displaying ML coefficients using bootstrapping to obtain standard errors.

Variable	Coefficient	Standard Error	z	Pr >| t|
SELF	**-0.06718**	**0.02293**	**-2.92979**	**0.00339**
POV	**0.64626**	**0.16813**	**3.84381**	**0.00012**
BLACK	0.08501	0.16092	0.52827	0.59731
HISPANIC	-0.32439	0.16559	-1.959	0.05011
CHILDAGE	-0.00386	0.09715	-0.03973	0.96831
DIVORCE	-0.10598	0.15294	-0.69295	0.48834
GENDER	**-0.56116**	**0.11667**	**-4.80981**	**<.0001**
MOMAGE	0.02076	0.02724	0.76211	0.44599
MOMWORK	0.21896	0.14659	1.49369	0.13526

It can be seen from the above table that the coefficients are the same as the ML estimates in Table 4.35. Further, in order to calculate the z-statistics in Table 4.36 above, the coefficients should be divided by its standard error. The variables that are significant at the 0.05 level and are highlighted in bold above include SELF, POV and GENDER. Hence, a conclusion can be made that antisocial behaviour is linked to low levels of self-esteem, being in poverty as well as being male.

Next, Table 4.37 shows the standard errors obtained from the complete data set, listwise deletion, as well as the EM algorithm.

Table 4.37: Table displaying the standard errors obtained from (a) the complete data set, (b) listwise deletion and (c) the EM algorithm.

Standard Errors			
Variable	Complete Data	Listwise Deletion	EM Algorithm
SELF	0.01863	0.03135	0.02293
POV	0.13896	0.23739	0.16813
BLACK	0.14114	0.24918	0.16092
HISPANIC	0.15335	0.25537	0.16559
CHILDAGE	0.10049	0.17072	0.09715
DIVORCE	0.14395	0.24499	0.15294
GENDER	0.11728	0.19844	0.11667
MOMAGE	0.02815	0.04611	0.02724
MOMWORK	0.12925	0.21751	0.14659

It can be seen from the above table that the standard errors obtained from the EM algorithm (using bootstrapping) are reasonably close to the standard errors obtained from using the complete data set. Again, listwise deletion produces the largest standard errors out of all three methods. Since the EM algorithm has difficulty calculating accurate standard errors, alternative methods such as direct maximum likelihood or multiple imputation should be used.

4.9 Multiple Imputation

The next method, multiple imputation, is one of the most popular methods used to handle missing values. Multiple imputation in SAS requires three steps (imputation, analysis and pooling) and these three steps can all be performed in one single software package. This is much preferred over a standalone imputation package, as repeatedly moving the data sets between packages can become a monotonous and time consuming task (Allison (2001)).

In the imputation step, the statement PROC MI is used in order to trigger the MI procedure. Here, the user will decide on the imputation model that will be used as well as the number of imputed data sets that will be created. The NIMPUTE statement stipulates the number of imputations that will be carried out. Based on the relative efficiency results in Table 3.2, 15 imputations will be used. The OUT= statement outputs the imputed data sets and then stacks them together in a data set called "mi_miss_out". A variable called _IMPUTATION_ is created automatically and is used to number each imputed data set - in this case the data sets will be numbered from 1-15. The VAR statement provides a list of all the specific variables that will be analysed in the imputation procedure (Allison (2001)). An option to use a SEED is not compulsory, however, because multiple imputation yields different results each time, using this option will allow the user to obtain the same results each time the MI procedure is run.

Next, the MI procedure organises the data into groups. These groups are created based on whether the analysis variables contain complete or incomplete data. The "missing data patterns" table is one of the outputs from running the PROC MI statement. This table provides distinct missing data patterns together with their corresponding frequencies and percentages (Inc. (2013)). The value "X" represents an observed variable and a "." indicates that the variable is incomplete. This table also displays the group specific variable means. Table 4.38 below shows a snapshot of the output when PROC MI is run. Here, only the missing data patterns are displayed and not the group means.

Table 4.38: Table displaying distinct missing data patterns from the PROC MI statement.

Group	ANTI	SELF	POV	BLACK	HISPANIC	CHILDAGE	DIVORCE	GENDER	MOMAGE	MOMWORK	Freq	Percent
1	X	X	X	X	X	X	X	X	X	X	225	38.73
2	X	X	X	X	X	X	X	X	X	.	41	7.06
3	X	X	X	.	.	X	X	X	X	X	48	8.26
4	X	X	X	.	.	X	X	X	X	.	8	1.38
5	X	X	.	X	X	X	X	X	X	X	77	13.25
6	X	X	.	X	X	X	X	X	X	.	8	1.38
7	X	X	.	.	.	X	X	X	X	X	22	3.79
8	X	X	.	.	.	X	X	X	X	.	4	0.69
9	X	.	X	X	X	X	X	X	X	X	71	12.22
10	X	.	X	X	X	X	X	X	X	.	13	2.24
11	X	.	X	.	.	X	X	X	X	X	22	3.79
12	X	.	X	.	.	X	X	X	X	.	3	0.52
13	X	.	.	X	X	X	X	X	X	X	24	4.13
14	X	.	.	X	X	X	X	X	X	.	9	1.55
15	X	.	.	.	.	X	X	X	X	X	6	1.03

Another output from PROC MI is a list of all 15 imputations. In Table 4.39 below, only the first 2 rows of each imputation can be seen. The variables highlighted in bold are the variables that contain missing values.

Table 4.39: Table displaying the first two rows of output for each of the 15 imputations.

IMPUTATION	ANTI	**SELF**	**POV**	**BLACK**	**HISPANIC**	CHILDAGE	DIVORCE	GENDER	MOMAGE	**MOMWORK**
1	1	21	1	0	0	8	0	1	21	0
1	0	20	0.24084	0.29614	0.84022	8.41667	0	1	22	1
2	1	21	1	0	0	8	0	1	21	0
2	0	20	0.91592	0.60237	0.95511	8.41667	0	1	22	1
3	1	21	1	0	0	8	0	1	21	0
3	0	20	0.54945	0.36996	0.13863	8.41667	0	1	22	1
4	0	20	0.19947	0.41803	0.34498	8	0	1	22	1
4	5	21	0	0.30704	0.5591	8.41667	1	0	18	0
5	1	21	1	0	0	8	0	1	21	0
5	0	20	0.40577	0.11495	0.47427	8.41667	0	1	22	1
6	1	21	1	0	0	8	0	1	21	0
6	0	20	0.5842	0.50363	0.19556	8.41667	0	1	22	1
7	1	21	1	0	0	8	0	1	21	0
7	0	20	0.48376	0.10324	0.33994	8.41667	0	1	22	1
8	1	21	1	0	0	8	0	1	21	0
8	0	20	0.96724	0.47045	0.29026	8.41667	0	1	22	1
9	1	21	1	0	0	8	0	1	21	0
9	0	20	0.68294	0.61397	0.54849	8.41667	0	1	22	1
10	1	21	1	0	0	8	0	1	21	0
10	0	20	0.75271	0.29465	0.90009	8.41667	0	1	22	1
11	1	21	1	0	0	8	0	1	21	0
11	0	20	0.95188	0.11386	0.72906	8.41667	0	1	22	1
12	1	21	1	0	0	8	0	1	21	0
12	0	20	0.05019	0.26484	0.24604	8.41667	0	1	22	1
13	1	21	1	0	0	8	0	1	21	0
13	0	20	0.62255	0.37334	1.21218	8.41667	0	1	22	1
14	1	21	1	0	0	8	0	1	21	0
14	0	20	1.13663	0.36784	1.01145	8.41667	0	1	22	1
15	1	21	1	0	0	8	0	1	21	0
15	0	20	1.20675	0.10881	0.62583	8.41667	0	1	22	1

Next, in order to determine the desired regression model, the PROC REG statement uses the "mi_miss_out" data set (Allison (2001)). According to Inc. (2013), the BY statement provides the user with the option of having separate analyses on observations in groups. These groups are determined by the BY variables and when this statement is used, the sorting of the input data set in order of the BY variables is very important. The OUTEST statement stores the parameter estimates resulting from the regression model into a new data set called "mi". The COVOUT statement provides the user with the option of incorporating the estimated covariance matrix in the "mi" data set.

Table 4.40 as well as Table 4.41 below show the regression results for the first two imputations only and the variables highlighted in bold are significant at the 0.05 level:

Table 4.40: Table displaying regression results for the variable ANTI for the first imputation.

Number of observations read	581
Number of observations used	581

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	2.90681	1.19237	2.44	0.0151
SELF	**1**	**-0.08014**	**0.01855**	**-4.32**	**<.0001**
POV	**1**	**0.69942**	**0.14225**	**4.92**	**<.0001**
BLACK	1	0.05626	0.1466	0.38	0.7013
HISPANIC	**1**	**-0.47371**	**0.14938**	**-3.17**	**0.0016**
CHILDAGE	1	-0.01802	0.09973	-0.18	0.8566
DIVORCE	1	-0.12064	0.14193	-0.85	0.3957
GENDER	**1**	**-0.56717**	**0.11563**	**-4.9**	**<.0001**
MOMAGE	1	0.02705	0.02775	0.97	0.33
MOMWORK	1	0.17706	0.1302	1.36	0.1744

Table 4.41: Table displaying regression results for the variable ANTI for the second imputation.

Number of observations read	581
Number of observations used	581

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >| t|
Intercept	1	2.60452	1.20665	2.16	0.0313
SELF	**1**	**-0.05676**	**0.01853**	**-3.06**	**0.0023**
POV	**1**	**0.69424**	**0.13432**	**5.17**	**<.0001**
BLACK	1	0.14875	0.13954	1.07	0.2869
HISPANIC	**1**	**-0.37125**	**0.154**	**-2.41**	**0.0162**
CHILDAGE	1	-0.02272	0.0999	-0.23	0.8201
DIVORCE	1	-0.11944	0.14214	-0.84	0.4011
GENDER	**1**	**-0.58718**	**0.11644**	**-5.04**	**<.0001**
MOMAGE	1	0.01871	0.02768	0.68	0.4993
MOMWORK	1	0.18601	0.12359	1.51	0.1329

Note, that there will be parameter estimates for each of the 15 imputations, however, only the results from the first two imputations will be shown in this dissertation.

The last step is known as the pooling phase which requires the user to make use of the PROC MI-ANALYZE statement. This statement uses the "mi" data set that contains the parameter estimates, as well as the covariance matrices for each of the 15 imputed data sets in order to make valid inferences (Inc. (2008c)). The variance/covariance matrix is required in order to estimate the standard errors.

Next, combining the parameter estimates into a single set of statistics will appropriately reflect the uncertainty associated with the imputed values (Graham et al. (2007)). The coefficients are calculated to be the mean of the individual coefficients that were estimated for each of the 15 regression models. Parameter estimates are then averaged, as this results in reducing the variance, which in turn increases efficiency and decreases sampling variation. Table 4.42 below displays the final regression estimates for ANTI:

Table 4.42: Table displaying regression results for the variable ANTI using multiple imputation.

Variable	Parameter Estimate	Standard Error	t for H0: Parameter=Theta0	Pr >| t|
Intercept	2.6579	1.22754	2.17	0.0304
SELF	**-0.06695**	**0.0215**	**-3.11**	**0.002**
POV	**0.63098**	**0.1625**	**3.88**	**0.0001**
BLACK	0.08813	0.15065	0.59	0.5586
HISPANIC	**-0.35921**	**0.1673**	**-2.15**	**0.0323**
CHILDAGE	-0.00727	0.1013	-0.07	0.9428
DIVORCE	-0.10476	0.14592	-0.72	0.4728
GENDER	**-0.56757**	**0.11848**	**-4.79**	**<.0001**
MOMAGE	0.01974	0.0282	0.7	0.484
MOMWORK	0.20564	0.13475	1.53	0.1272

It can be seen from the above, that SELF POV, HISPANIC and GENDER are significant at the 0.05 level and hence a conclusion can be made that antisocial behaviour is linked to low levels of self-esteem, being in poverty as well as being male. High levels of antisocial behaviour are also linked to whether the child is HISPANIC, where HISPANIC lowers the level of antisocial behaviour. Next, in order to illustrate the effectiveness of multiple imputation, consider Table 4.43 below which shows a comparison of the standard errors produced by the complete data set with no missing values, listwise and pairwise deletion, the EM algorithm as well as multiple imputation:

Table 4.43: Table displaying the standard errors for (a) the complete data set, (b) listwise deletion, (c) pairwise deletion, (d) EM algorithm and (e) multiple imputation.

Variable	Complete Data	Listwise Deletion	Pairwise Deletion	EM Algorithm	Multiple Imputation
SELF	0.01863	0.03135	0.02198	0.0224	0.02150
POV	0.13896	0.23739	0.16334	0.16621	0.16250
BLACK	0.14114	0.24918	0.16454	0.16812	0.15065
HISPANIC	0.15335	0.25537	0.17613	0.16313	0.16730
CHILDAGE	0.10049	0.17072	0.11593	0.10336	0.10130
DIVORCE	0.14395	0.24499	0.16659	0.15028	0.14592
GENDER	0.11728	0.19844	0.13510	0.11491	0.11848
MOMAGE	0.02815	0.04611	0.03242	0.02801	0.02820
MOMWORK	0.12925	0.21751	0.14985	0.14578	0.13475

It is clear that the standard errors produced by multiple imputation are very similar to those produced by the complete data set and much smaller than those produced by the listwise and pairwise deletion data sets. It can also be seen that there is a small inflation in the standard errors but this is to be expected since multiple imputation is designed to capture any additional uncertainty in the estimates (Allison (2001)).

In order to further illustrate how effective multiple imputation is, consider the graphs below where each graph shows how some of the missing data methods can affect confidence intervals. It is evident that the complete data set (no missing data) produces the shortest confidence intervals. Using the listwise deletion technique, however, results in the largest confidence intervals. Multiple imputation produces the smallest confidence intervals out of all three methods.

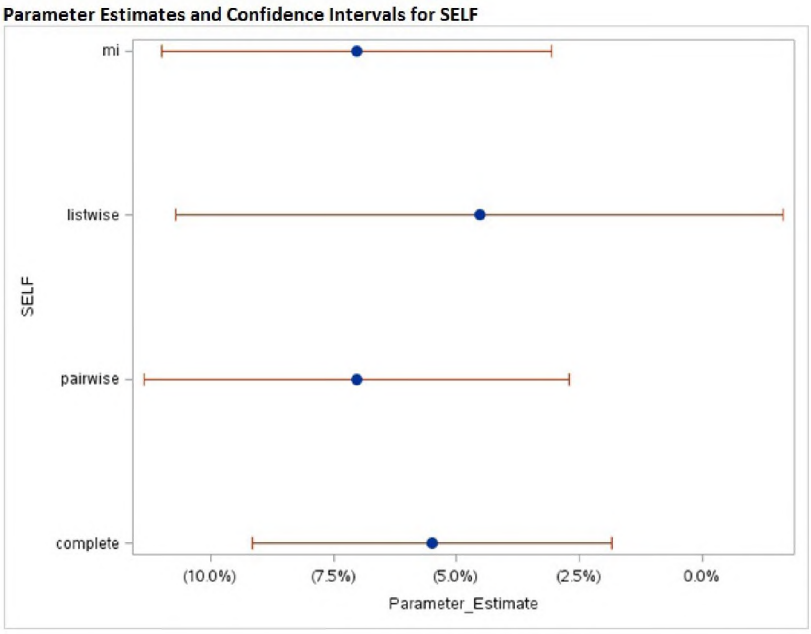

Figure 4.6: Parameter estimates and confidence intervals using the complete data set, pairwise deletion, listwise deletion and multiple imputation for the variable SELF.

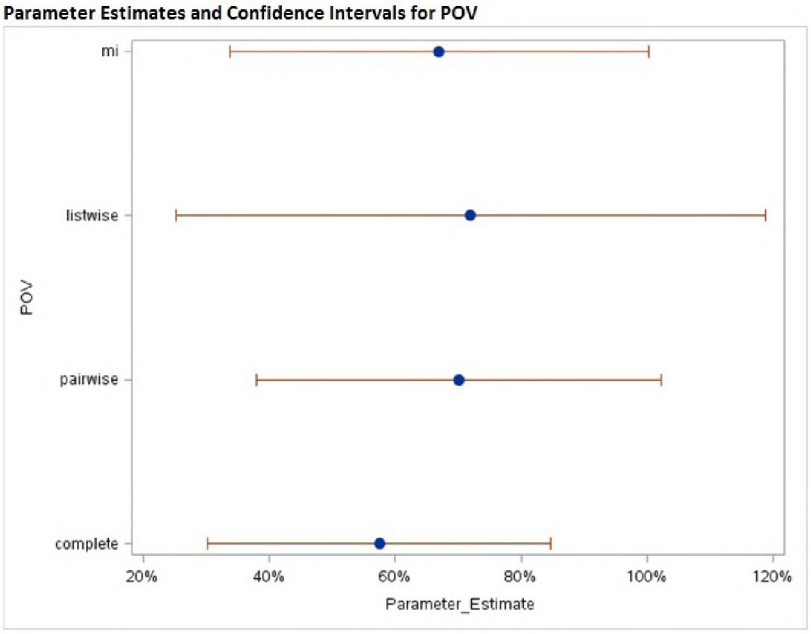

Figure 4.7: Parameter estimates and confidence intervals using the complete data set, pairwise deletion, listwise deletion and multiple imputation for the variable POV.

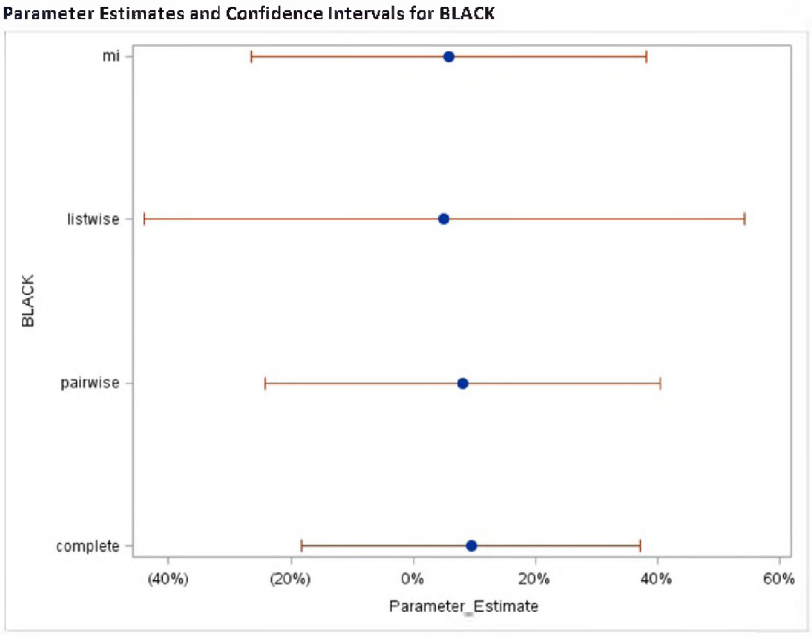

Figure 4.8: Parameter estimates and confidence intervals using the complete data set, pairwise deletion, listwise deletion and multiple imputation for the variable BLACK.

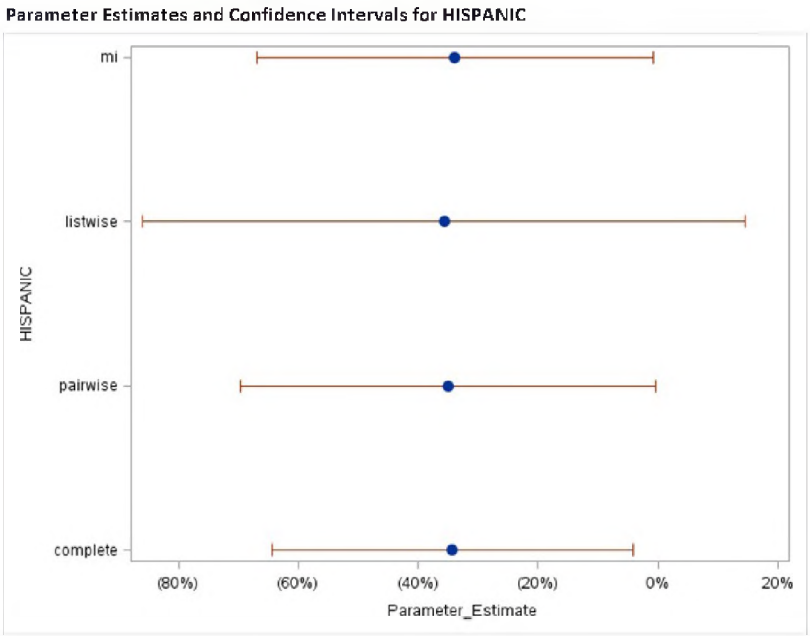

Figure 4.9: Parameter estimates and confidence intervals using the complete data set, pairwise deletion, listwise deletion and multiple imputation for the variable HISPANIC.

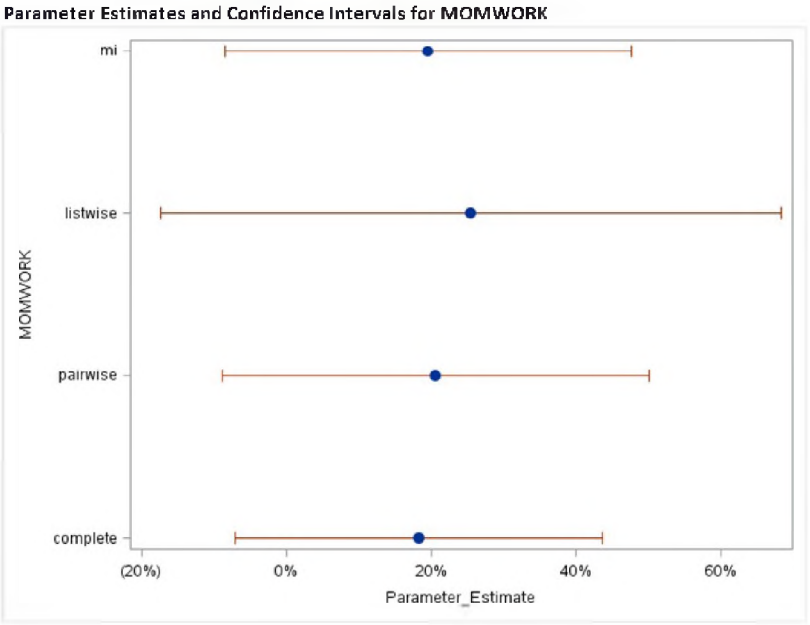

Figure 4.10: Parameter estimates and confidence intervals using the complete data set, pairwise deletion, listwise deletion and multiple imputation for the variable MOMWORK.

It is submitted that multiple imputation is much preferred over the conventional methods when treating missing values for many of the reasons mentioned in Chapter 3. This technique by far stands out as single imputation methods have many limitations. According to Graham et al. (2007), multiple imputation has the option of allowing plausible values to be used instead of missing values in such a way that will allow parameter estimates to be unbiased. Also, allowing the uncertainty of parameter estimation to be estimated in a reasonable manner is another big advantage of multiple imputation. Further, unlike the EM algorithm, multiple imputation produces standard errors that can be relied upon. There are, however, other maximum likelihood techniques such as direct maximum likelihood that can be investigated. Thus, it is primarily up to the individual when deciding on using either maximum likelihood or multiple imputation in order to handle missing values.

Chapter 5

Conclusion

Missing data is not a new concept and the ways in which researchers have dealt with this issue have continuously evolved over the years. This dissertation sought to examine the reasons that missing data seem to appear so frequently in most works and how to combat the potential problems it raises as a result.

The most common methods of dealing with missing data were discussed in this dissertation and these included deletion methods, such as pairwise and listwise deletion, which attempted to replace missing values with substituted data. Listwise deletion will produce biased estimates unless the data is MCAR. This method also discards a lot of the data which results in large standard errors, a loss of power in hypothesis testing as well as wider confidence intervals. Although pairwise deletion preserves all of the data, this technique yields biased parameter estimates. This is because different analyses are based on different subsets of the data and hence there is a lack of consistency. Pairwise deletion also has difficulty in computing accurate standard errors and correlations as a sample size needs to be specified which is not always easy. Further, pairwise deletion is only valid when the data is MCAR and will produce biased estimates otherwise.

Single imputation techniques such as mean/mode/median, regression imputation as well as k-nearest neighbour were also studied. However, the results from these types of methods were not conclusive as they tended to yield biased parameter estimates, biased standard error estimates or both. They also showed that they could result in problematic underestimates of standard errors and p-values. This is because using a single value to replace missing values does not account for the variation that would most probably have been present in the data set had the variables been observed.

However, all was not lost, as this dissertation showed that maximum likelihood and multiple imputation methods displayed far better results when dealing with missing values. These methods demon-

strate almost perfect statistical properties and at the same time possess these properties under assumptions that were usually weaker than those used in more conventional methods. For example, although the standard software used for maximum likelihood and multiple imputation assumes that the data is MAR, these techniques still performed well and produced reasonable standard errors even though the data was not MAR. Unlike the single imputation techniques, multiple imputation allowed plausible values to be used to replace missing values in such a way that resulted in unbiased parameter estimates. Also, the uncertainty of parameter estimation was estimated in a reasonable manner.

Since missing data software is now widely available and because maximum likelihood and multiple imputation can be easily implemented, this dissertation strongly suggests using these modern approaches to handle missing data and not the traditional/conventional techniques. Maximum likelihood and multiple imputation both have their advantages and disadvantages and hence deciding on which of these two methods to use depends on the individual's preference.

Possible future scope may be to test the performance of each model that was fitted to the imputed data by making use of training and holdout samples, where the root mean square error may be used as a measure to evaluate this performance. Also, instead of only looking at the EM algorithm and applying the bootstrapping technique to yield accurate standard errors, an investigation into the direct maximum likelihood method can be done.

References

Acuña, E. & Rodriguez, C. (2004). *Classification, Clustering, and Data Mining Applications*, chapter The Treatment of Missing Values and its Effect on Classifier Accuracy, (pp. 639–647). Springer Berlin Heidelberg.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Allison, P. D. (2001). *Missing Data*. Thousand Oaks, CA: Sage.

Allison, P. D. (2012). Handling Missing Data by Maximum Likelihood. Keynote presentation at the SAS Global Forum, April 23, 2012, Orlando, Florida.

Banerjee, A., Chitnis, U. B., Jadhav, S. L., Bhawalkar, J. S., & Chaudhury, S. (2009). Hypothesis Testing, Type I and Type II Errors. *Industrial Psychiatry Journal*, 18(2), 127–131.

Baneshi, M. & Talei, A. (2012). Does the Missing Data Imputation Method Affect the Composition and Performance of Prognostic Models? *Iranian Red Crescent Medical Journal*, 14(1), 31–36.

Baraldi, A. N. & Enders, C. K. (2010). An Introduction to Modern Missing Data Analyses. *Journal of School Psychology*, 48(1), 5–37.

Batista, G. E. A. P. A. & Monard, M. C. (2003). An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Applied Artificial Intelligence*, 17(5-6), 519–533.

Berglund, P. & Heeringa, S. (2014). *Multiple Imputation of Missing Data Using SAS®*. SAS Institute Inc., Cary, North Carolina, USA.

Chan, Y. H. (2005). Multinomial Logistic Regression. *Singapore Medical Journal*, 46(6), 259–268.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods*, 6(4), 330–351.

Dempster, A. P., Laird, N., & Rubin, D. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.

Dempster, A. P. & Rubin, D. B. (1983). Incomplete Data in Sample Surveys. *Theory and Bibliographies*, 2, 3–10.

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: A Gentle Introduction to Imputation of Missing Values. *Journal of Clinical Epidemiology*, 59(10), 1087–1091.

Enders, C. K. (2001). A Primer on Maximum Likelihood Algorithms Available for Use with Missing Data. *Structural Equation Modelling*, 8(1), 128–141.

Enders, C. K. (2010). *Applied Missing Data*. New York: Guilford Press.

Farber, O. & Kadmon, R. (2002). Assessment of Alternative Approaches for Bioclimatic Modelling with Special Emphasis on the Mahalanobis Distance. *Ecological Modelling*, 160(2003), 115–130.

Graham, J. W. (2009). Missing Data Analysis: Making it Work in the Real World. *Annual Review of Psychology*, 60, 549–576.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science*, 8(3), 206–213.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. MIT Press Cambridge, MA, USA.

He, Y. (2010). Missing Data Analysis using Multiple Imputation. Getting to the Heart of the Matter. *Circulation: Cardiovascular Quality and Outcomes*, 3(1), 98–105.

Horton, N. J. & Lipsitz, S. R. (2001). Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables. *The American Statistician*, 55(3), 244–254.

Inc., S. I. (2008a). *SAS/STAT® 9.2 Users Guide*, chapter The DISCRIM Procedure, (pp. 1384–1481). Cary, NC: SAS Institute Inc.

Inc., S. I. (2008b). *SAS/STAT® 9.2 Users Guide*, chapter The REG procedure, (pp. 76–113). Cary, NC: SAS Institute Inc.

Inc., S. I. (2008c). *SAS/STAT® 9.2 Users Guide*, chapter The MIANALYZE Procedure, (pp. 3834–3884). Cary, NC: SAS Institute Inc.

Inc., S. I. (2013). *SAS/STAT® 9.2 Users Guide*, chapter The MI Procedure, (pp. 3738–3831). Cary, NC: SAS Institute Inc.

Jönsson, P. & Wohlin, C. (2004). An Evaluation of k-Nearest-Neighbour Imputation Using Likert Data. In *Proceedings of the 10th International Symposium on Software Metrics* (pp. 108–118).

Khattree, R. & Naik, D. N. (2000). *Multivariate Data Reduction and Discrimination with SAS Software*. SAS Institute Inc., Cary, NC, USA.

Little, R. J. A. (1992). Regression with Missing X's: A Review. *Journal of the American Statistical Association*, 87(420), 1227–1237.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley and Sons, Inc., 2nd edition.

Luengo, J., García, S., & Herrera, F. (2012). On the Choice of the Best Imputation Methods for Missing Values Considering Three Groups of Classification Methods. *Knowledge and Information Systems*, 32(1), 77–108.

Massell, P. B. (2000). Latent Variable Models for Analysis of Survey Data. In *Proceedings of the Survey Research Methods Section Alexandria, VA: American Statistical Association* (pp. 203–208).

Munguía, T. & Armando, J. (2014). Comparison of Imputation Methods for Handling Missing Categorical Data with Univariate Pattern. *Journal of Quantitative Methods for Economics and Business Administration*, 17(1), 101–120.

Myung, I. J. (2003). Tutorial on Maximum Likelihood Estimation. *Journal of Mathematical Psychology*, 47(1), 90–100.

Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London. Series A*, 236(767), 333–380.

Neyman, J. & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, 231, 289–337.

Osborne, J. W. (2013). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. Sage Publications, Inc.

Pal, N., Jin, C., & Lim, W. K. (2006). *Handbook of Exponential and Related Distributions for Engineers and Scientists*. Chapman & Hall/CRC, Boca Raton, FL.

Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. H. (2006). *Real Data Analysis*, chapter Advances in Missing Data Methods and Implications for Educational Research, (pp. 31–78). Greenwich, CT: Information Age Publishing, Inc.

Pigott, T. D. (2001). A Review of Methods for Missing Data. *Educational Research and Evaluation*, 7(4), 353–383.

Raghunathan, T. E. (2004). What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. *Annual Review of Public Health*, 25(1), 99–117.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3), 581–592.

Rubin, D. B. (1978). Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse. In *Proceedings of the Survey Research Methods Section Alexandria, VA: American Statistical Association* (pp. 20–34).

Russell, D., Stern, H. S., & Sinharay, S. (2001). The Use of Multiple Imputation for the Analysis of Missing Data. *Psychological Methods*, 6(4), 317–329.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall/CRC.

Schafer, J. L. & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2), 147–177.

Schafer, J. L. & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing Data Problems: A Data Analyst's Perspective. *Multivariate Behavioural Research*, 33(4), 545–571.

Schluchter, M. D. (2014). *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd.

Searle, S. R. (1971). *Linear Models*. John Wiley & Sons, Inc.

Soley-Bori, M. (2013). *Dealing with Missing Data: Key Assumptions and Methods for Applied Analysis*. Technical report, Boston University.

Truxillo, C. (2005). Maximum Likelihood Parameter Estimation with Incomplete Data. In *Proceedings of the Thirtieth Annual SAS Users Group International Conference, Philadelphia, PA.* (pp. 1–19).

Walli, G. M. (2010). Bayesian Variable Selection in Normal Regression Models. Master's thesis, Johannes Kepler University Linz.

Xie, L. (2012). KNN Classification and Regression using SAS®.

Zhu, X. (2014). Comparison of Four Methods for Handling Missing Data in a Longitudinal Data Analysis through a Simulation Study. *Open Journal of Statistics*, 4(11), 933–944.

Appendix A - SAS code

A.1 Importing the data

```
/*Importing the NLSY missing data set*/
data nlsy_miss;
  input anti self pov black hispanic childage divorce gender momage momwork;
  datalines;
1 21 1 0 0 8 0 1 21 0
0 20 . . . 8.4166666667 0 1 22 1
5 21 0 . . 8.0833333333 1 0 18 0
2 23 0 0 0 8.25 0 0 24 0
1 22 0 0 0 9.3333333333 0 1 22 0
1 . 0 0 0 8.5833333333 0 0 24 0
3 24 0 0 0 9.25 1 1 23 .
4 19 0 . . 8.5 1 0 18 0
1 21 . . . 8.0833333333 0 0 24 0
4 9 0 0 0 9.1666666667 1 0 20 .
3 20 1 . . 8.8333333333 1 1 23 1
3 15 . 0 1 9.1666666667 1 1 20 0
3 . . 0 0 8.5833333333 0 0 19 .
1 . 0 0 0 8.75 0 0 22 1
3 21 0 . . 9.8333333333 0 1 23 .
2 16 . 1 0 9 0 1 20 0
1 18 . 1 0 8.6666666667 0 0 19 .
0 . 0 . . 9.3333333333 0 0 22 1
3 19 1 1 0 8.1666666667 0 1 18 0
5 13 0 0 0 8.4166666667 0 0 20 1
4 . 0 0 0 9.3333333333 0 0 23 0
1 21 1 0 0 8.5833333333 1 0 24 .
```

```
0 20 . . . 8 0 1 24 1
2 . 0 0 0 9.5 0 0 23 0
0 23 0 0 0 9.5 0 1 22 1
5 18 0 0 0 8.75 0 1 22 0
2 22 0 0 0 9.1666666667 0 0 17 1
2 22 0 0 0 10 0 1 21 0
1 18 0 0 1 8.25 0 0 24 1
0 20 . 0 0 9.75 0 1 20 1
1 . . 0 0 8 0 1 25 1
4 24 1 0 0 8.9166666667 1 1 19 0
2 . 0 0 0 9.75 1 0 18 0
3 . 0 0 0 9.8333333333 0 0 19 0
3 19 0 0 0 8.4166666667 0 1 21 0
0 20 0 0 1 9.25 0 1 19 0
0 19 . 1 0 8.1666666667 1 1 22 0
4 21 1 0 0 9.75 1 0 19 1
2 21 0 0 0 8.6666666667 0 1 22 0
0 20 1 0 0 9 1 1 21 .
1 24 0 0 0 9.4166666667 0 1 18 1
1 22 1 0 0 8.5833333333 1 1 19 0
2 24 . . . 9.3333333333 0 0 22 0
0 24 . 0 0 8.6666666667 0 1 25 1
0 19 0 0 0 9.25 0 0 22 0
1 24 0 0 0 8.5 0 0 24 1
1 24 . 1 0 8.5 0 1 23 1
3 19 . 0 0 9.9166666667 0 1 22 .
3 24 1 . . 8.1666666667 1 0 24 0
1 . 0 . . 9.0833333333 1 0 22 0
0 21 . 1 0 8.0833333333 1 1 23 0
1 . 0 1 0 10 0 0 19 .
2 24 1 0 0 9.6666666667 1 1 16 1
0 21 . 1 0 8.8333333333 0 0 23 1
5 20 1 1 0 8.9166666667 0 0 20 0
0 19 . 1 0 8.5833333333 0 1 17 0
2 16 1 1 0 8.9166666667 0 1 21 .
5 15 . 1 0 9.1666666667 0 0 21 0
```

```
3 20 . . . 8.25 0 0 24 0
0 . . 0 0 8.0833333333 0 1 24 1
2 13 . 0 0 8.8333333333 0 1 23 1
3 20 1 . . 8.0833333333 1 1 18 0
0 . . 0 0 8.9166666667 0 0 18 0
1 22 1 1 0 9.6666666667 0 1 23 1
2 24 1 1 0 8.6666666667 0 1 20 0
1 . 0 0 0 9.75 0 1 22 1
1 . . 0 0 8.5833333333 0 1 23 .
0 . . . . 8.75 0 1 21 0
4 19 0 0 0 8.4166666667 0 0 19 0
2 20 1 0 1 9.4166666667 1 1 20 .
0 . 1 0 0 9.4166666667 1 0 19 1
0 17 0 0 1 9.9166666667 0 1 19 0
2 . 0 . . 8.8333333333 0 1 23 0
0 24 0 . . 8.75 0 1 18 0
0 21 0 . . 9.9166666667 0 0 16 .
0 21 0 0 0 9.0833333333 0 0 17 1
1 23 0 0 0 9.9166666667 0 1 20 .
0 22 0 0 0 8.8333333333 0 1 23 0
2 18 . 1 0 8.9166666667 1 0 20 0
3 15 0 0 0 8.5 0 1 17 1
0 . . 0 1 8.8333333333 0 0 22 0
0 22 . 0 0 8.6666666667 0 1 23 0
4 19 0 0 0 9.25 0 0 20 .
3 20 1 0 0 9.25 0 1 21 1
3 . 0 . . 8.9166666667 0 1 24 0
0 24 . 0 0 9.5833333333 1 1 23 1
0 . 0 . . 9.9166666667 0 0 23 .
4 24 0 0 0 8.1666666667 0 0 24 .
2 . 1 0 0 8.4166666667 1 1 22 1
1 17 . 0 1 8.25 0 1 17 0
4 14 0 . . 8.9166666667 1 0 19 .
0 21 . 1 0 9.5 0 1 20 0
2 . 0 . . 8.5 0 0 21 0
2 24 0 0 0 8.0833333333 0 0 23 .
```

```
2 . 0 0 0 10 0 1 18 0
2 19 0 0 0 9.9166666667 1 1 17 1
0 . 0 0 0 8.5833333333 1 1 18 1
0 21 . 0 0 9.25 0 1 20 0
0 . 1 1 0 9.6666666667 1 0 19 1
3 20 0 0 1 8.25 0 1 20 0
2 20 . . . 9 0 1 24 .
0 . . 0 0 8.9166666667 0 0 22 0
2 21 0 0 0 9.8333333333 0 1 23 0
1 . 0 0 0 8.1666666667 0 1 24 0
1 23 0 . . 9.0833333333 0 1 20 0
0 23 0 0 0 9.1666666667 0 0 24 0
0 . 0 0 0 9.75 0 0 19 0
0 21 . 0 0 8 0 1 21 0
0 24 1 0 0 8.6666666667 1 1 20 .
0 24 0 0 0 8.25 0 0 21 .
3 10 1 0 0 8.0833333333 1 1 24 0
0 24 0 0 0 9.0833333333 0 0 19 0
2 21 0 0 0 8.4166666667 0 1 20 .
0 21 0 0 0 9 0 0 24 1
0 . 0 0 0 8.25 0 1 25 0
3 22 . 0 0 8.9166666667 0 1 23 .
1 . 0 0 0 8.1666666667 0 0 20 0
1 24 0 0 0 9.9166666667 0 0 17 0
0 23 0 . . 8.1666666667 0 1 22 1
1 16 0 0 0 9.5 0 1 23 0
1 20 1 . . 9.8333333333 1 0 18 0
1 21 0 . . 9 0 1 20 0
0 22 . 1 0 8 0 0 20 0
4 17 1 1 0 9.0833333333 0 0 18 1
2 22 0 0 1 9.1666666667 1 0 17 0
1 22 0 1 0 10 0 1 19 0
3 19 1 1 0 9.1666666667 0 0 21 1
4 22 1 0 0 9.3333333333 0 0 23 1
1 21 0 0 0 8.4166666667 0 1 23 1
6 20 1 1 0 9.8333333333 0 0 16 1
```

```
1 . 0 0 0 8.75 0 0 22 0
1 24 0 0 0 8.0833333333 0 0 19 1
1 24 0 . . 9.6666666667 0 1 22 0
2 16 . 0 0 9.3333333333 1 0 20 0
0 . . 0 0 9 1 0 23 .
0 18 0 0 0 8.6666666667 1 1 17 0
2 . 0 0 0 9 0 0 24 0
5 18 1 0 0 8.8333333333 1 0 17 1
1 20 0 0 0 9.1666666667 0 0 18 .
0 20 . . . 8.8333333333 0 1 18 0
2 22 0 . . 9.75 0 1 20 0
1 22 0 0 1 9.25 1 0 18 0
0 19 . 0 1 8.75 1 0 19 0
4 19 1 0 1 9.8333333333 0 1 17 1
0 . . 0 1 8.8333333333 0 1 22 0
5 22 1 0 1 9 0 0 20 1
3 19 0 0 1 9.75 0 0 18 0
1 21 . . . 9.5833333333 0 0 20 0
2 . 0 . . 9.4166666667 0 0 19 .
0 . 0 0 0 10 1 1 16 .
1 14 0 0 0 9.8333333333 0 0 19 0
0 23 0 0 0 8.3333333333 0 0 22 0
1 . 0 . . 8.5833333333 0 1 24 0
4 20 0 . . 9.5833333333 0 0 20 .
1 21 0 0 0 8 0 1 23 1
1 18 . 1 0 8.1666666667 1 0 24 0
0 18 . . . 9.5 1 1 19 0
0 . 0 . . 8.3333333333 0 0 23 0
4 24 1 0 0 9.25 0 1 20 0
0 20 0 0 0 9.8333333333 0 0 22 0
2 . 1 0 1 9.3333333333 0 1 22 .
0 . 0 . . 8.8333333333 0 0 24 0
0 . 1 0 1 9.6666666667 1 1 23 1
1 17 1 0 1 8.4166666667 1 0 24 1
1 22 0 0 0 9.8333333333 0 0 18 0
1 . 0 0 0 8.75 0 0 23 0
```

```
3 23 1 1 0 9.75 0 0 23 0
0 23 1 1 0 8.0833333333 0 1 17 .
2 21 . . . 8.1666666667 0 0 19 .
1 21 0 0 0 8.1666666667 1 1 19 0
4 22 0 0 0 8.4166666667 0 1 23 1
0 20 . 1 0 9.6666666667 1 0 17 0
0 . 0 1 0 8.1666666667 1 0 18 0
2 24 . . . 8 0 1 24 0
0 . 0 0 0 8.6666666667 0 1 25 0
1 19 0 0 0 9.9166666667 0 0 22 0
3 21 0 0 0 9.25 0 1 19 0
1 20 0 . . 8.8333333333 0 0 22 0
1 19 . 0 0 9.5833333333 0 1 17 0
0 . 0 0 0 9.0833333333 0 1 23 .
2 . 0 0 0 8.0833333333 0 0 22 0
4 22 1 0 0 9.9166666667 0 0 21 .
4 22 0 0 0 9.6666666667 0 0 20 1
1 . 0 0 0 9.75 0 1 22 0
0 21 0 . . 9.1666666667 0 1 21 0
3 19 1 0 0 9.5833333333 0 0 20 1
0 . 0 0 0 9.4166666667 0 0 19 0
3 . 1 1 0 8.4166666667 0 0 21 0
0 . 1 0 1 9.5833333333 1 0 21 0
0 16 1 0 1 8.6666666667 1 1 22 0
2 19 . 0 0 9 0 0 21 0
0 . 0 0 0 8.9166666667 0 0 22 0
4 18 . 0 0 9.3333333333 0 0 23 0
0 . 0 0 0 8.4166666667 0 1 20 0
0 . . 0 0 9.9166666667 0 0 23 1
3 18 1 1 0 8.5833333333 0 1 23 0
1 18 1 1 0 9 1 0 20 0
0 . . 0 0 9.5 0 0 19 0
3 17 0 0 0 8.5 0 1 22 .
4 17 1 0 0 10 1 1 20 1
2 . 0 1 0 9 0 0 23 1
4 15 0 1 0 8.4166666667 0 0 22 1
```

```
6 20 . 0 0 8.5 1 0 23 0
2 21 0 1 0 8.8333333333 1 0 22 0
1 24 0 0 0 9.0833333333 1 0 19 0
0 . . 0 0 9.6666666667 0 0 23 1
5 22 . . . 8.8333333333 0 0 18 0
0 . 0 0 0 8.9166666667 0 0 23 1
0 19 0 . . 8.1666666667 0 0 23 0
0 . 0 1 0 9.3333333333 0 1 19 0
2 17 1 1 0 9.75 0 0 23 1
3 20 1 1 0 9.9166666667 0 1 17 .
2 24 . 0 0 9.4166666667 0 1 19 0
1 17 . 0 0 9.75 0 1 23 0
0 . 0 0 0 8.5 0 1 21 .
0 . 0 0 0 8.1666666667 1 0 23 .
1 . . . . 9 0 1 21 0
0 . 1 0 1 9.0833333333 1 0 18 0
3 24 0 . . 9.0833333333 0 0 23 1
3 24 . 1 0 9.9166666667 0 0 19 1
1 17 0 1 0 8.1666666667 0 0 21 1
0 22 0 0 0 8.5 0 1 24 0
2 18 0 0 0 9.1666666667 0 0 20 1
3 . 0 0 0 8.0833333333 0 0 21 .
0 23 0 0 0 9.8333333333 0 1 22 0
0 . 0 0 0 8.4166666667 0 0 24 .
1 20 1 . . 9.0833333333 1 0 22 .
1 22 . 0 1 9.3333333333 1 0 23 1
1 24 0 0 0 9.6666666667 0 0 19 .
2 21 . 0 0 8.4166666667 0 0 20 .
0 20 . 0 0 8.5 0 1 22 1
1 . . 0 0 8.5 0 1 22 1
5 19 1 1 0 9.8333333333 1 0 20 0
4 19 1 0 1 9.1666666667 1 0 20 0
2 20 0 0 1 10 0 0 17 0
2 . 0 0 0 9.5833333333 0 0 21 0
4 20 . 0 0 8.25 1 1 21 0
2 24 0 0 0 9.8333333333 0 0 21 0
```

```
2 21 0 0 0 9.0833333333 0 1 21 0
1 . 0 1 0 8.8333333333 0 0 21 0
3 . 0 0 0 10 0 0 20 0
2 22 0 0 0 8.4166666667 0 1 22 0
6 14 1 0 0 8 1 1 20 1
0 21 0 0 0 9.1666666667 0 0 23 0
0 . . 0 0 8.75 1 1 20 0
1 21 1 . . 8.3333333333 0 1 19 0
0 21 1 . . 9.6666666667 1 1 17 0
1 18 0 0 0 8 0 1 21 .
0 24 0 0 0 8.1666666667 0 0 19 0
0 . 0 . . 8.4166666667 0 1 20 1
3 18 0 0 0 9.3333333333 0 1 20 1
2 22 0 0 0 8.5833333333 1 1 23 0
3 20 0 1 0 9.75 0 1 21 0
0 . . 0 0 8.4166666667 0 1 21 0
0 23 0 0 0 8.8333333333 0 0 24 0
1 18 0 . . 8.0833333333 0 1 22 0
2 22 0 0 0 8.1666666667 0 1 25 1
0 . . 0 0 9.1666666667 0 0 17 .
0 . 1 0 0 8 0 1 18 0
0 24 0 0 0 9.3333333333 0 1 22 .
3 15 . 0 0 8.4166666667 1 1 20 .
1 24 0 0 0 9.75 0 0 23 0
3 21 . 0 0 8.9166666667 1 0 23 0
2 . 0 0 0 8.0833333333 0 0 19 0
0 . . 0 0 8.0833333333 0 1 25 .
0 20 . 0 0 9.6666666667 0 1 22 1
1 15 0 0 1 9.25 0 0 19 0
4 . 0 0 1 8.0833333333 0 0 20 0
0 21 1 0 0 9.75 0 1 18 1
1 24 1 0 0 9.8333333333 0 0 20 0
1 22 0 0 0 9.75 0 1 18 .
0 . 0 0 0 9.75 0 1 22 1
2 20 0 0 0 9.5 0 1 23 1
6 24 1 0 0 8.6666666667 0 0 23 .
```

```
2 23 0 0 0 9 0 0 22 0
4 22 0 0 0 8.1666666667 0 0 20 0
1 . 0 0 0 9.25 0 0 20 1
1 20 0 0 0 9.9166666667 0 0 23 0
3 24 . 0 0 10 0 0 20 0
2 16 0 0 0 8.5 0 0 17 0
3 24 1 0 0 9.5 0 0 19 1
0 . 0 0 0 9.3333333333 0 1 19 .
2 24 0 . . 8.25 0 1 22 0
3 19 0 1 0 8 0 0 20 0
1 . 0 . . 9.6666666667 0 0 23 0
1 . . . . 9.0833333333 0 1 24 0
0 21 . 0 0 9 0 1 23 0
0 21 0 0 0 8 0 0 24 .
3 20 0 0 0 8.5833333333 0 1 22 1
2 17 0 1 0 8.1666666667 0 1 17 0
0 14 0 0 0 8.5833333333 0 0 19 0
0 23 . 0 0 8.5833333333 1 1 19 0
1 14 . . . 9.5833333333 1 1 21 0
2 . . 0 0 9.25 0 1 21 0
2 24 . 0 0 9 0 0 18 1
1 13 0 0 1 8.8333333333 0 0 21 0
1 . . . . 9.5833333333 0 1 19 0
1 . . 0 1 8.1666666667 0 1 19 0
1 16 1 . . 9.9166666667 0 0 19 0
4 12 1 0 1 8 0 0 21 0
0 21 0 . . 8.75 0 0 23 0
3 12 0 0 1 9 1 1 22 0
0 . 0 0 1 8.4166666667 0 1 22 1
1 24 0 0 1 8.5 0 0 24 0
4 22 . 0 1 8.25 0 0 19 1
2 14 0 0 1 8.1666666667 0 0 19 1
0 24 . 0 1 8.8333333333 0 1 19 0
2 24 . 1 0 8 0 0 20 0
2 22 . 1 0 9.3333333333 0 1 22 1
5 24 1 1 0 10 0 0 18 0
```

```
0 21 0 1 0 8.6666666667 0 1 19 0
2 17 1 1 0 9.25 0 0 19 0
1 24 1 1 0 9.25 0 1 23 1
1 20 1 1 0 8.25 0 1 20 1
1 . 1 . . 9.9166666667 1 0 17 1
3 21 1 1 0 9 0 0 20 1
1 21 0 0 1 9.8333333333 0 0 21 0
1 22 0 0 1 8 0 1 23 0
1 21 1 . . 9.5833333333 0 0 19 .
2 21 1 1 0 8.0833333333 0 0 21 0
1 18 1 1 0 8.3333333333 0 0 19 0
1 15 1 1 0 9.0833333333 1 1 20 1
1 . 0 . . 8.5833333333 0 1 21 1
0 22 . . . 9.3333333333 0 0 23 0
0 18 0 1 0 8 0 1 18 1
3 19 1 1 0 8.8333333333 1 0 23 0
2 14 . . . 8.6666666667 1 1 21 0
5 15 1 1 0 9.1666666667 0 0 19 0
2 22 1 1 0 9.8333333333 1 1 23 0
0 12 1 1 0 8.8333333333 1 0 24 0
1 . 0 1 0 9.8333333333 1 0 22 0
0 19 . 1 0 8 0 0 21 0
3 18 1 1 0 8.1666666667 1 1 21 0
4 18 0 1 0 9 0 1 23 .
1 . . 1 0 9.1666666667 0 1 20 0
0 21 1 1 0 9.3333333333 1 1 17 .
1 16 1 1 0 9.8333333333 0 1 19 0
5 17 0 1 0 9.25 0 1 23 0
2 22 0 1 0 8.3333333333 0 0 24 0
2 . . 1 0 10 1 0 18 .
2 21 1 1 0 9.75 0 1 19 0
1 21 . 1 0 8.25 0 1 22 0
1 18 1 1 0 9.1666666667 1 0 20 0
0 . 0 . . 9.6666666667 0 1 18 0
3 18 1 1 0 8.75 1 0 19 0
3 17 1 1 0 9 0 1 17 1
```

```
3 24 1 1 0 8.1666666667 0 1 22 1
3 21 . 1 0 9.6666666667 0 0 22 1
3 20 1 . . 8.5833333333 0 0 22 1
1 . 1 . . 8.0833333333 1 0 19 0
0 20 1 1 0 8.3333333333 0 1 21 1
1 24 . 1 0 9.8333333333 0 0 16 0
2 17 . 1 0 8.75 0 0 18 0
2 . 0 1 0 8.0833333333 0 0 23 0
2 . . 1 0 9.0833333333 0 0 20 1
1 . 0 1 0 8.5833333333 0 1 18 .
5 20 . 1 0 9.6666666667 0 0 18 .
4 15 1 1 0 8.75 1 1 22 1
0 23 1 1 0 8.5 1 1 23 0
3 14 0 1 0 9.1666666667 0 0 17 .
1 18 1 0 1 8.8333333333 1 1 23 .
2 18 1 1 0 8.9166666667 1 1 22 1
2 . 0 0 1 8 1 0 19 1
0 19 . 1 0 9.5 0 0 20 0
1 . 1 0 1 9.1666666667 1 1 21 0
0 12 0 0 1 8.25 0 1 24 0
0 . . 1 0 9.0833333333 1 1 22 0
2 23 1 1 0 8.3333333333 0 0 18 1
1 21 1 1 0 9.4166666667 0 1 20 1
1 . 1 0 1 8.75 0 1 21 1
1 . 0 1 0 8.6666666667 0 1 23 .
2 17 . . . 9.75 0 1 20 0
5 . 0 1 0 8.3333333333 0 0 21 0
2 . . . . 8.25 0 1 24 0
1 21 0 1 0 8.9166666667 1 0 21 0
0 17 . . . 8.9166666667 1 0 23 .
0 23 . 1 0 8.9166666667 0 1 22 0
2 21 1 . . 9.0833333333 0 0 21 0
0 22 . 1 0 9.1666666667 0 1 19 0
1 . 1 . . 9.9166666667 1 1 17 1
4 18 0 1 0 8.25 1 0 20 0
1 18 1 1 0 8.6666666667 0 1 19 1
```

```
5 20 1 1 0 9.3333333333 1 1 21 1
5 23 1 . . 8 1 1 22 1
4 . 1 1 0 9.5 1 0 23 0
4 20 1 . . 8.1666666667 0 0 19 1
0 . 1 1 0 9.25 0 0 18 0
1 16 1 1 0 8.3333333333 0 1 17 .
5 20 0 1 0 9.0833333333 0 0 22 0
2 22 1 . . 9.3333333333 1 1 22 1
1 . . 1 0 9.1666666667 0 0 20 0
0 24 0 1 0 9.8333333333 0 1 20 1
3 22 0 1 0 8.75 0 0 21 .
0 12 0 1 0 8.5833333333 0 1 18 0
0 . 0 1 0 8.25 0 0 19 1
2 16 . 1 0 9 0 1 18 0
4 24 1 1 0 9.25 0 0 21 1
0 22 1 1 0 9.4166666667 0 1 18 0
2 16 0 1 0 8.25 0 0 18 .
2 23 . 1 0 9.4166666667 1 1 22 0
1 24 0 1 0 8 0 1 20 0
2 22 1 1 0 8.9166666667 0 1 17 0
0 24 1 1 0 8.1666666667 0 1 22 0
0 23 0 . . 8 0 1 21 0
0 . 0 . . 9.25 1 1 24 .
3 17 . 1 0 8.5833333333 1 0 24 0
4 22 0 1 0 8.6666666667 0 1 24 .
0 . . 1 0 8.8333333333 0 1 19 1
4 22 1 . . 9.5833333333 1 0 19 0
0 23 0 1 0 9.5833333333 0 0 16 0
4 20 1 0 1 8.3333333333 1 0 19 1
2 19 0 1 0 10 1 1 22 .
1 19 0 . . 8.0833333333 0 1 19 0
3 20 1 0 1 8.9166666667 0 0 19 1
0 . 0 0 1 8.5833333333 0 1 20 0
0 15 0 0 1 9.4166666667 0 0 22 0
1 22 0 1 0 8 0 1 19 0
3 . 1 1 0 9.9166666667 0 0 17 1
```

```
1 21 1 1 0 8.8333333333 0 0 18 1
0 21 1 0 1 8.3333333333 0 1 18 1
0 15 0 0 1 9.5 0 0 20 1
2 24 0 1 0 9.3333333333 0 1 23 1
2 20 0 0 1 8.25 0 0 22 0
1 . 0 0 1 8.3333333333 0 0 19 0
5 24 . 1 0 9.5833333333 0 0 23 1
2 21 1 0 1 9.3333333333 1 1 23 1
0 18 1 0 1 8.75 0 1 19 1
0 19 0 1 0 9.8333333333 0 0 19 0
3 . 1 . . 8.0833333333 0 1 21 1
2 21 . . . 8.25 1 1 22 1
0 24 1 0 1 9.3333333333 0 1 18 1
0 21 . 0 1 9.1666666667 0 1 18 0
4 15 1 . . 9.9166666667 0 0 20 1
0 . . 0 1 8.75 0 0 22 0
3 19 0 0 1 9.9166666667 0 0 20 1
2 23 0 0 1 8.8333333333 0 0 21 1
1 17 1 0 1 8.0833333333 1 1 19 0
0 . 0 0 1 9.0833333333 0 1 22 .
1 . 0 0 1 8.25 0 0 24 1
1 18 0 . . 8.25 0 1 24 1
1 21 . . . 8.0833333333 0 1 20 1
3 12 0 0 1 8.6666666667 0 1 24 0
1 22 . 0 1 8.25 0 1 19 1
2 19 . . . 8.9166666667 0 0 21 1
0 . . 0 1 9.1666666667 0 1 20 0
0 . . . . 9.9166666667 0 1 19 0
0 . . 0 1 8.5833333333 0 1 23 0
0 . 1 . . 8.5833333333 0 1 24 0
1 19 . 0 1 8.3333333333 0 0 17 1
3 21 0 0 1 9.6666666667 0 1 17 0
3 19 0 0 1 8.3333333333 0 1 18 0
3 17 1 0 1 9.1666666667 0 0 24 1
4 18 1 . . 9.5833333333 0 0 20 1
1 22 0 0 1 10 1 0 22 0
```

```
1 19 . . . 8.6666666667 0 1 19 0
3 18 0 1 0 8.5833333333 0 1 21 0
1 17 0 . . 9.5833333333 0 0 19 0
0 21 . . . 8.25 0 1 21 0
2 20 0 1 0 8.1666666667 1 0 22 0
0 . 0 . . 8.5833333333 0 1 21 0
4 19 0 1 0 8.4166666667 0 1 23 0
2 23 0 1 0 8.6666666667 0 1 19 0
4 . 0 1 0 9.75 0 0 22 0
2 22 1 . . 8.75 0 0 18 0
3 24 0 1 0 9.6666666667 0 0 21 0
1 . 0 1 0 8.5833333333 0 0 22 0
1 15 . 1 0 10 1 1 16 0
3 13 1 1 0 9.5 0 1 17 1
0 22 0 0 1 8.8333333333 1 0 17 1
1 22 0 0 1 8.4166666667 0 0 25 0
2 13 1 1 0 8 0 1 17 0
1 19 1 1 0 9.25 0 1 19 0
0 24 . . . 9.0833333333 0 1 24 0
0 . 1 0 1 9.25 1 1 20 1
0 . 0 0 1 8.9166666667 0 0 23 1
0 . 0 . . 8 0 0 20 0
2 21 0 0 1 8.5833333333 0 1 23 0
2 24 . . . 9 1 0 19 0
2 15 0 . . 8.8333333333 0 1 22 0
2 20 0 0 1 9.5 0 0 19 .
0 . . 0 1 10 1 0 19 0
0 23 0 0 1 8.9166666667 0 1 22 0
0 18 0 0 1 9.4166666667 0 1 16 1
2 17 . 1 0 9.75 0 1 21 0
2 24 1 . . 9.4166666667 0 1 18 0
0 22 . 1 0 8.9166666667 0 1 18 0
3 24 1 . . 9.4166666667 0 0 21 0
2 20 . 1 0 9.0833333333 0 1 19 0
6 21 1 1 0 9.5833333333 0 1 21 0
0 17 . 1 0 9.3333333333 0 1 20 1
```

```
1 11 1 . . 9.75 0 0 18 0
4 24 1 . . 8.9166666667 0 0 19 1
1 20 1 0 1 8.25 0 1 22 .
2 21 1 0 1 8.8333333333 1 0 21 1
0 22 0 0 1 9.6666666667 1 0 20 0
0 . 0 0 1 8.1666666667 1 0 21 0
1 . 0 0 1 8.0833333333 0 0 19 0
1 22 1 0 1 9 0 0 18 1
1 24 0 1 0 9.1666666667 0 1 23 0
3 24 0 1 0 8.9166666667 0 0 21 0
1 21 0 1 0 8.5833333333 1 1 24 0
0 . . 1 0 9.75 1 1 21 .
0 14 0 1 0 8.25 1 0 23 0
0 22 0 0 1 8.5833333333 0 0 23 0
2 . 0 1 0 9.75 0 0 20 0
4 22 . 1 0 8.5833333333 0 1 25 0
0 24 . 0 1 8.0833333333 0 0 25 0
1 24 0 0 1 9.9166666667 0 1 16 1
1 23 0 0 1 8.4166666667 0 0 18 1
1 19 1 1 0 8.0833333333 0 1 21 1
1 21 0 . . 8.5 0 1 22 0
0 . 0 0 1 8.5 0 1 22 1
0 21 . . . 8.75 0 1 17 .
0 . 0 1 0 10 1 1 18 0
3 19 1 1 0 8.3333333333 1 1 18 0
6 17 . 0 1 9 0 0 21 1
2 22 0 . . 9.4166666667 0 0 21 0
3 18 0 . . 8.6666666667 0 1 19 1
0 . 0 . . 9.75 0 1 20 0
1 21 0 0 1 9.9166666667 0 1 23 0
2 . 1 0 1 9.5 0 0 18 1
1 . 0 0 1 8.75 0 1 23 .
2 20 . 0 1 8.4166666667 0 0 23 0
1 . . 0 1 8.6666666667 0 1 19 .
4 20 0 0 1 8.6666666667 0 1 19 0
3 . 0 0 1 9.5833333333 0 1 23 0
```

```
1 . 0 0 1 8.9166666667 0 0 24 0
3 20 0 0 1 8.3333333333 1 0 23 0
1 20 1 1 0 9.5 0 1 21 .
1 19 1 0 1 10 0 1 16 0
0 . 0 0 1 9.1666666667 0 0 20 0
1 23 0 . . 8.9166666667 1 1 23 1
1 21 0 0 1 9.0833333333 1 0 21 0
2 18 1 . . 9.25 1 0 17 1
2 16 1 0 1 9 0 0 24 .
3 17 . 0 1 8 0 1 25 1
1 22 1 0 1 9.5833333333 0 0 20 1
0 24 . 0 1 8.75 0 0 24 0
1 19 0 1 0 8 0 1 20 0
2 17 0 1 0 8.8333333333 0 1 23 0
0 23 . 1 0 9.5833333333 0 0 20 1
1 16 0 . . 9.8333333333 0 1 19 .
1 . 0 1 0 8.25 1 1 20 0
1 20 0 1 0 9.25 1 1 18 0
3 24 0 1 0 9.5833333333 0 0 18 .
0 16 0 1 0 9.1666666667 0 1 20 0
0 24 0 . . 8.25 1 1 18 0
0 24 0 . . 9.6666666667 1 1 23 0
1 24 . 0 1 8.25 1 0 22 .
0 23 . 1 0 8.3333333333 0 1 22 1
0 22 0 0 0 9.0833333333 0 0 22 0
1 . 0 . . 8.0833333333 1 0 24 1
0 9 0 0 0 8.3333333333 0 0 22 0
4 24 1 1 0 10 0 0 18 0
1 . 0 . . 9.4166666667 0 0 17 1
4 18 1 0 0 8.0833333333 0 1 18 .
1 20 . . . 8.4166666667 0 1 20 1
2 24 0 0 0 9.4166666667 1 1 18 0
4 14 . 0 0 8 1 0 19 0
2 15 . 0 0 8.75 1 1 17 1
0 23 . 0 0 8.5 0 0 21 0
4 20 0 0 0 8 0 0 20 1
```

```
2 24 0 0 0 9 0 1 22 0
2 15 1 . . 9 0 1 19 .
3 24 0 0 1 9.0833333333 0 0 24 1
1 . . 0 1 8.1666666667 1 0 20 .
1 19 1 0 0 9.0833333333 1 1 20 0
0 . . 0 0 8.3333333333 0 1 21 0
3 14 0 0 1 9.4166666667 0 0 20 1
3 17 . 0 1 8.5 0 0 21 1
0 24 1 0 0 9.8333333333 1 0 18 .
1 17 0 0 0 9.1666666667 0 0 21 0
2 17 . 1 0 8.25 0 0 20 1
2 24 1 1 0 8.25 0 0 20 .
4 16 0 . . 9.0833333333 0 1 24 1
1 . 0 1 0 8 0 1 25 1
3 . 1 0 1 8.75 0 0 20 1
0 18 0 0 1 9.4166666667 0 1 18 1
2 18 . 1 0 8.0833333333 1 1 19 0
2 21 0 0 1 8.5 0 0 21 0
1 21 . 1 0 9.6666666667 0 0 20 .
;
run;
```

A.2 Listwise Deletion

```
/*----------------------Listwise Deletion----------------------*/
data nlsy_miss;
  set nlsy_miss;
run;

/*Fitting a regression model with ANTI as the dependent variable on
  the resulting data set that has removed all cases that have
missing values - the data set will now be much smaller than it was originally*/
proc reg data = nlsy_miss;
  model anti = self pov black hispanic childage divorce gender momage momwork;
run;
```

```
quit;
```

A.3 Pairwise Deletion

```
/*---------------------Pairwise Deletion---------------------*/
data nlsy_miss;
  set nlsy_miss;
run;

proc corr data = nlsy_miss out = pair_corr;
  var anti self pov black hispanic childage divorce gender momage momwork;
run;

/*Fitting a regression model with ANTI as the dependent variable
using the pairwise correlation matrix as input*/
proc reg data = pair_corr;
  model anti = self pov black hispanic childage divorce gender momage momwork;
run;
quit;
```

A.4 Mean Imputation

```
/*---------------------Mean Imputation---------------------*/
data nlsy_miss;
  set nlsy_miss;
run;

/*Performing a proc means on nlsy_miss to identify the mean for each variable*/
proc means data = nlsy_miss mean;
  var anti self pov black hispanic childage divorce gender momage momwork;
  output out = mean_imp;
run;

/*Creating a dummy indicator called "merging" in order to merge the mean
```

```
  onto the data set*/
data nlsy_miss_2;
    set nlsy_miss;
    merging = 1;
run;

/*Renaming the variables from the proc means output in order to merge later on */
data mean_imp;
  set mean_imp;
  merging = 1;
  where _STAT_    = "MEAN";
  rename anti     = anti_imp;
  rename self     = self_imp;
  rename pov      = pov_imp;
  rename black    = black_imp;
  rename hispanic = hisp_imp;
  rename childage = childage_imp;
  rename divorce  = div_imp;
  rename gender   = gender_imp;
  rename momage   = momage_imp;
  rename momwork  = momwork_imp;
run;

/*Merging the two data sets together in order to create one complete data set that has
  no missing values*/
data match_mean;
merge nlsy_miss_2 (in = a)
      mean_imp    (in = b);
  by merging;
  if a;
run;

/*Replacing missing values with the mean*/
data miss_mean;
  set match_mean;
  if anti     = . then anti     = anti_imp;
```

```
  if self     = . then self     = self_imp;
  if pov      = . then pov      = pov_imp;
  if black    = . then black    = black_imp;
  if hispanic = . then hispanic = hisp_imp;
  if childage = . then childage = childage_imp;
  if divorce  = . then divorce  = div_imp;
  if gender   = . then gender   = gender_imp;
  if momage   = . then momage   = momage_imp;
  if momwork  = . then momwork  = momwork_imp;
run;

/*Fitting a regression model with ANTI as the dependent variable on the
  resulting complete data set*/
proc reg data = miss_mean;
  model anti = self pov black hispanic childage divorce gender momage momwork;
run;
quit;
```

A.5 Mode imputation

```
/*---------------------Mode Imputation---------------------*/
data nlsy_miss;
  set nlsy_miss;
run;

/*Determining the mode for each variable*/
proc univariate data = nlsy_miss modes;
  var anti self pov black hispanic childage divorce gender momage momwork;
run;

/*Creating a dummy indicator called "merging"  in order to merge the mode onto
  the data set*/
data nlsy_miss_2;
```

```
  set nlsy_miss;
  merging = 1;
run;

/*The mode is listed below for each of the variables*/
data mode_imp;
  merging       = 1;
  anti_mode     = 0;
  self_mode     = 24;
  pov_mode      = 0;
  black_mode    = 0;
  hisp_mode     = 0;
  childage_mode = 8.25;
  divorce_mode  = 0;
  gender_mode   = 1;
  momage_mode   = 20;
  momwork_mode  = 0;
run;

/*Merging the two data sets together in order to create one complete data set that
  has no missing values*/
data match_mode_1;
  merge mode_imp    (in = a)
        nlsy_miss_2 (in = b);
  by merging;
  if b;
run;

/*Replacing missing values with the mode*/
data miss_mode_1;
  set match_mode_1;
  if anti     =  . then anti     = anti_mode;
  if self     =  . then self     = self_mode;
  if pov      =  . then pov      = pov_mode;
  if black    =  . then black    = black_mode;
  if hispanic =  . then hispanic = hisp_mode;
```

```
  if childage =  . then childage = childage_mode;
  if divorce  =  . then divorce  = divorce_mode;
  if gender   =  . then gender   = gender_mode;
  if momage   =  . then momage   = momage_mode;
  if momwork  =  . then momwork  = momwork_mode;
run;

/*Fitting a regression model with ANTI as the dependent variable
  on the resulting complete data set*/
proc reg data = miss_mode_1;
  model anti = self pov black hispanic childage divorce gender momage momwork;
run;
quit;
```

A.6 Median imputation

```
/*----------------------Median Imputation----------------------*/
data nlsy_miss;
  set nlsy_miss;
run;

/*Creating a macro to identify the median for each variable*/
%macro med_imp(variable);
   proc univariate data = nlsy_miss noprint;
   var &variable.;
   output out = med_imp_&variable.
   pctlpts = 50 pctlpre=med_&variable._;
 run;
%mend;

/*Calling the macro to get the median for each variable*/
%med_imp(anti);
%med_imp(self);
%med_imp(pov);
```

```
%med_imp(black);
%med_imp(hispanic);
%med_imp(childage);
%med_imp(divorce);
%med_imp(gender);
%med_imp(momage);
%med_imp(momwork);

/*Creating one final data set with the medians for each variable*/
data match_med;
  merge
   med_imp_anti
   med_imp_self
   med_imp_pov
   med_imp_black
   med_imp_hispanic
   med_imp_childage
   med_imp_divorce
   med_imp_gender
   med_imp_momage
   med_imp_momwork;
   merging = 1;
run;

/*Merging the two data sets together in order to create one complete data set
  that has no missing values*/
data match_med_2;
  merge   nlsy_miss (in = a)
          match_med (in = b);
  by merging;
if a;
run;

/*Replacing the variables that contain missing values with the median
  for that specific variable*/
data miss_med_1;
```

```
  set match_med_2;
  if anti     =  . then anti     = med_anti_50;
  if self     =  . then self     = med_self_50;
  if pov      =  . then pov      = med_pov_50;
  if black    =  . then black    = med_black_50;
  if hispanic =  . then hispanic = med_hispanic_50;
  if childage =  . then childage = med_childage_50;
  if divorce  =  . then divorce  = med_divorce_50;
  if gender   =  . then gender   = med_gender_50;
  if momage   =  . then momage   = med_momage_50;
  if momwork  =  . then momwork  = med_momwork_50;
run;

/*Fitting a regression model with ANTI as the dependent variable
on the resulting complete data set*/
proc reg data = miss_med_1;
  model anti = self pov black hispanic childage divorce gender momage momwork;
run;
quit;
```

A.7 Regression Imputation

```
/*---------------------Regression Imputation---------------------*/
/*---------------------SELF---------------------*/
data nlsy_miss;
  set nlsy_miss;
  merging = 1;
run;

data nlsy_full;
  set nlsy_miss;
run;
```

```
/*Fitting a linear regression with dependent variable SELF*/
proc reg data = nlsy_miss outest = self_parameter;
  model self = childage divorce gender momage anti;
run;
quit;

/* Renaming the variables in order to merge on later*/
data self (drop = self);
  set self_parameter;
  merging = 1;
  rename childage = childage_est;
  rename divorce  = divorce_est;
  rename gender   = gender_est;
  rename momage   = momage_est;
  rename anti     = anti_est;
run;

/*Merging the two data sets together*/
data nlsy_full;
  merge self        (in = a)
        nlsy_full   (in = b);
  by merging;
  if b;
run;

/*Filling in the missing values for SELF*/
data nlsy_full (drop = childage_est divorce_est gender_est momage_est anti_est
                       intercept _: );
    set nlsy_full;
   if self = . then self = intercept + childage * (childage_est) + divorce *
                           (divorce_est) + gender * (gender_est) + momage *
                           (momage_est) + anti * (anti_est);
run;

/*---------------------POV---------------------*/
/*Using proc logistic regression on categorical variables where POV is the
```

```
  dependent var*/
proc logistic data = nlsy_miss outest = pov_parameter;
  model pov = childage divorce gender momage anti;
run;

/* Renaming the variables in order to merge on later*/
data pov (drop = pov);
   set pov_parameter;
   merging = 1;
   rename childage = childage_est;
   rename divorce  = divorce_est;
   rename gender   = gender_est;
   rename momage   = momage_est;
   rename anti     = anti_est;
run;

/*Merging the two data sets together*/
data nlsy_full;
  merge  pov      (in = a)
         nlsy_full(in = b);
   by merging;
   if b;
run;

/*Filling in the missing values for POV*/
data nlsy_full (drop = childage_est divorce_est gender_est momage_est anti_est
                       intercept _:);
    set nlsy_full;
  if pov = .
  then pi_pov = exp(intercept + childage * (childage_est) + divorce *(divorce_est)+
                  gender *(gender_est) + momage *(momage_est) + anti *(anti_est))/
                  (1+exp(intercept + childage *(childage_est) + divorce *(divorce_est)+
                  gender *(gender_est) + momage *(momage_est) + anti *(anti_est)));
run;

/*Creating a rule to fill in the missing values:
```

```
  if probability > 0.5 then POV = 1 and if probabilty <= 0.5 then POV = 0*/
data nlsy_full (drop = pi_pov);
   set nlsy_full;
 if pi_pov ne .
    then do;
       if pi_pov > 0.5 then pov = 1;
       else pov = 0;
    end;
run;

/*---------------------MOMWORK---------------------*/
proc logistic data = nlsy_miss outest = momwork_parameter;
   model momwork = childage divorce gender momage anti;
run;

/* Renaming the variables in order to merge on later*/
data momwork;
   set momwork_parameter;
   merging = 1;
   rename childage = childage_est;
   rename divorce  = divorce_est;
   rename gender   = gender_est;
   rename momage   = momage_est;
   rename anti     = anti_est;
run;

/*Merging the two data sets together*/
data nlsy_full;
   merge  momwork   (in = a)
          nlsy_full (in = b);
  by merging;
  if b;
run;

/*Filling in the missing values for MOMWORK*/
data nlsy_full(drop = childage_est divorce_est gender_est momage_est anti_est
```

```
                        intercept _: );
   set nlsy_full;
  if momwork = .
  then pi_momwork = exp(intercept + childage *(childage_est) + divorce *(divorce_est)+
                  gender *(gender_est) + momage *(momage_est) + anti *(anti_est) )/
                 (1+exp(intercept + childage *(childage_est) + divorce *(divorce_est)+
                 gender *(gender_est) + momage *(momage_est) + anti *(anti_est)));
run;

/*Creating a rule to fill in the missing values:
  if probability > 0.5 then MOMWORK = 1 and if probabilty <= 0.5 then MOMWORK = 0*/
data nlsy_full(drop = pi_momwork);
   set nlsy_full;
  if pi_momwork ne .
     then do;
        if pi_momwork > 0.5 then momwork = 1;
        else momwork = 0;
      end;
run;

/*---------------------RACE---------------------*/
/*Creating a rule on the missing data set to determine RACE*/
data race;
  set nlsy_miss;
 if black = 1          then race = "B";
  else if hispanic = 1 then race = "H";
    else race = "W";

 if black = . then race = "";
  else if hispanic = . then race = "";
 merging = 1;
run;

data nlsy_full;
  set nlsy_full;
 if black = 1 then race = "B";
```

```
  else if hispanic = 1 then race = "H";
  else race = "W";

 if black = . then race = "";
  else if hispanic = . then race = "";
run;

/*using 10% percent level of significance*/
proc logistic data = race outest = race_reg;
model race = anti |gender| divorce| momage |childage/selection = forward slentry = 0.1;
run;

/*Keeping only the significant variables (ANTI and MOMAGE) and renaming them
  for merging later*/
data race1(keep = momage_est anti_est merging Intercept_B intercept_H);
  set race_reg;
  merging = 1;
  rename momage = momage_est;
  rename anti = anti_est;
run;

data nlsy_full;
  merge race1     (in = a)
        nlsy_full (in = b);
  by merging;
  if b;
run;

data nlsy_full;
  set nlsy_full;
 if race eq "" then do;

/*NON-HISPANIC WHITE*/
   pi_white = 1/(exp(intercept_B + momage *(momage_est) + anti *(anti_est))
              + exp(intercept_H + momage *(momage_est) + anti *(anti_est))+ 1);
```

```
/*BLACK*/
   pi_black = pi_white *exp(intercept_B + momage *(momage_est) + anti *(anti_est));

/*HISPANIC*/
   pi_hisp = pi_white * exp(intercept_H +  momage * (momage_est) +  anti * (anti_est));

/*Determining RACE based on the highest probability*/
  if pi_white = max(pi_white, pi_black, pi_hisp)
                then filled_in_race = "W";
   else if pi_black = max(pi_white, pi_black, pi_hisp)
                then filled_in_race = "B";
   else filled_in_race = "H";
  end;
run;

/*Creating a new data set with the filled in values*/
data fill_in(drop = race pi_black pi_white pi_hisp filled_in_race) ;
  set nlsy_full;
  if race = "" then do;
    if filled_in_race = "W" then do;
           black = 0;
           hispanic = 0;
    end;
  if filled_in_race = "B" then do;
           black = 1;
           hispanic = 0;
  end;
  if filled_in_race = "H" then do;
           hispanic = 1;
            black = 0;
  end;
end;
run;

/*Fitting a regression model with ANTI as the dependent variable on
  the resulting complete data set as all of the variables no longer
```

```
  contain missing values*/
proc reg data = fill_in;
  model anti = self pov black hispanic childage divorce gender momage momwork;
run;
quit;
```

A.8 KNNI

```
/*----------------------KNNI----------------------*/
/*SELF*/
data nlsy_miss_D;
  set nlsy_miss;
run;

data nlsy_miss_DM nlsy_miss_DC;
  set nlsy_miss_D;
  if self = . then output nlsy_miss_DM;
  else output nlsy_miss_DC;
run;

/*Performing PROC DISCRIM analysis*/
proc discrim data = nlsy_miss_DC test = nlsy_miss_DM testout = _score_self
                    method = npar k = 4 testlist threshold = 0;
  class self;
  var childage divorce gender momage anti;
run;

/*Replacing the missing values*/
data full_data(keep=anti self pov black hispanic childage divorce gender momage momwork);
  set nlsy_miss_DC
      _score_self;
/*If values are the same - can choose either neighbour - in this case 17 was chosen*/
  if _into_ = . then _into_ = 17;
```

```
  if self   = . then self = _into_;
run;

/*POV*/
data nlsy_miss_D(keep = merging anti self pov black hispanic childage divorce gender
                        momage momwork);
  set full_data;
run;

data nlsy_miss_DM nlsy_miss_DC;
set nlsy_miss_D;
if pov = . then output nlsy_miss_DM;
    else output nlsy_miss_DC;
run;

/*Performing PROC DISCRIM analysis*/
proc discrim data = nlsy_miss_DC test = nlsy_miss_DM testout = _score1_pov
                                        method = npar k = 5 testlist threshold = 0;
class pov;
var childage divorce gender momage anti;
run;

/*Replacing the missing values*/
data full_data(keep =   anti self pov black hispanic childage divorce gender
                        momage momwork);
  set nlsy_miss_DC
      _score1_pov;
if pov = . then pov = _into_;
run;

/*MOMWORK*/
data nlsy_miss_D;
  set full_data;
run;

data nlsy_miss_DM nlsy_miss_DC;
```

```
  set nlsy_miss_D;
  if momwork = . then output nlsy_miss_DM;
  else output nlsy_miss_DC;
run;

/*Performing PROC DISCRIM analysis*/
proc discrim data = nlsy_miss_DC test = nlsy_miss_DM testout = _score_momwrk
                                        method = npar  k = 5 testlist threshold = 0;
class momwork;
var childage divorce gender momage anti;
run;

/*Replacing the missing values*/
data full_data(keep = merging anti self pov black hispanic childage divorce gender
                        momage momwork);
    set nlsy_miss_DC
     _score_momwrk;
if momwork = . then momwork = _into_;
run;

/*RACE*/
data full_data;
  set full_data;
  if black = 1 then race = "B";
  else if hispanic = 1 then race = "H";
  else race = "W";

  if black = . then race = "";
  else if hispanic = . then race = "";
  merging = 1;
run;

data nlsy_miss_DM nlsy_miss_DC;
  set full_data;
if race = "" then output nlsy_miss_DM;
    else output nlsy_miss_DC;
```

```
run;

/*Performing PROC DISCRIM analysis*/
proc discrim data = nlsy_miss_DC test = nlsy_miss_DM testout = _score1_race
                                        method = npar k = 5 testlist threshold = 0;
class race;
var childage divorce gender momage anti;
run;

/*Replacing the missing values*/
data full_data;
  set nlsy_miss_DC
      _score1_race;
 if race = "" then race = _into_;
run;
/*Filling in the values for RACE*/
data full_data (keep = merging anti self pov black hispanic childage divorce gender
                        momage momwork);
  set full_data;
  if race = "B" then black = 1;
  else black = 0;

  if race = "H" then hispanic = 1;
  else hispanic = 0;
run;

/*Fitting a regression model with ANTI as the dependent variable on the resulting
  complete data set*/
proc reg data = full_data;
  model anti = self pov black hispanic childage divorce gender momage momwork;
run;
quit;
```

A.9 EM Algorithm

```
/*Computing EM algorithm*/
```

```
proc mi data = nlsy_miss nimpute = 0;
em itprint outem = nlsyem;
var anti self pov black hispanic childage divorce gender momage momwork;
run;

proc print data = nlsyem;
run;

/*Performing proc reg using inputs from covariance matrix*/
proc reg data = nlsyem outest = a;
model anti= self pov black hispanic childage divorce gender momage momwork;
run;

/*Bootstrapping Technique*/
proc surveyselect data = nlsy_miss method = urs n = 581 reps = 1000
out = bootsamp outhits;

proc mi data = bootsamp nimpute=0 noprint;
var anti self pov black hispanic childage divorce gender momage momwork;
em outem=nlsyem;
by replicate;
proc reg data  =nlsyem outest = a noprint;
model anti = self pov black hispanic childage divorce gender momage momwork;
by replicate;
run;

proc means data = a std;
var self pov black hispanic childage divorce gender momage momwork;
run;

/*Calculate obtain P-values*/
data test;
p_self =     2 *  probnorm (-2.92979);
p_pov =      2 * (1-probnorm(3.84381));
p_black =    2 * (1-probnorm(0.52827));
p_hisp =     2 *  probnorm (-1.959);
```

```
p_childage = 2 *  probnorm (-0.03973);
p_divorce =  2 *  probnorm (-0.69295);
p_gender =   2 *  probnorm (-4.80981);
p_momage =   2 * (1-probnorm(0.76211));
p_momwork =  2 * (1-probnorm(1.49369));
run;
```

A.10 Multiple Imputation

```
/*---------------------Multiple Imputation---------------------*/
proc mi data = nlsy_miss out = mi_miss_out nimpute = 15;
  var anti self pov black hispanic childage divorce gender momage momwork;
run;

/*Fitting a regression model with ANTI as the dependent variable on the resulting
data set*/
proc reg data = mi_miss_out outest = mi covout;
  model anti = self pov black hispanic childage divorce gender momage momwork;
  by _imputation_;
run;

proc mianalyse data = mi;
  var intercept self pov black hispanic childage divorce gender momage momwork;
run;
```