An Analysis of the Missing Data Methodology for Different Types of Data

A THESIS PRESENTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE
OF

MASTER OF APPLIED STATISTICS

AT MASSEY UNIVERSITY, ALBANY
NEW ZEALAND

Judith-Anne Scheffer

2000

Abstract

Missing data is an eternal problem in data analysis. It is widely recognised that data is costly to collect, and the methods used to deal with missing data in the past relied on case deletion. There is no one overall best fix, but many different methodologies to use in different situations.

This study was motivated by the writer's time spent analysing data in the nutrition study, and realising how much data was wasted by case deletion, and subsequently how this could bias inferences formed from the results. A better method (or methods), of dealing with missing data (than case deletion) is required, to ensure valuable information is not lost.

What is being done: What is in the literature? The literature on this topic has exploded with new methods in recent times. Algorithms have been written and incorporated based on these methods into a number of statistical packages and add-on libraries.

Statistical packages are also reviewed for their practicality and application in this area. The nutrition data is then applied to different methodologies, and software packages to assess different types of imputation.

A set of questions are posed; based on type of data, type of missingness, extent of missingness, the required end use of the data, the size of the dataset, and how extensive that analysis needs to be. This can guide the investigator into using an appropriate form of imputation for the type of data at hand.

A comparison of imputation methods and results is given with the principal result that imputing missing data is a very worthwhile exercise to reduce bias in survey results, which can be achieved by any researcher analysing their own data.

Further to this, a conjecture is given for using Data Augmentation for ordinal data, particularly Likert scales. Previously this has been restricted to either person or item mean imputation, or hot deck methods. Using model based methods for imputation is far superior for other types of data. Model based methods for Likert data are achieved by means of inserting the linear by linear association model into standard missing data methodology.

Acknowledgements

I wish to offer my sincerest thanks to my supervisor, Doctor Barry W. M^{c}Donald, for all his helpful advice, comments and efforts on my behalf, and also for his encouragement and mentoring throughout the course of this degree.

My thanks also go to Doctor Howard P. Edwards for his assistance in 'Matters Bayesian', Ms Katya Ruggiero for her ability to challenge practices and ideas, Mrs Kay Rowbottom for her assistance with the production of the flowcharts, and Synthia for her encouragement.

Thanks also go to Mrs Patsy E. Watson for providing via my supervisor, the nutrition dataset; and also to Ms Janet Norton for providing her dataset, via Professor Graham R. Wood.

Lastly but not least, I would like to thank my family (the thesis orphans) for putting up with my frequent absences for long periods to do this work.

Blessed is the man who perseveres under trial,
because when he has stood the test,
he will receive the crown of life that
God has promised to those who love him.
James 1:12

Table of Contents

List of Tables and Figures

Notation and Abbreviations

BLR	Binary Logistic Regression
CD	Case Deletion
EM	Expectation Maximisation (algorithm)
EM Imp	Imputation via the EM algorithm
GLM Imp	General Location Model Imputation
HD	Hotdeck (Imputation)
iid	Independent identically distributed
LUM	Look up methods
LVCF	Last Value Carried Forwards
MCAR	Missing Completely at Random
MAR	Missing at Random
Mean Imp	Mean family of Imputation
MI	Multiple Imputation
MI BB	Multiple Imputation Bayesian Bootstrap
MICE	Multiple Imputation by Chained Equations
MI DA	Multiple Imputation via Data Augmentation
MI EM	Multiple Imputation via the EM algorithm
N.Neighbour	Nearest Neighbour
N Nets	Neural Networks
NLR	Nominal Logistic Regression
NMAR	Not Missing at Random (Informatively Missing)
OLR	Ordinal Logistic Regression
PMM	Predictive Mean matching
Reg Imp	Regression Imputation
SHHD	Sequential and/or Hierarchical Hotdeck
SI	Single Imputation
St Reg	Stochastic regression Imputation

W	Indicator for Missingness
X	Co-variate in model
Y	Variable of interest
$\hat{\alpha}$	Gamma Parameter (Ch 8)
$\hat{\beta}$	Gamma Parameter (Ch 8)
$\hat{\beta}$	Regression Coefficient Estimate (Ch 9)
θ	Distribution Parameter
$\hat{\theta}$	Maximum Likelihood Estimate of the Parameter
ψ	Missingness Parameter in Model

1 Introduction: Is Ignorance Bliss?

1.1 The thesis

This thesis is concerned with a perennial problem of survey data, namely what to do with it when it is missing, or unobserved. It has been developed with human nutrition, and genetically modified foods in mind.

1.1.1 An overview of the thesis

Firstly the literature developments, and software advances of recent times are looked at, together with the methodology for handling missing data. A set of suggestions are given for when to use each type of imputation. Different software is applied to different types of data, the completed datasets are run through a regression analysis, and then the resulting coefficients and standard deviations are examined for consistency.

The problem of imputing Likert data is then considered. This is a gaping hole in the literature on missing data, and needs to be addressed.

1.1.2 Background

Missing data analysis really took off after the fundamental paper by Donald Rubin was published in 1976 defining the important terms Missing At Random (MAR), and Observed At Random (OAR). Rubin is the number one pioneer in this field, and has remained so for more

than 20 years. Another important investigator here is Roderick Little, and together they have written the monograph 'Statistical Analysis with Missing Data', published in 1987. This is probably the single greatest authority on this subject. There are a number of people who have made important contributions in the field (of multiple imputation), and without exception, all have been Rubin's PhD students, or spent a time studying with him. Much of the work in this thesis is based on Rubin's framework for missing data inference, and also on fitting the linear by linear association model into the work done by Joe Schafer, a student of Rubin's.

1.1.3 The Remaining Chapters

Chapter 2 is all about data collection, and what can be done to avoid missing data and just what can be done to prevent it. The causes of missing data are also considered.

Chapter 3 is concerned with the literature and methodology surrounding the classification of types of missingness, of the concept of ignorability, and non-ignorability, and also cures in the literature for missing data, some good, some not so good.

Chapter 4 gives the motivation for this study, and the problems addressed by this work. Also here is a description of the datasets used.

Chapter 5 gives a brief description of many types of imputation currently available, and classifies them according to the type of imputation that they are.

Chapter 6 is an overview of the software available to deal with this problem. Included is commercial statistical software, as well as commercial non-statistical software, and non-commercial routines for use with the major statistical packages, as well as other useful packages.

Chapter 7 is a set of rules for dealing with missing data, how to fit which imputation method to which kind of data, and what is the best method to use, although usually more than one method is presented as being suitable for any given data situation. The appendix gives these rules in the form of a flowchart.

Chapter 8 gives the background for a suggested approach for ordinal data, and an account of how the linear by linear association model fits into the frame work of missing data theory, particularly for log-linear models. This is particularly suitable for Likert data.

Chapter 9 gives an account of various types of imputation, when applied to the nutrition data set. These are then recombined back into datasets and the complete datasets are then analysed using the same regression model. The multiply imputed datasets have their estimates and standard errors recombined using Rubin's rules. The resulting regression coefficients are then compared for consistency.

Chapter 10 has a section on the ethics of imputation,(and the abuses), and then a part for the conclusion with suggestions for further study.

2 Literature Review of Data Collection Methodology

2.1 What is Missing Data?

Data is generally arranged in data sets with each case forming a row, and each variable forming a column, to provide a rectangular matrix. Missing data is the term used to describe the event that any of the cells of this matrix are empty.

Each observation may be continuous and the values may represent values of say height, weight or age. Other variables may be categorical; either ordinal, for example represent numbers of years of education, place in family of children, income bands etc; or nominal variables such as ethnicity and gender. All these may be contained in the same data matrix.

Ideally for standard methods of analysis, i.e. regression, ANOVA etc, all values should be observed. When they are not observed due to, for example refusal to report income or inability to take a measurement, then these unobserved cases are known as missing values.

Usually the main Statistical Packages, Minitab, SAS, S-PLUS, SPSS tend to deal with these missing values by excluding the cases for which any of the observations are missing. This is known as case-deletion. If the non-observed values are across a number of variables, the usual manner of case deletion tends to discard a lot of the available data. Often too the completely observed units are not a random sample of the original sample; this is described as not missing completely at random (Little and Rubin, 1987).

2.1.1 Ways in which Missing Data Arise

The ways in which missing data may arise are many and varied.

Firstly it is a common problem in Longitudinal Analysis - people simply disappear, due to moving etc; and with the best will in the world the researcher cannot find them. Often chunks of data that were there at the outset become missing at the later observations.

In agricultural experiments, very often crops may fail over a particular block of the experiment, for example a river flooded, or the sheep (or chickens) got in and ate the experiment.

In surveys, as well as the census, there is also the problem of deliberate non-response, that is refusal to answer questions like income, smoking and drinking (alcohol) habits, and religion. Sometimes answers given in a survey are completely incompatible with what the data analyst considers believable, but was not checked at the time of the interview.

Censoring can occur when an experimental animal dies, or a child is born etc; and the time to these events is measurable.

'Non-Applicable' questions (due to a 'no' to a question causing a 'go to' in a survey) may cause missing values, quite legitimately, as in the question "How long have you been married?" for a respondent who has never married, or the amount of alcohol consumed by a teetotaller. These are occasionally answered with a zero, as a means to getting

around this problem. However this does not have the same meaning as a measured zero.

Missing Data may be a process independent of the respondent that is due to data entry errors, or data collection errors. Careful checking, and re-checking of data generally can correct these. If however, it is on the part of the respondent, i.e. a refusal to answer, then that will be a missing value.

If some form of imputation is not applied, then any observation with a variable not observed will be deleted from the analysis - this is called case deletion. The consequence, occurring often in survey research, is that what started off as an adequate sample size, now becomes insufficient for the analysis.

Missing data are referred to as ignorable missing data (see Chapter 3.5) if essentially the missingness is at random (See Chapter 2.3.3). If patterns are found in the missing observations, then they cannot be considered random.

Censoring (for example, moving away or deaths) and legitimately not applicable responses are forms of missingness, which are often ignorable, but not always.

2.1.2 Inference and missing data

Rubin, particularly through his 1976 Biometrika paper, dominates the literature around inference and missing data. In this paper he defines the weakest possible conditions under which missing data may be

ignorable, for making inferences from samples about a population. These conditions are intuitive, and non-parametric in the sense that they are not tied to any particular distribution.

Rubin defines the concept of Missing At Random (MAR), and Observed At Random (OAR - See Chapter 2.3.3). That is, the missingness does not depend on the actual value of the missing variable. Rubin also described the concept of parameter distinctness. When parameter distinctness and MAR hold (see Chapter 2.3.3, for definitions), then the missingness mechanism may be ignored, for either Bayesian or Likelihood inference. Little and Rubin (1987) further develop the idea of Missing Completely at Random (MCAR), which is when both Missing At Random (MAR), and Observed At Random hold. This is in fact a far more stringent condition (see Chapter 3.4) and very rarely holds, but is necessary for case deletion to be unbiased. Heitjan and Basu (1996) distinguish between MAR and MCAR in their paper.

2.1.3 Consequences of Missing Data

The consequence of having missing data is to generally throw away cases that contain these missing data. This is wasteful of data, which may be expensive to collect. Worse still the sample may not match the target population, and this leads to bias.

2.1.4 Bias

Missing data occurs for all kinds of reasons, and the assumptions made about missing values are often incorrect, for example MAR. The

resulting bias can be quite large, especially when the analysis requires complete case data, where cases are only used if each variable is completely observed. To choose an appropriate method for dealing with missing values requires knowledge of why the data are missing in the first place.

To investigate the reason for the missing data, validation studies should be carried out to understand whether missing values are random across the study population, or occur more frequently in specific subgroups. In case control studies, very often the case information is more complete than that of the controls.

Depending on the reason for the missing data, we may decide to use imputed data, or stick with complete case data. Complete case data is only unbiased if the missingness is MAR. In the case of categorical data a test can be performed by comparing the odds ratio for some factor; for the category of missing values, as against the odds ratio for the non-missing category. The other factors can show whether a variable is indeed MCAR. If the odds ratios are different, then a complete case analysis will be biased.

If a variable is a confounder, then the analysis will be seriously biased if the values for the confounding variable are missing. Groves (1989) gives a formula for Non Response Bias (calculated by using the observed incomplete table, or from subjects with complete data only).

$$y_r = y_n + nr/n\ (y_r - y_{nr})$$

The bias for characteristic Y is $Y_r - Y_n$ where

n = The total sample size for the survey (respondents and nonrespondents)

r = The group of respondents

nr = the group of non respondents

nr/n = proportion of nonrespondents

y_r =the characteristic value obtained from the survey (for example the mean)

y_n = the characteristic value which would be obtained with a total response for the survey,

Y_{nr} = the characteristic value for the noresponders (not observed), and

$y_r - y_{nr}$ = difference in characteristic value.

Different patterns of nonresponse are associated with different biases, with a cumulative and possibly confounding effect (Gray, Campanelli, Deupchat, Prescott-Clarke, 1996).

2.1.5 Omitting covariates

Sometimes covariates containing missing data are omitted, as a means of producing a rectangular complete case data matrix: again this may be throwing away valuable information. It is well known in the regression literature that omitting variables may result in biased regression parameters and predicted fits.

2.2 Forms of Nonresponse.

Non response may be by design, or unit non-response, or item non-response. What actually causes non-response? Unit or Item non-response can occur in many forms. (For example, do the don't knows really not know?) One way it can come about is through poorly organised questionnaires, where the respondent may be incapable of

answering the questions, or may not know the answers. This is a particular problem with telephone surveys, e.g. if the respondent is deaf, or has some other disability. In this case the interviewer (by phone, mail or face to face) cannot expect a reasonable answer from the respondent if they are incapable of answering the questions. The respondent may say "I don't know" because it is easier. There is also a situation called Panel non-response (Longitudinal, see Chapter 2.4.4).

2.2.1 Unit Nonresponse.

Unit Nonresponse - The entire case is missing. These are the non-contactable cases (non-coverage). Unit Non Response can occur even when surveys are properly designed, with the primary sampling units being chosen at random, and within whatever sampling scheme is used. A particular household may be chosen for study, but the interviewer can have great difficulty making contact with the respondents. It may not be because they do not want to participate (refusal), but just that they are uncontactable (see Chapter 2.4.5).

Refusals: What causes refusals? In the case of unit non-response these include people with an attitude which is against surveys for whatever reason. People in this category are often older (for example > 70), who have a great mistrust of younger people asking questions or they may be from immigrant groups, who have survived wars etc., and are very careful about whom they tell what. Careful explanation of what the survey is about can go some way to alleviating this problem. Refusal can be due to the respondent taking offence at some sensitive questions (the prime one is personal income), which can lead to both

item non-response (refusal to answer individual questions) or unit non-response (refusal to answer any questions at all).

Unit non-response can be dealt with by weighting, i.e. post stratification, when a population estimate is required from a survey. This is probably the only legitimate means of dealing with unit nonresponse, although field substitution is sometimes used. Field substitution is also known as Non Respondent Substitution. This occurs when a particular case, selected by the sampling scheme which is used, is uncontactable (often after repeated attempts on the part of the data collector). A demographically similar case is chosen using the same sampling scheme as the originally chosen case was selected under (Lessler, Kalsbeek,1992). However field substitution can lead to bias if the reason for the unit nonresponse is not understood and known to be random.

2.2.2 Item Nonresponse

Item nonresponse again can be due to the incapability of the respondent to respond to the particular question. For example:

1. A person who works multiple part-time on call casual jobs has very little idea of their annual income until they actually file a tax return. Rather than asking weekly income a better question could be 'what did you earn in the last financial year?'

2. In a survey on pregnancy, asking a pregnant mother, "When was your Last Monthly Period?" is not always the best question. If it is known it furnishes an excellent predictor of the baby's expected date of

delivery, but most mothers simply don't know. So though they are happy to answer other questions, this one becomes one of Item non-response.

3. Again the interviewer may ask just one question which is outside the respondents knowledge.

It may not be the case that "Don't know" is correct. It can be that the questions are of a sensitive nature, and "Don't know" may be a convenient way of covering up a refusal. 'Don't Knows' are often treated as ignorable missing data, or even as a valid response category in their own right. In many cases 'Don't Knows' are common among respondents with low education and low income - particularly if it is an issue that they do not particularly care about. However Rubin, Stern and Vehovar(1995) show that if 'Don't Knows' are conditionally dependent on some other variable, then this is MAR, which is ignorable if distinctiveness also holds (see next section, and Chapter 3.5) and more correctly should be treated using methods suitable for MAR. Therefore likelihood based inferences will be correct, such as MLE's or Bayesian Posterior Distributions.

2.3 Missing Data Mechanism

The missing data mechanism is a description of the probability distribution of the pattern of missing observations. Suppose Y is an n×p matrix, n rows of observations, on p variables, and W an n×p matrix that is an indicator for missingness, so that $y_{ij} = 1$ if missing, and 0 otherwise. This data is then fully modelled with its missing data mechanism by a distribution f (Y|θ) for Y with parameter θ, and a

distribution f(W|Y, ψ) for W given Y, with parameter ψ. We denote $Y=(Y_{obs},Y_{mis})$, where Y_{obs} = Observed values of Y, Y_{mis} = missing values of Y. If the probability Y is observed depends on the value of X_i , then missing data may:

Case1: depend on Y (that is; on the unobserved value)

Case 2: depend on X but not Y

Case 3: be independent of X and Y. Cases 2 and 3 are ignorable but case 1 is not. Here the standard case deletion methods used by the major statistical software are inappropriate, as use of case deletion will lead to biased inferences.

2.3.1 Parameter distinctiveness

"Parameter distinctiveness holds where there are no a priori ties between the parameters of the missingness model, and those of the data model." (Heitjan et al, 1996). If the data are MAR, and θ and ψ are distinct (independent) the likelihood inference for θ can be based on integrating out the density $f(Y_{obs},Y_{mis}|\theta)$ without including a term for the missing data mechanism.

2.3.2 MCAR

If all the observed data are observed at random, and the non-observed data are missing at random, then this is known as Missing Completely at Random. MCAR holds if $f(W|Y_{obs}, Y_{mis}, \psi = f(W|\psi)$ for all Y_{obs} and Y_{mis} - that is the missingness does not depend on observed or missing values of Y.

An example of MCAR is where the unobserved units are a random sub-sample of the total sample. In this case the reason for the missing data is ignorable. Rubin (1976) suggests that case 3 be called Missing At Random, and Observed At Random, (that is, independent of x and y) and these two concepts together be known as Missing Completely at Random. A simple example would be to imagine all the data cards being shuffled, and then the middle fifth being discarded, this would be missing completely at random. When this condition holds, all inferences to the population of complete case methods, available case methods (see Chapter 5), non-parametric methods and generalised estimating equations are valid (Heitjan, 1997).

Park and Lee (1997) proposed a test for Missing Completely at Random for longitudinal data with missing observations. In this test the data are stratified according to the missing value pattern (by weighted least squares), then tested for homogeneity across each strata (of the demographic variables on which the strata are constructed, as well as the indicator for strata membership). If a regression coefficient for the indicator variable, for the strata, is not zero and if sufficiently different, then the missing data mechanism is not MCAR. Chen and Little (1999) provide a test for MCAR for Generalised Estimating Equations, a Wald type test. This is particularly appropriate for longitudinal data. The statistic is of the form:

$$d = \sum_{k=1}^{W} n_k \{\hat{g}_k(\theta) - g_k(\hat{\theta}_c)\}' V_k^{-1} \{\hat{g}_k(\theta) - g_k(\hat{\theta}_c)\}$$

where W = the number of estimating equations according to the missing value pattern;

$\sum_{k=1}^{W} n_k$ = the number of observations in each missing data pattern, of which there are W;

d = the test statistic;

$g_k(\theta_c)$ is the maximum identifiable parameter;

$\hat{\theta}_c$ = the locally consistent parameter estimate;

θ is the parameter of interest.

g_k is the kth function (for the kth missing data pattern) $\hat{g}_k$ is the estimate of the kth function (for the kth missing data pattern) and V_k is the variance.

However if the estimating equations are likelihood scoring equations, the test is a Wald test. It is not necessary to test for MCAR when using likelihood methods, since likelihood based methods remain valid under the weaker MAR condition.

2.3.3 MAR

If case 2 applies this is just missing at random. Most analyses assume the mechanisms that caused the data to be missing are in someway accidental, and that each value is equally likely to be missing (Rubin, 1976).

If the probability that an observation is missing depends on the values of other observed items, but not the value of the missing variable itself, the missing data are Missing at Random.

We can denote the situation when MAR holds by $f(W|Y_{obs}, Y_{mis}, \psi) = f(W|Y_{obs}, \psi)$ for all Y_{mis}, where W is an indicator for missingness. That is missingness does not depend on the missing value Y_{mis}, but may depend on another (other) observed value(s) in the dataset.

Maximum Likelihood estimates are valid, even if the parameters of the missing data mechanism are not estimated at the same time. This is the assumption used in Multiple Imputation procedures (Heitjan,1997).

When the outcomes are categorical, Park and Brown (1994) suggest that a log linear model be used to adjust for the non-response. Essentially the frequency table has additional columns relating the frequency of the missing values. Then the likelihood function is maximised using the additional columns. This is called ignorable Nonresponse when the Nonresponse mechanisms are independent of the actual missing value, Y_{mis}. It is called not ignorable when the probability of the variable being missing is different for different responses. Park and Brown also suggest that even a rough estimate of the uncertain probabilities of the parameter (of particular interest) can result in improved estimates of the cells.

A non-response model is defined as a loglinear model for the full array of X Y and W, where W is an indicator variable for missingness. It is assumed where non-ignorable non-response is present, and so is the YW (interaction) term. The EM algorithm is used, the E step to allocate frequencies to the non observed cells, via the marginal totals, and the M step to estimate the maximum likelihood. A problem here can be when a model is unidentifiable because there are fewer degrees of freedom than the independent parameters.

It is obvious that the reasons behind the missing observations - particularly in surveys- are often not random, but due to a particular cause, the question being asked, or that a group or strata respond

differently to being asked a certain question. This leads to non-response bias.

2.3.4 NMAR

If case 1 applies, then the data are not missing at random, or not observed at random as the unobserved values Y depend on subclasses defined by X.

This occurs when the missingness depends on the variable value that is missing. For example suppose there is a pattern behind the non response, where all respondents with high income do not reply to a particular question, (say about income), this is nonignorable nonresponse. Correct analyses are problematic, as the probability of missingness depends on the missing values, and this needs to be estimated. Very often one can only estimate it using external (validation) sources of data, such as Census data. Even so the resulting estimates may well be unreliable, and certainly very difficult to compute.

At the present time nothing is available in terms of statistical software to deal with this problem (Heitjan, 1997).

2.3.5 Patterns of Missing Data

Often it is impossible to tell from the data whether the mechanism leading to missing data is MCAR, MAR or nonignorable. Attention

needs to be paid to this in the data collection phase, to allow this to be assessed, so informed choices may be made in the analysis phase.

If only a few data points are missing, (<1%) then all methods will give good results, be they case deletion, right through to multiple imputation! In a case such as this however, the less complicated the imputation method used, the better (particularly when the complete sample size is required to be preserved for later analysis). That is methods such as look up tables, or mean substitution, or hotdecking will be adequate, as this will have minimal effect on the variance estimate (see Chapter 7).

When the missingness is less than 5%, and ignorable (see Chapter 3.5.1) methods involving single imputation may be used (see Chapter 7). The advantage of using single imputation methods is that the filled in (or imputed) dataset may be then used in other (later) analysis as if it were a complete data set. (The implications of this are discussed in Chapter 10). For many investigators, this is what they require of those doing the data collection phase. In Europe single complete clean data sets are required and the data collectors are paid accordingly to 'come up with the goods'. (This is disturbing, and is discussed in Chapter 10).

When in doubt about the type of missingness, a sensitivity analysis will be illuminating of this point (Heitjan, 1997). Producing two extreme imputations, one of high and the other of low values can achieve this. The complete data analyses are then compared to determine the range of potential sensitivity. When this is sufficiently narrow, then the simple single imputation methods are acceptable. However if an explicit imputation model is suspect (it provides imputed values outside the known reasonable range), then an implicit model may be used,

such as predictive mean matching. This has been suggested for use with Likert-type data (see Chapter 3.5.8).

2.4 Types of data in Surveys

Monotone data patterns: When analysing survey data containing missing values, It is helpful to arrange the data so that the first variable(s) are the most observed, and the last are the least observed.

2.4.1 Surveys

In Sample surveys there are three types of non response

1 Non contact, the case selected is uncontactable. This is usually replaced by another similar case.

2 Refusal- the respondent refuses part or all of the survey, -if all, then field replacement is an option, provided the reason for the missingness is known, and the replacement is selected under the same sampling scheme as the original respondent, otherwise bias becomes an issue.

3 Incompetency - language barriers, illness, and inability to participate for whatever reason.

The reality is however that no survey ever achieves remotely close to a 100% response rate. Indeed when a very high response rate is achieved, say 95%, a more cynical view would be - Is there something wrong with the selection procedure - Did the data collection involve too

many incentives (that is, is the data honestly answered, and trustworthy?)

To clarify terminology here a nonrespondent is a case that is selected but not measured. The response rate and the different characteristics between responders, and nonresponders, on the other hand, determine nonresponse bias (Steel, Vella and Harrington.1996).

2.4.2 Occurrences of Nonresponse in Surveys

Nonresponse in surveys may be due to genuine don't knows, uncontactability, inability to answer, refusal, attrition or just poor technique.

2.4.3 Inevitable missingness in Surveys

This can be lead to by 'don't know' responses. If a value is unobserved, say in an opinion poll, due to the respondents inability to choose, this is more correctly a don't know response. A 'don't know' may or may not be inevitable, but if it really is 'don't know' then it will inevitably be missing. According to Rubin, Stern, Vehovar (1995) a 'don't know' response can be considered a true missing value, (but if it is a refusal to answer, it has an underlying response). Don't Knows are ignorable nonresponses.

2.4.4 Longitudinal drop out mechanism

In most regression situations, when real data is from real studies, particularly longitudinal studies then missing values are a major problem. Incompletely observed data in longitudinal studies is generally due to attrition. Attrition may occur because of difficulty to contact a respondent, refusal, inability to be traced (moved etc), and death. This occurs in longitudinal data such as a nutrition study. Here data may be collected on a subject at the beginning of the study (or when a particular respondent enters the study) with X follow-ups over a Y year period. A person may participate the first time, which is good because very important demographic data is collected at this point. However respondents moving, or being too ill to take part in the follow-up interviews hamper later data collection efforts. They may miss one 'wave' of data collection, but then come back in on a later one; (this was true in the Nutrition study for a number of women) or they may drop out altogether.

Mobility is a major cause of nonresponse in these longitudinal studies, and very often tracing the respondents becomes an impossible task particularly when they 'go to ground' and don't wish to be found.

If analyses must be based on using fully observed data (in a large longitudinal study) - then the method to obtain fully observed data is case deletion. Apart from being wasteful of valuable data, case deletion implies that relationships within the data are seriously compromised, and variances are underestimated. So if the uncertainty due to the missing value mechanism is not taken into account, then inaccurate results are obtained. For this reason response bias needs

to be fully investigated. A comparison can be obtained by using complete case analysis and comparing with the imputed results, to record the impact of the missing data on the analyses. If the nonresponse is unit nonresponse, then weighting adjustments can be used. Imputation is used for item nonresponse, but in longitudinal studies, (such as the NHANES [National Health and Nutrition Examination Survey] study), there is also a special case called wave nonresponse (an individual may co-operate some of the time, but one set of measurements may be left out, for whatever reason). From the longitudinal perspective this is item nonresponse, and from the cross-Sectional perspective this is unit nonresponse. Model based methods for item nonresponse are discussed in Little and Rubin(1987) and Schafer(1997).

2.4.5 Quota Sampling:

This tends to be used in:

1 Commercial polls, i.e. the population may be systematically sampled until the required number is reached.

2 Surveys that dictate only volunteers will be used for ethical reasons.

3 Establishment surveys often have a percentage quota that must be filled. (Establishment surveys are those of businesses, or establishments, not households but often government surveys).

2.4.6 Telephone Surveys

Telephone surveys are having increasing difficulty with nonresponse, due to things such as caller identification - the caller may chose not to answer the phone - This is known as call screening. Many people also use answering machines for this purpose and so increase the nonresponse problem.

It is only a matter of time before landline telephones will no longer be the main source of contact in a household. Increasingly personal cellphones are in use, and for many this is a cheaper option than a conventional landline telephone (Finland has a 60% personal cell phone rate for adults). So these people are lost to survey sampling of this type. Of course, traditional families still have landline phones as their primary means of communication, however it is often the minority groups, which don't fit this category, which are the groups of interest in the first place, and therefore bias is introduced.

2.4.7 Call Backs for the Noncontactables

This group of people tends to be young, less than 30, who tend to move about. Follow up visits and calls (if phone numbers are available) are very useful at getting around the problem of noncontactability, using a different time of day, or different day of the week for subsequent visits. Very often a card to say the interviewer has called, is a great way around noncontactability.

Some surveys can be dealt with by self-reporting (i.e. leaving the survey for the respondent to fill out, and then to post back - having a

stamped self addressed envelope. But this does not always work, as leaving a form to post back is much more subject to potential nonresponse bias).

2.4.8 Sensitive questions.

Difficult questions (those of a sensitive nature) will produce item nonresponse. In a United States survey of pregnant women, they were actually asked questions such as:

1. What was your weight prior to pregnancy? (They may not know).

2. How many cigarettes a day did you smoke in the three months prior to pregnancy?

3. How many alcohol binges did you have in the three-month period prior to pregnancy?

4. Was your pregnancy intended?

5. What is the frequency in the three-month period prior to pregnancy of physical abuse?

In that study (Hughes et al, 1999) all these questions induced item nonresponse of at least 5% over and above that of all the other variables. Even the usual income question was more likely to be answered than these sensitive questions. (Perhaps this is an indicator that women may not object to this question quite as much). Demographic variables produced significant results when these sensitive questions were tested for missingness. Interestingly the

physical abuse question had the highest rate of nonresponse, indicating it was perhaps the most sensitive question. However very definite patterns of nonresponse emerged, the demographic variables being very good predictors of nonresponse.

This sort of question, or rather the approaches and alternatives to them, needs to be very carefully considered and maybe a less offensive way of asking the questions can be developed.

2.4.9 Coercion

Factors affecting non-response may well be external, as in the case of the Good Friday exit poll for the referendum in Northern Ireland. A large proportion of the predominantly Protestant areas refused to participate because the Rev. Iain Paisley told them not to! A good many people did not want to respond to this particular poll (run by a commercial polling organisation commissioned by the BBC), because they were apprehensive about so doing. There were difficulties for interviewers too, particularly in this type of poll. Two of the interviewers were abducted when doing street sampling - until the polling organisation was contacted and asked who they were and who was sponsoring the research (in this case the BBC). When it was proven to be 'non-governmental' the interviewers were released. Since then this organisation has only used local people with VERY local dialects in places like Belfast. The interesting generalisation from this here is that collecting data will usually be more successful if it is done by someone locals perceive to be part of 'their community' as opposed to someone 'imported in' to do the job. Speaking the right

'street speak' to the respondents is very important to put the respondent at ease.

2.4.10 Methods of Interviewing

An interesting survey result relates to the case of sampling more than one person in a household but asking all the questions of one respondent. It concerns to topic based questioning. Topic based questioning is where all the questions relating to a topic - age, income or employment - were asked together, about all individuals in the household. Person based surveying, on the other hand asks all questions about a person and then moves on to the next person in the household. The topic based approach was faster and had a lower overall nonresponse rate. However the person based approach was slower and had a higher nonresponse rate overall. The income question actually had a lower item nonresponse rate with the person approach. Why is this? Possibly it is because the sensitive question (in this case income) has been diffused through the whole interview, not concentrated in one lump causing major offence and 'interviewee fatigue'. This supports the call for randomising the order of questions. (Loomis, 1999; Borgers, Hox,1999) (see Chapter 2.6)

2.4.11 Incentives

These are used to encourage people to respond:

1. Money, gift certificates or consumer goods reward the time and effort responding.

2. Lottery type (a chance to win a ... with each returned survey) do not show particularly promising results.

However instant gratification can sometimes work. For 'street sampling' within a university main thoroughfare, the use of a chocolate fish can induce a few people to be respondents - this can work on a young age group (18 to 25 year olds).

2.4.12 Double Sampling

Double sampling: (Cochran, 1977)This is used in sample surveys, where some variables are recorded for all units in a sample, and then additional variables are recorded for some of the cases, but not all. If the variables being recorded depends on the original recorded, then this is MAR. Sampling for Nonresponse follow up: Follow up is used in Censuses, and large surveys, where unit nonresponse is present, due to non contactability in face to face situations; or in the case of mail surveys, failure to mail back forms. Where continuing follow up is economically unfeasible, it may be possible to sample the nonresponding units, and therefore call the remaining missing units MAR.

Matrix sampling for test items: Here a questionnaire is divided up into sections, and the sections are randomly administered to the cases. The resulting missing data will be MAR (but more correctly missing by design).

2.5 Special Types of Data

Data is usually continuous, categorical or mixed. Other designs requiring special calculations include some types of experimental design, longitudinal data, and case control studies.

2.5.1 Experimental design

Missing data occurs in experimental design for two reasons

1 Intentionally missing by design, and therefore unbalanced.

Balanced designs are preferable to unbalanced designs, as they are easier to analyse than unbalanced designs. If balance is not possible, then the data has to be analysed as unbalanced. This can be remedied by saying that if it were a balanced design, it would have X missing cases. Although this approach is not recommended, it occurs in say medical screening where one group may be given a test, but it is not given to all cases. This is very similar to double sampling, (Cochran, 1977) and is MAR.

2 Missing data is unintentionally missing.

This occurs, for example when the cows got in and ate the plants in one of the split plots in an agricultural experiment, so part of the experimental design is missing; or the river flooded and washed away the crops in a particular field. These are nonrandomly missing. Although the event itself is random, the fact that there is a pattern or bunching in the data means that is classified as nonrandomly missing. With biological studies if the technician accidentally drops the tray of samples collected that morning, this is MAR. Another problem is losing a vital part of the experiment, such as the only treatment-block combination in the experiment.

Correcting the Sums of Squares in the analysis phase -The usual Analysis of variance (ANOVA) used for tests of significance in experimental design, requires a complete balanced design. (Two way ANOVA). When there is missing data, balance no longer holds. In SAS the sums of squares are corrected for by using Type III Sums of squares. (This is now available in S-PLUS also.)

Sums of Squares in ANCOVA

Bartlett (1937) suggested that the missing values (in the covariates) be filled in with estimated values (guesses). The ANCOVA then proceeds with an indicator for each missing value. The fitted value is then calculated. This approach is favoured because it is not an iterative method. Little and Rubin (1987) provide a proof that this method leads to correct least squares estimates adjusted for the missing value covariates.

2.5.2 Case Control Studies

Case Control studies differ from the usual data analysis in that the observations are paired to another observation (the case control), and if the observation of interest is missing and case deletion is used, then the 'pair', (the control) is also deleted, even though it is observed. This increases the probability of bias, since the type of missingness is nearly always not MCAR. When this type of study is carried out - usually in clinical trials, then results such as relative risk estimates will not be valid. If the variable is continuous then a simple method is to replace the missing value with two extreme values, in which interval the true value lies. If the variable is categorical, then a missing value indicator can be used, with the relative risk for this giving an indicator of bias. Vach and Blettner (1991) stated in their paper that these simple methods always will give biased results.

2.6 Ways to prevent Nonresponse

If the problem is item nonresponse the way the questions are worded may well be the answer to this problem. When surveys are properly pretested and piloted on a group similar to the target respondent group, item nonresponse would show up in the early pretesting pilot stage. Often through time or financial constraints these stages are dispensed with, only to provide grief later.

Questionnaire design is also important. It has been shown in overseas studies that randomising the position of the item on the questionnaire can be a very good way of reducing item nonresponse.

For example, interviewee fatigue at being asked questions on a particular topic, often repeating the same question under a different disguise, can have the effect of inducing item nonresponse. Many researchers realise that a question such as personal income is sensitive, and so to get around this they ask 'proxy' questions which give good indicators for the sensitive question. Often at this stage the respondent will tell the interviewer to 'Get lost!' It is never a good idea to annoy or badger the respondent in this way. The most successful surveys are those which are short, to the point and inoffensive. Any repetition of questions is unwise. So to get good results the first step has to be good survey design. Time spent at this stage will be rewarded later, many times over.

3 Literature Review of Methodology for Analysing Missing Data

3.1 Cure for Missing data

Essentially there are four methods available

- Complete available case
- Imputation
- Reweighting
- Model based methods

Ways of dealing with missing data are described in the following sections.

3.1.1 Complete and Available Case Analysis

1 Use only Complete Data observations. - Anything else is excluded from the analysis. This only works if the assumption MCAR is true. However this is not necessarily optimal as it is wasteful of the data, which may be expensive to collect.

2 Case Deletion - Where a dependent variable is not observed, the case is deleted from the analysis. This is the default method for most of the major statistical packages.

3 Imputation methods (See chapter 5). If SPSS Pairwise is used, this replaces missing values in Y, by imputing data based on all available valid observations in X. This means that if a particular combination of variables is available, in a particular model, these will not be excluded from that model, even though they may have other missing values in that case. This is also known as available

case analysis. This is catered for in BMDP's CORPAIR, COVPAIR, or Allvalue options. These however are based on correlations between X and Y, and assume that the correlation with other variables are zero. As this is nearly always not the case, computational difficulties can arise; i.e. singular matrices etc.

3.1.2 Imputation (see chapter 5, for a more detailed description of methods used)

Older, more known methods include:

Replacement (Imputation) of data

1 Case Substitution- where observations with missing values are replaced by another non-sampled unit. e.g. In a survey, replacing one household with another, hopefully similar household - this is most widely used in commercial surveys, ultimately providing complete case data.

2 Mean Substitution - This is an often used method, where the missing value is replaced by the mean of that variable, based on the other responses, so the valid responses in the sample provide the basis for the calculation of the imputed values. This has a disadvantage of reducing the variance for that variable, and also reducing the correlation observed with the other variables, because all missing values will then have one single observation. It could also distort the observed distribution for that variable.

3 Hot Deck Imputation - This is where a value is input into the missing value, by means of selecting a value from similarly corresponding units. This is fairly commonly practised; obviously it

doesn't have the variance reducing problems which mean substitution has.

4 Cold Deck Imputation - this replaces the missing value by means of an external source (a previous survey, say). There is no real justification for this method, unless the researcher is satisfied that the externally imputed value is more likely to be correct than the internally computed mean.

5 Regression Imputation - This is used to predict the missing values based on the other variables, and the fitted values are put into the model.

3.1.3 Reweighting

Weighting is a natural idea that is an extension of stratified sampling.

This is primarily used when unit nonresponse is present, and what it does is to say a sampling unit π_i represents π_i^{-1} units of the population. If n_j units chosen from N_j units in stratum J, then $\pi_i = \frac{n_j}{N_j}$ for units i in stratum j. In this way each sampling unit becomes a stratum within the population (Little and Rubin,1987). If these strata fit into larger (coarser) strata groups, it is possible to assign the missing units into strata containing observed units, and weight these observed units according their known occurrences in the population. Here the unit nonresponse is assumed to be a form of random subsampling. In practice this works when the missingness mechanism is MAR, but not MCAR, as the observed at random part

of MCAR does not hold under the weighting procedure. Within strata the MAR condition will hold however.

3.1.4 Model Based Methods

1. Likelihood

One advantage of likelihood based methods is that essentially there is no difference whether the data is missing or observed. Of course there needs to be observed data to estimate the model, but given the model chosen, say multivariate normal, the Maximum Likelihood of the estimate(s) are calculated from the available data (or equivalently the log-likelihood). Finding the standard errors is more of a problem, as any estimate of the variance needs to reflect the uncertainty due to the missing values. If an indicator (W) for missingness is included in the model, then the model is specified over the joint distribution of W and Y (Y being the data matrix). W is treated as a random variable. The product of the conditional distribution of W given Y, and that of Y is (Little and Rubin,1987)

$$f(Y,W/\theta,\psi) = f(y|\theta)\ f(W|Y,\psi)$$

If the missing (Y_{mis}) are integrated out, then the distribution of the observed data is obtained, so

$$f(Y_{obs},W|\theta,\psi) = \int f(Y_{obs},Y_{mis}|\theta)\ f(W|Y_{obs},Y_{mis},\psi)\ dY_{mis}$$

The likelihood of θ and ψ is proportional to a function of θ and ψ as in the above equation. So the likelihood is

$$L(\theta,\psi|Y_{obs},W) \propto f(Y_{obs},W|\theta,\psi)$$

When the missing data does not depend on actual values of Y, that is if:

$$f(Y_{obs},Y_{mis}|\psi) = f(W|Y_{obs},\psi)$$

then:

$$f(Y_{obs},W,\theta,\psi) = f(W|Y_{obs},\psi) \times \int f(Y_{obs},Y_{mis},\theta) dY_{mis}$$
$$= f(W|Y_{obs},\psi)\ f(Y_{obs}|\theta).$$

This is the condition MAR.

That is the parameters are distinct. The likelihood's are proportional. Therefore likelihood based inferences only require MAR, not the more stringent MCAR condition. (Little and Rubin,1987)

2. Bayesian

An alternative to the Likelihood approach is the Bayesian approach. When the data are complete, applying a (class of) prior distributions, which are conjugate to the likelihood function, gives an appropriate posterior distribution, from the same class from which Bayesian inferences may be drawn. A conjugate class means that any prior $\pi(\theta)$ in that class gives a posterior $P(\theta|Y) \propto \pi(\theta)L(\theta|Y)$ that is in the same class. Some common ones are: (De Groot, 1970)

A When both μ and Σ are unknown, the conjugate class for the multivariate normal distribution is the normal inverted-Wishart distribution.

B When the data is categorical, the conjugate class for the multinomial distribution is the Dirichlet distribution.

However, when some of the data are not observed, then a prior is chosen to reflect the knowledge about the distributed data. Further, what is appropriate for complete data will also be appropriate for incomplete data. In practice for missing data problems, if little is known about the parameter(s) of the distribution, a diffuse prior is used, which is suitable for the complete data, and which will also be suitable for missing data problems. When a diffuse prior is used, the Bayesian and frequentist inferences for missing data usually agree, which is rather reassuring.

Rubin (1987) shows that when both the sampling and the response mechanisms of the posterior distribution of Y_n (and thus of Y_{mis} and Q) [Q being the scalar denoting the posterior distribution of $Q = \overline{Y}$] can be obtained from the observed values and the specification for the Pr(X,Y):

$$Pr\ (Y_n|X,\ Y_{obs},\ W_{inc},\ I) = Pr\ (Y_n|X,\ Y_{obs})$$

$$= Pr\ (X,\ Y)/\int Pr\ (X,\ Y)\ d\ Y_n$$

Where W_{inc} is an indicator for missingness (the response indicator),

I is the indicator for inclusion into the survey (the sampling indicator),

X is a covariate,

Y_n is the number of observed responses,

Y_{obs} is the observed response.

The logarithm of the observed -data posterior is

$$\log P\ (\theta|Y_{obs}) = l\ (\theta|Y_{obs}) + \log \pi(\theta).$$

The completed data posterior distribution = the complete data posterior distribution × an adjustment factor.

3.2 *Older Methods used an 'ad hoc approach': Early Literature on Missing Observations*

A review by Afifi and Elashoff (1966) shows that in the early literature, complete case analysis was used, and mean imputation, with regression imputation the most complicated method (of which mean imputation is a special case.) The 1970's saw mainly case deletion and single imputation. The 1980's gave likelihood based imputation, and the EM algorithm. Prior to this missing data was something to be gotten rid of, but now it was seen as something to be averaged over. Paulin et al (1996) showed case deletion studies are biased, and recommended regression imputation.

Ibrahim (1990) showed that imputation may be achieved via the EM algorithm by weighting, and expressing the E-step as a weighted complete data log-likelihood; when the unobserved covariates come from a discrete finite domain, then the M-step, was achieved by MLE. The 1990's saw the introduction of Multiple Imputation, Markov Chain Monte Carlo methods, and other Bayesian methods. Azen, Van Guilder and Hill (1989) give a comparison of early methods.

3.2.1 Performance of Different Methods:

In this section the performance of different methods of dealing with missing data in regression analysis is considered:

1: Complete case analysis; no missing values.

2: The Estimation-Maximisation (EM) Algorithm; as described above, EM Algorithm.

3: The Allvalue Imputation method.
Allvalue Imputation - is based on a BMDP program 8D, which essentially uses pairwise deletion (described earlier), imputing means calculated from complete cases. This works when the numbers of missing values are small, and appeared to be similar to complete case results. The EM Algorithm performs best, with censored and non random data (Azen, Van Guilder, Hill, 1989).

To use Maximum Likelihood for the inference of regression models, the distribution of the covariates needs to be known. If the response is categorical, then loglinear models can be used to estimate the variable values. Usually the MCAR condition is assumed, and the cases with missing values are deleted from the analysis (the imputation model). ML can be used to test the estimates for violations of the MCAR principle (see chapter 3.4). It is sufficient to assume MAR if the probability of the occurrence of missing values does not depend on the outcome variable, but does depend on those covariates that are completely observable.

Vach and Blettner show the important result that creating an additional category for the missing values always yields biased results (Vach and Blettner 1991). The Mantel-Haenszel estimator shows this. The 'additional category' approach is often used in Epidemiological studies.

The reason for the missing data needs to be known if the correct method is to be applied: that is to decide whether to use imputed data, or complete case data. Complete case data is only unbiased if the missingness is MCAR. This can be tested by comparing the odds ratio for the category of missing values, with that of the observed values. Vach and Blettner (1995), describe the missingness odds ratio used in categorical analysis as:

$$MOR_{y,\,x(a)\,(k)} = \frac{m(y, x_a, k)}{1 - m(y, x_a, k)} \div \frac{m(y, x_a, 1)}{1 - m(y, x_a, 1)}$$

where $m(y,x_a,k)$ is the conditional probability to observe a missing value, in category k.

3.3 More Modern Methods

More modern methods of single imputation include Hotdecking, Colddecking, and Substitution, Last value carried forward, Mean Imputation, Regression Imputation with and without stochastic error, and so on, (see chapter 5).

Some of the newer methods of dealing with Item Nonresponse by single imputation are imputations based on Box-Cox transformations. These allow for transformations of the imputed values. A criticism of these sort of regression imputations is that it is essentially mean imputation, which is then backtransformed, also that the imputed data sets are then analysed as if they are completely observed, and this does not lead to accurate variance estimates. This method is better suited to continuous data. The previous way of dealing with this was to add a random (stochastic) component, to try and preserve the variance estimate.

3.3.1 Imputation using Box-Cox Transformations

Cassel et al (1999) proposed a method of estimating the variance associated with this imputation. They propose that when an imputed variable is transformed (to do a calculation), the retransformation will introduce bias, as the retransformation will tend to estimate the median response rather than the mean response. It is possible to adjust for this bias (Sakia,1990). For a semi-log transformation, the imputed values are multiplied by $\exp(\sigma^2/2)$. For the Box-Cox transformation a suitable estimate is $(1+ (\sigma^2/2) * (1-\lambda) / y^{\lambda})$ where $y = (1+\lambda*(\beta_0+\beta_1x_1+\beta_2x_2+....))^{1/\lambda}$, where these parameters are estimated by ML according to the Box-Cox transformation approach. (When $\lambda=1$ this reduces to simple linear regression). In

Cassel et al's (1999) paper a geometric mean is considered. Cassel et al (1999) propose the use of 'prediction error estimates' which are $(1/n)\ \Sigma_n$ (observation-prediction)^2 for the n observations, and if weights are used, this becomes $(1/\Sigma_n w)\ \Sigma_n w$(observation-prediction)^2. As using regression to impute leads to lower variance estimates than using just true observations, this can lead to incorrect variance estimates. An estimator which accounts for imputation is needed. Sarndal (1992) suggested that using imputed values actually increases the variance estimator. The increase in variance can be estimated, as can the total variance, and therefore the increase in variability is due to imputation. The estimated variances are as follows $V_{tot} = V_{sam} + V_{imp}$ (The component of the variance due to imputation) and the sampling variance estimates can be further broken down into $V_{sam} = V_{*p} + V_{diff}$. The estimate V_{*p} is the variance observed by traditional estimates. $\hat{V}_{diff}$ is the estimated difference between V_{sam} and V_{*p}.

For SRS the estimate $\hat{V}_{diff} = (1/n(n-1))\left\{\sum_1^n (y_k - \bar{y})^2 - \sum_1^n (y_{*k} - \bar{y}_*)^2\right\}$

where $\bar{y}_* = \left\{w_r \bar{y}_r + (1 - w_r)\bar{y}_{*(s-r)}\right\}$

$$\bar{y} = \left\{w_r \bar{y}_r + (1 - w_r)\bar{y}_{(s-r)}\right\}$$

where w_r = the response rate

y_r = responding units

y_{*s-r} = nonresponding units

$\bar{y}$ = imputed data

$\bar{y}_*$ = available (observed case only)

which gives an estimate of

$$V_\xi\left(\hat{\bar{y}}_{(s-r)}\right) = \frac{1}{(n-r)^2}\sum_1^{n-r}\left(\lambda x_k \hat{\beta} + 1\right)^2 \left(\frac{1-\hat{\lambda}}{\hat{\lambda}}\right)\hat{\sigma}^2$$

Where ξ shows the auxiliary information model is used, this involves the variance estimate of $V_{\xi}(Y_k) = \sigma_k^2$ (by estimating from the response set, it is possible to predict the y-values for the nonresponse set). The parameters λ, β, σ are all estimated when the Box-Cox model is estimated.

3.3.2 More on Regression Imputation

Knaub (1999) has used regression imputation to predict 'missing values' (i.e. impute a fitted value) using standard software - SAS Proc Reg. Using this software different predictor variables may be chosen, and different weights for imputation. Here data can be collected under one set of categories, and published under a completely different set. [The writer finds this methodology to be very suspect, as it could be misleading to the end-user]. However for this to work, predictions should be made using as large as possible a set of homogeneous data. (Homogeneous data is where the same model with the same coefficients would be appropriate for all members of a given subgroup in the data).

Assuming independence, variance estimates for each strata (homogeneous sub-group) are then added to give an overall estimate of the variance. This method is very restrictive in that it relies on the homogeneity of each sub-group. If this assumption were to not hold true, (as in most sampling surveys) then the variance would tend to be very much under-represented. This method is designed for Establishment surveys (Establishment surveys are those of Businesses, and commercial establishments, rather than individuals or households) as opposed to household surveys (which in themselves are generally very skewed) and

possibly may be stratified in this way. This method is not very generalisable.

3.3.3 Imputation using Coarsening, or Discretising Data

Another method used is by Heeringa, Little, and Raghunathan (1999) to impute income data (net worth). Here the problem of nonresponse is tackled at the data collection phase by asking the respondents who are unwilling or unable to give actual figures of net worth to provide data in 'coarsened bands'. So the resulting data set is a mixture of 'coarsened' data (interval censored responses) and exact observed data, and missing values. This is then analysed using a mixed normal model for the 'coarsened data' using a Bayes multivariate, multiple imputation approach, based on this normal mixed model. The continuous variables may be skewed and require transformations.

Two comparative methods used here with good results are:

1 Sequential Regression Method (IVEWARE). This is SAS callable and is a general purpose imputation algorithm for mixed categorical and continuous variables (also used for mixture modelling). This program multiply imputes from a sequence of regressions (linear-continuous, Logistic-binary, multinomial-logistic, and Poisson-count) and uses Bayesian methods when categorical variables are more observed than continuous. The method can incorporate restrictions and bracketing of imputed values.

2 Gibbs Sampler algorithm (Heeringa, et al 1999) This yields draws from the posterior predictive distribution of coarsened data, given the observed data. It conditions on the observed data, and multiply imputes from the predictive distribution. Heeringa et al (1999) compared the results of the above methods to different imputation methods. Complete case analysis is found to underestimate the distribution (particularly the variance) relative to the other methods. Mean/median substitution has the same problem. Hot-deck introduces shrinkage towards the mean (underestimating of the variance). Bayes methods and sequential regression yield similar results, in that the estimates using these methods are similar.The distribution estimates are shrunken due to non-ignorable missing data. Only the Gibbs (Bayes) method and the IVEWARE method appear to give valid (in this case - the income example) estimates (fortunately they agree with each other). In the income example, as the net worth increases, so does the likelihood of using intervals as opposed to observed data. The methods appear to go some way to addressing nonignorable missing data.

3.3.4 Multiple Imputation

The efficiency of complete case analysis is greatly reduced in a lot of applied research situations, when there are a large number of variables and not a large number of cases (Little and Rubin, 1987). If the data are MCAR, there is a loss of efficiency, and if MAR the analysis will also be biased. When there are a large number of variables, and a smaller number of cases, even a small number of missing observations in each variable can result in a high

percentage of incomplete cases. (If 10% are randomly missing where there are 20 variables and 100 cases, then the probability of one case 'surviving' is $0.9^{20} = 0.12$, or on average only 12% of the cases would be complete). In reality it is usually not quite that bad, but it is not uncommon to lose 50% of data in this way. Multiple Imputation (MI) provides an alternative that represents the uncertainty due to missingness.

If the data is MAR, and distinctness holds, and the Imputations are 'proper' (Rubin 1987) then the missing data mechanism is ignorable (Rubin 1987). If all available information is included, then differences of estimates between completely and partially observed cases are reduced, including covariate information, meaning the differences are less if all available information is used (Meng 1994, Rubin 1996). If the sample size is small, parsimony is a problem and the model may well be over parameterised. Schafer (1997) suggested a ridge prior distribution to handle multivariate normal data. (A Ridge prior is a normal limiting Wishart prior-generalised k dimensional χ^2 distribution, sum of squares of normal variates). The procedure is analogous to Ridge Regression (Evans, Hastings and Peacock, 1993).

Factor analysis may be used to reduce the number of parameters, for the covariates. Factor Analysis works by ignoring factors corresponding to small eigenvalues, thus reducing the number of parameters in the imputation model. The assumption is that the number of factors is unknown, and factor analysis provides a means of including the variables without overparameterisation. (However if the imputation model is underparameterised, serious bias can occur).

Rubin's 3-step process for Multiple Imputation is as follows:

Step1: Create m completed data sets, m>1, by imputing the unobserved data m times using m independent draws from a stochastic imputation model (or m draws from a posterior distribution).

Step 2: The m complete data sets are analysed conventionally (for example regression models computed and so on) - treating the data sets as complete data sets (using standard software).

Step 3: The results from the m complete data sets are combined into one set of results.
The problems of non-compliance, censoring or attrition are the most visible form of missing data, but the problem of bias between observed and unobserved data is invisible, and potentially much more serious. If the observed data represent a biased sample of what is intended, then the results are biased, and any inferences taken from these are incorrect. The methods for analysing the data can be as computationally sophisticated as they like, but if they are based on a biased sample, they may be of very little value. The problem with available case analysis (case deletion) is that it is essentially invisible. A lot of investigators do not even realise that there is a problem, and how serious non-response bias is (Little and Rubin 1987).

Imputation is used to get around the problems of censored data, attrition and non-compliance (item nonresponse). This can produce results that are reasonable as long as two criteria are met:

1 The analysis needs to take into account the uncertainty of imputed values, because they are still not real observations, no matter how sophisticated the imputation technique.

2 The imputation method must include the distribution relationships between unobserved and observed (this needs to take into account the missingness mechanism). Combining the m data sets (the most straightforward part of MI). The estimate is

$$\overline{\theta}_m = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$$

and its variance is $\mathrm{Var}(\overline{\theta}_m) = T_m = \overline{U}_m + (1+\frac{1}{m})B_m$,

where T_m is the variance of $\overline{\theta}_m$ (the total variance)

$\overline{U}_m$ = the average of $\{U_i ; i=1.....m\}$, where U_i is the variance of $\hat{\theta}_i$ within an individual imputation, and $B_m = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \overline{\theta}_m)(\hat{\theta}_i - \overline{\theta}_m)^T$ estimates the between imputation variability. An inflation factor $(1+m^{-1})$ accounts for the additional variability for using 'm' imputations, as opposed to an infinite number (Rubin, 1987, page 76).

For constructing p-values and confidence intervals, the normal approximation does not hold, unless the sample size n is large, and the number of imputations m is also large. For small m (as this is nearly always the case [$5 \leq m \leq 10$]) other methods are used which take into account the degrees of freedom (Barnard and Meng, 1999; Rubin and Schenker, 1991)

A good imputation model will combine all available data, and include what is known about the missing data mechanism. However it is important to keep the model fitting reasonable. There is a 'trade off' between simplicity and accuracy. Something oversimplified such as mean imputation is clearly inadequate. (Mean imputation includes within it the same biases they are supposed to reduce). However if the model is overfitted, then the power of prediction is

poor, and in practice is less acceptable and more prone to error: so a more parsimonious model is preferable.

3.3.5 Uncongenial sources of input.

Another principle is that the imputation model should not be in conflict with the model for analysing the data (Meng, 1994). When an analysis procedure does not correspond to the imputation procedure, this is known as uncongenial. It occurs when the analyst and imputer use different sources and amounts of information. If the imputer's assessment is incorrect then all the methods of handling nonresponse will be incorrect (Rubin, 1996) If however the imputer's model is accurate, then serious nonresponse bias is avoided for the analyst. Congeniality is thus the agreement between imputation model and the analysis. This can be a problem in the case of 'Public Use' data files (Imputed first and analysed later by others).

There are three Cases of Uncongeniality:

1 When the imputation model is unknown to the person analysing the data.

2 There are different purposes for imputing missing values and for analyses, as different models serve their needs (this can occur when the aims of the analyst differ from those of the imputer).

3 Several models are considered for imputation or analysis.
In uncongeniality, the imputer can assume more than the analyst, or the analyst may assume more than the imputer.

A frequentist view of all of this is that:

Disagreement between repeated imputation analysis (usually Bayesian) and the best possible incomplete data analysis does not invalidate the repeated imputation inference. It can in fact show up nonresponse bias with the complete case analysis especially if the analyst has used less information than the imputer (Meng, 1994; Rubin, 1987). This is a good indicator that the missingness mechanism is not MCAR.

Proper Imputation (second moment proper) is based on conditional unbiasedness of three quantities

1. [conditional unbiasedness with respect to conditional randomisation of the distribution of the non-response indicator W given the sampling indicator I, (and
2. X and
3. Y)];

described in chapter 3.1.4. The following gives Rubin's rule for proper imputation. It is proper (for the set of complete-data statistics {Q, U} (Q is the mean, and U is the variance of the posterior distribution) if the following three conditions are satisfied (Rubin, 1987; Little and Rubin, 1987; Meng 1994; see chapter 3.7).

Based on the infinitely many imputations $m \to \infty$, let $\overline{Q}_\infty = \lim_{m\to\infty} Q_m$, let $\overline{U}_\infty = \lim_{m\to\infty} U_m$, and let $\overline{B}_\infty = \lim_{m\to\infty} B_m$, and $\cong$ is approximately equal to; the statistics $(\overline{Q}_\infty, \overline{U}_\infty, \overline{B}_\infty)$ yield valid inferences for the complete-data statistics $\hat{Q}$ and U; under the response mechanism, where U= the total variance, and B = the between imputation

component of the variance. Here Z_c denotes the complete imputed data, and Z_0 is the complete observed data.

1 $\overline{Q}_\infty$ and $\hat{Q}(Z_c)$ have the same expectation

$$E\left[\overline{Q}_\infty | X, Y\right] \cong E\left[\hat{Q}_\infty (Z_c | X, Y)\right];$$

2 $\overline{U}_\infty$ estimates the variance of $\hat{Q}(Z_c)$

$$E\left[\overline{U}_\infty | X, Y\right] \cong V\left[\hat{Q}(Z_c) | X, Y\right];$$

3 B_∞ estimates the variance of $\overline{Q}_\infty$ - $\hat{Q}(Z_c)$

$$E\left[B_\infty | X, Y\right] \cong E\left[\left(\overline{Q}_\infty - \hat{Q}(Z_c)\right)^2 | X, Y\right]$$

- Only with respect to the analysts complete data procedure;
Congeniality is with respect to complete data and incomplete data procedures (Meng, 1994).

An Imputation model g is said to be better (than the analysts congenial imputation model) for $\hat{Q}(Z_c)$ if:

$$E\left[\left(\overline{Q}_\infty - \hat{Q}(Z_c)\right)^2 | X, Y\right] \leq E\left[\left(\overline{Q}(Z_0) - \hat{Q}(Z_c)\right)^2 | X, Y\right]$$

This compares the repeated imputation estimator with the incomplete-data estimator $\hat{Q}(Z_0)$._A further comparison is when the imputer does a better imputation job, than the secondary analyst can do (Rubin, 1987). This commonly occurs as the imputer often has a greater knowledge of the missingness mechanism (Meng, 1994).

Uncongeniality can occur when the imputer has used one model for imputation, but the analyst would have preferred it to have been imputed another way, more in line with the model the analyst would be using.

If the following conditions hold:

1 The analyst's complete-data estimator $\hat{Q}(Z_c)$ is self -efficient. (Self efficiency is when an estimation procedure for Q is such that there is no $\lambda \in (-\infty,\infty)$ for which the mean squared error of $\lambda \hat{Q}(Z_0) + (1-\lambda)\hat{Q}(Z_c)$ is less than that of $\hat{Q}(Z_c)$].

2 The Imputer's model is information regular for estimating Q using $\hat{Q}(Z_c)$ (Information regular is using the extra information on efficiency available, to the imputer. Meng, 1994)

3 The Imputer's model is second - moment proper with respect to the analysts complete data procedure.

4 The Imputer's model is better for $\hat{Q}(Z_c)$

Then the following hold:

A The repeated-imputation estimator is consistent for Q, and is at least as efficient as the analyst's incomplete-data estimator.

B For any nominal level, the corresponding repeated imputation confidence interval has at least the nominal coverage, but at most the same width as the confidence interval from the analyst's incomplete-data procedure with the same nominal coverage (Meng, 1994).

3.3.6 EM Based, MCMC Based Methods

Software Comparison for handling missing data

A comparison of more recently available software for handling missing data is discussed in the paper by Wiggins et.al. (1999) In particular Norm (Schafer, 1997) is compared to the SPSS module Missing Value Analysis. (SPSS Inc. 1997)

Regression analysis may be used to assess the impact of ignoring missing data. For imputation the likelihood-based estimation procedures in SPSS and Norm are used. SPSS uses the EM algorithm, for multivariate normal data (Dempster, Laird and Rubin 1977) and imputes missing values. (SPSS also imputes using multiple linear regression with prescribed predictors of missing values). Norm uses multiple imputation (Rubin, 1987; Schafer, 1997) to generate m complete data sets (m is often equal to 5). These are analysed separately, and then combined using Rubin's Rule (1987) for scalar estimates (see chapter 3.9.1). Likelihood estimation procedures in SPSS and Norm assume that the data are missing at random. Is this reasonable? MAR is less restrictive than MCAR. The test suggested by Rubin is to take an indicator function and then to test that the distribution of missingness does not depend on the missing value? A simple Anova, or a logistic regression using the missingness indicator as an outcome, or even a two sample t-test will achieve this, for continuous data. (An indicator function is used to describe the two groups, say 1 = Y_{mis}, and 0 = Y_{obs}. Then a two-sample t-test is done on the continuous variable of interest, using the two groups defined). A two by two table can achieve this for binary data.

If it doesn't (depend on the missing value; that is, there are no differences between the two groups) then it is said to be ignorable. Ignorable is defined by Schafer as 'not that the propensity to respond is completely unrelated to the missing data, but this relationship can be explained by data that are observed'. The actual

patterns of missingness are explained within the observed data. To get a barometer of efficiency, an indicator for the degree of imputation for any case can be used.

For both MI and EM imputation, multivariate normality is assumed. Perhaps this really is not valid - particularly in the case of categorical variables such as gender. An interesting exercise would be to simulate complete data, then delete to form particular missingness patterns. This could be tested against the complete data in some way. Little and Rubin (1987) state that the knowledge of what led to the missing mechanisms in the first place is essential to choosing an appropriate analysis.

3.4 Little's test for MCAR

(Little,1988) developed a test for MCAR and showed that the null distribution is asymptotically χ^2. This test is too restrictive for testing whether the model is ignorable for likelihood inferences (that is, the usual MAR). However it is useful for testing the validity of simple data methods, for example whether pairwise and complete case methods are valid, as these require the MCAR assumption. Assume the data y_i is multivariate normal with mean μ, and covariate matrix Σ. Let J = the number of distinct missing data patterns in the w_i in the dataset. Fully observed cases count as a pattern as do all the set of cases with missing data pattern j (j=1,2....J). This is shown as S_j, m_j is the number of cases in S_j, and $\Sigma\, m_j = n$.

Little (1988) gives his test in three forms:

3.4.1 Σ known.

$$\text{Let } d_0^2 = \sum_{j=1}^{J} m_j (\bar{y}_{obs,j} - \mu^*_{obs,j}) \Sigma^{-1}_{obs,j} (\bar{y}_{obs,j} - \mu^*_{obs,j})^T$$

Under the null hypothesis (MCAR), conditioning on w_i,

$$(y_{obs,i} \mid w_i) \overset{\sim}{_{ind}} N(\mu_{obs,j}, \Sigma_{obs,j}) \; i \in S_j \;\; 1 \le j \le J$$

And under the alternative hypothesis,

$$(y_{obs,i} \mid w_i) \overset{\sim}{_{ind}} N(\nu_{obs,j}, \Sigma_{obs,j}) \; i \in S_j \;\; 1 \le j \le J.$$

That is, the means of the observed variables can vary across the patterns where μ^* = the MLE of μ; and ($\nu_{obs,j}$ j = 1, 2, ..., J) are mean parameters for observed variables (not $\mu_{obs,\,j}$), which are distinct for each pattern j.

3.4.2 Σ unknown.

The μ^* are replaced with $\hat{\mu}$, and Σ with $\tilde{\Sigma}$ giving :

$$d^2 = \sum_{j=1}^{J} m_j \left(\bar{y}_{obs,j} - \hat{\mu}_{obs,j}\right) \tilde{\Sigma}^{-1}_{obs,j} \left(\bar{y}_{obs,j} - \hat{\mu}_{obs,j}\right)^T.$$

3.4.3 Monotone missing

When the data are monotone missing this becomes (given here for two variables) [see chapter 3.4.4]

$$d^2 = n_2 \begin{pmatrix} \bar{y}_1 - \hat{\mu}_1 \\ \bar{y}_2 - \hat{\mu}_2 \end{pmatrix}^T \tilde{\Sigma}^{-1} \begin{pmatrix} \bar{y}_1 - \hat{\mu}_1 \\ \bar{y}_2 - \hat{\mu}_2 \end{pmatrix} + (n_1 - n_2) \left(\bar{y}_1^* - \hat{\mu}_1\right)^2 \; / \tilde{\sigma}_{11}$$

which reduces to

$$n_2 \frac{\left(\bar{y}_1 - \hat{\mu}_1\right)^2}{\tilde{\sigma}_{11}} + n_2 \frac{\left[\bar{y}_2 - \hat{\mu}_2 - \hat{\beta}_{21.1} (\bar{y}_1 - \hat{\mu}_1)\right]^2}{\tilde{\sigma}_{11}} + \frac{(n - n_2)\left(\bar{y}_1^* - \hat{\mu}_1\right)^2}{\tilde{\sigma}_{11}}$$

where $\beta_{21..1}$ is the between (distinct missingness) pattern,
Or alternatively from an ANOVA table:

$$d^2 = \frac{\left[n_2(\bar{y}_1 - \hat{\mu}_1)^2 + (n - n_2)(y_1^* - \hat{\mu}_1)^2\right]}{\tilde{\sigma}_{11}}$$

$$= SSB_1 / MST_1 = \frac{(n-1)F}{(n-2+F)}$$

Where SSB_1 is the between-groups sums of squares, MST_1 is the total mean square, and F is the F-statistic from an ANOVA table of Y_1 on the missing data pattern. In this case of two variables, the F test = t^2 , so effectively d^2 is equivalent to a t-test for comparing means. Under the null hypothesis of MCAR, (assuming normality) the test statistic F has an F distribution with 1 and n-2 degrees of freedom.

A limitation of this test is that the missingness is allowed to affect the means, but in all three forms of it, the variance and covariance patterns are required to be the same. (Perhaps this could be extended for unknown, and unequal Σ). Also normality is assumed here, but if the sample size is large, normality departures are accommodated via the asymptotic null distribution [χ^2]. Using f degrees of freedom, where f = Σp_j - p, where p = the number of values for a case in the absence of missing data, and p_j = number of observed variables for cases in S_j. A test for MCAR provides guidance as to when the standard errors based on the expected information matrix are adequate. Anything less than MCAR and they are not.

3.4.4 Monotone data patterns

One of the manipulations necessary for multiple imputation is to arrange the variables in 'missingness' order, i.e. to have the most

observed variable on the left hand side, and proceeding to the least observed on the right hand side, i.e. so they are in a monotone pattern of missingness. Just about all the software for particularly multiple imputation (but this applies to all imputation, especially if the variables are imputed sequentially) will run much better if this is the case. - Of course in the case of Joe Schafer's 'MIX' program for mixed models (i.e. both categorical and continuous, see chapter 6) the categorical variables are put on the extreme left of the data set. (Again following the same monotone pattern). Here the advantage is that multiple imputations can be simulated via single variable imputations, i.e. variable by variable, but each time using the maximum available information with each 'simulation'. Real data sets do not present themselves in this way and have to be manually sorted to produce this pattern.

Dempster, Laird and Rubin (1977) in JRSS Series B suggest that a monotone pattern is important as does Rubin (1974) JASA. However the idea originated with Anderson (1957) JASA and Wilks (1932) Annals of Mathematical Statistics, and Mariani, Olsen and Rubin(1980) - Sociological Methodology. The idea of this ordered approach is that missing values may be imputed by a series of case available regressions. This works well if the percentage of missingness is not large. An alternative is to specify one multivariate model conditionally on completely observed variables, and use likelihood techniques for analysis under the model. This can predict a distribution for missing values under a specified model, and given the observed data values, the missingness very often depends upon a covariate if MAR.

The model for imputation really needs to be consistent with the model used for analysis. As said before, Meng (1994) uses the term congeniality, or lack of it to describe the relationship between the

two models. This concept is used in Joe Schafer's software by means of a Markov Chain Monte Carlo algorithm which provides Data Augmentation. Rubin (2000) states that this method is theoretically correct, as is King's software 'Amelia', provided the specified model is correct (again Meng's congeniality), as well as the likelihood analysis. Amelia and Norm involve iterative maximisation using the EM algorithm and use draws under an asymptotic normal assumption. Convergence can be a problem, particularly with large data sets as demonstrated with the nutrition dataset used in the case studies which follow. Also this technique is used by the incompatible Gibbs sampler software, of T.E. Rathuanathan (Institute for Social Research).

Rubin (2000) states that even the crudest Multiple Imputation is likely to be an improvement on case deletion (unless MCAR), available case, LVCF methods, (the ad hoc methods) because multiple imputation allows for the uncertainty about the values of the missing data to be reflected.

3.5 Ignorable Nonresponse

Ignorability means that likelihood based methods are appropriate, and may be used without the need to model the missing data mechanism.

When does Ignorability hold?
It holds when MAR holds, and when MCAR hold, and when the property distinctness holds.

Randomness Diagnostics.

There are three diagnostic methods available:

1 Consider the two groups, those with missing data, those without (that is those with all values valid). Test to see whether other variables, such as demographics are significantly different. If yes then the missingness is not random, and therefore non-ignorable.

2 Use an Indicator variable for valid data, that is, whether the data is valid or not. For categorical data, this can be cross tabulated for each variable. This gives a measure of the association between the indicator of complete data and the other variables, and if the association is slight, then random statistical tests on the cross tabulations provide a conservative estimate of the degree of randomness.

3 Overall test of randomness, to see if the missing data is MCAR. The approach looks at the pattern of missingness on all variables, and the pattern of a random missing process. If not significantly different, then the pattern can be classified as MCAR. (see chapter 3.4) (Hair et al, 1995).

Some examples of when ignorability does not hold:

1 When MAR does not hold, for example nonrespondents in a survey (unless they can be sampled later, and this converted to MAR).

2 When a Petrie Dish breaks, with a particular culture of interest in it, or a production run fails.

3 Where data is collected, but some of the outcomes are not recorded, because some variables are not available. - This can provide a response in itself as in censoring.

MAR does not hold just because the data are missing for reasons beyond the investigators control. The missing values may be available from an external source (or a sample of them) against which to test the available (and missing) data. Ignorability is relative, as is MAR, which is dependent on the actual observed data. For a sampling survey, deciding on ignorability may involve a follow up survey for the nonresponders. Often there are covariates present in an analysis which may explain, or predict missingness although the actual Nonresponse mechanism is not known. This can also be handled as MAR, as long as these covariates are present in the analysis.

The method used by most statistical packages - 'case deletion' (omitting all incomplete cases from the analysis) can introduce bias to the analysis if the assumption MCAR is not true. If ignorability holds, procedures based on likelihood techniques (or an observed data posterior) are valid. The EM algorithm is a good example of this. Case deletion does <u>not</u> fit this criterion (ignorability) because it leads to a different likelihood (or posterior), based on complete cases. A general ignorable procedure will not discard any known data, as the observed data likelihood conditions on all available observed data. This leads to 'proper' inferences under any missing data mechanism, as the marginal distributions are known. 'The assumption made by ignorable methods is not that the propensity to respond is completely unrelated to the missing data, but that the relationship can be explained by data that are observed' (Schafer, 1997)

3.5.1 EM algorithm: what is it applied to Missing data

For the EM imputation:

Suppose there are S missing data patterns appearing in a data matrix (all possible ways of observing 1 missing value among p variables, 2 missing values among p variables, and so on up to p-1 missing values from p variables). This is needed for the E(xpectation)- step.

Suppose Y_{miss} is predicted from Y_{obs} and θ (θ being the estimated parameter(s)). The distribution $(Y_{i\ (miss)}|Y_{i\ (obs)})$ is a multivariate normal linear regression of $Y_{i\ (miss)}$ on the $Y_{i\ (obs)}$ for the missingness pattern corresponding to i (Schafer, 1997 p164-166).

Once $Y_{i\ (miss)}$ has been predicted, the M(aximisation)-step is straightforward. θ is then estimated (the maximum likelihood estimate of the parameter, θ) again until the parameter converges. The rate of convergence depends on the degree of missingness, i.e. the greater the missingness the slower the convergence.

Multiple Imputation shares the same underlying philosophy as EM, solving the incomplete data problem by repeatedly solving the complete data version (a simulation approach to missing data). Y_{miss} are replaced by m versions of Y_{miss} where m>1. These m complete data sets are analysed using conventional methods, and the results combined using Rubin's Rule (Rubin 1987)

PX-EM algorithm.

Liu, Rubin, Wu (1998) have further developed this algorithm in an attempt to "speed-up" the EM algorithm. They use a covariance adjustment to adjust the M-step, using additional information contained in the imputed complete data.

This is called the Parameter Expanded EM algorithm. Its greatest advantage is that it is stable, but converges faster than the standard EM algorithm.

3.5.2 MLE for multivariate normal

When data is arranged in a matrix of rows and columns, generally the rows represent the cases (n), and the columns represent the variables (p), giving a n×p matrix. The complete data $Y = (Y_{obs}, Y_{mis})$ where Y_{obs} = the observed portion of Y, and Y_{mis} = the missing portion of the matrix. y_{ij} are the individual elements of Y, i = 1,2,.....,n, and j = 1,2,.....,p. The ith row of Y is expressed as a column vector, $y_i = (y_{i1}, y_{i2}, y_{i3},.....,y_{ip})^T$.

Assumed is that $y_1, y_2, y_3,.....,y_n \mid \theta \sim$ iid N (μ, Σ), where $\theta = (\mu, \Sigma)$ is an unknown parameter, and Σ must be positive definite. The multivariate normal density is:

$P(y_i|\theta) = |2\pi\Sigma|^{-1/2} \exp\{-1/2(y_i-\mu)^T \Sigma^{-1} (y_i-\mu)\}$

with a complete data likelihood of :

$$L(\theta|Y) \propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right\}.$$

Thus the log-likelihood is :

$$l(\theta|Y) = -\frac{n}{2}\ln|\Sigma| - \frac{n}{2}\mu^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\sum_{i=1}^{n}y_i - \frac{1}{2}tr\Sigma^{-1}\sum_{i=1}^{n}y_iy_i^T$$

which has the maximum likelihood estimates:

$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ for the sample mean vector, and

$S = n^{-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T$, for the sample covariance matrix.

Because this is a biased estimator, more commonly $n(n-1)^{-1}S$ is used, which although not the MLE, is unbiased (Ganesalingam, 1999).

Dependent variables (response)

If the dependent variable in an analysis has missing values, and the incomplete observations do not contain information on the predictor variables, then a ML estimate using only the complete cases will be the same as that obtained by iteratively using the EM algorithm applied to all observations (Little, Rubin 1987).

Independent variables (predictor)

Usually both the independent variables and the dependent variables have missing values. If MAR, then for continuous variables, they are modelled as multivariate normal for $(Y, X_1, X_2, ..., X_P)$. The ML estimate for Y on $X_1, X_2, ...X_p$ is found using the EM algorithm for multivariate normal.

3.5.3 Contingency Tables (Categorical)

Imputation of categorical variables (using a form of hot-deck Imputation)

In large surveys where imputation is used for large data sets, categorical variables are often used for imputing the numerical or continuous variables. However if those categorical variables themselves are subject to imputation then the impact of imputing

the categorical variables becomes much greater than that of continuous variables. Missing data is usually dealt with by reweighting (for unit nonresponse) or imputation (item nonresponse), as is continuous data.

Rubin, Schafer and Schenker (1988) modelled the missing patterns, and used ignorable, and nonignorable maximum likelihood estimation methods. Hotdeck Imputation does not rely on modelling the missingness patterns and may not preserve the distribution of the categorical variables, unless specifically taken into account, and more so if the nonresponse is not MCAR.

Liu and Rancourt (1999) show that ratio-adjustment (pro-rating) can be applied to imputed data, and this is done via an iterative imputation algorithm for categorical data. They do it by imputing from a set of records placed into categories. All nonrespondents are forced into categories, then the marginal distributions are checked with the imputed distributions, and the records exchanged between categories to satisfy the constraints by the marginal totals. This preserves the distributional associations of the auxiliary variables used.

Ratio adjustment can be used as a 'refining process' to better balance the categories post-imputation to restore the original distribution of the data. (while this process cannot be used where there are empty cells, as they remain empty, it works fine when there are at least some observations in a cell).

These methods are evaluated by three criteria:

The distribution of true and imputed values, the mean of the true and imputed values, and the variance of the true and the imputed values.

One way in which the marginal distributions may be preserved with categorical data is Look-up tables. (This relies on the assumption MCAR - and for this to be fulfilled is very unlikely - and so bias is introduced).

From Liu, Rancourt (1999), a look up table can be as follows:

where W^S = total survey weight.

W^R = Sum of weights of records in S^R

S^R = Full Response part of the Survey

W^I = Categorical weight distribution Imputed

W^L = Look up other source.

$$\Delta_{X*Y*Z*} = W^{S^R}{}_{X*Y*Z*} - W^L{}_{X*Y*Z*}$$

The aim here is to minimise Δ_{X*Y*Z*}

by:

1 Insert a record into the category

2 Force a record into a category, or force two identical records into two different categories and split their weights.

3 Shift imputed values from one category to another

4 Share two categories by replicating and splitting weights

where Shift = [find a new donor in another category]

(',' = where)

Adjusting between categories involves a 'nearest neighbour' approach. The 4 step categorical imputation algorithm is based on minimising the difference between the auxiliary totals (predicted or from outside source), and the imputed totals.

Table 3.1. Construction of a look-up table:

Name	Look-up table	Characteristics
Shrink	$\frac{W^S}{W^R} W^R{}_{X^*Y^*Z^*}$	Needs MCAR Make all categories similar
Equally likely	$W'_{X^*Y^*Z^*} = \sum_{ji \in X^*Y^*Z^*} w_{ji}, w_{ji} = \frac{w_i}{\sum_{k \in j} 1}$	Every Imputable category treated with same weight
Proportional	$W'_{X^*Y^*Z^*} = \sum_{ji \in X^*Y^*Z^*} w_{ji},$ $w_{ji} = w_i \cdot \frac{W^R_{X^*Y_j^*Z^*}}{\sum_{k \in j} W^R_{X^*Y_k^*.....Z^*}}$	Larger Imputable category more important
Other	$W^L{}_{X^*Y^*Z^*}$	Lookup table built from other source

3.5.4 MLE for Multinomial Model

We next consider MLE for various models, along with their extension to allow for missing data. For maximum likelihood estimation, the likelihood function for the multinomial parameter is:

$$L(\theta|Y) \propto \prod_{d=1}^{D} \theta_d^{x_d} I_{\Theta}(\theta) \ ,$$

where I_θ is an indicator function equal to 1 if $\theta \in \Theta$, and 0 otherwise. The log-likelihood is:

$$l(\theta|Y) = \sum_{d=1}^{D} x_d \ln \theta_d \ , \ \theta \in \Theta$$

with the expected value for the cell probabilities is:

$$\hat{\theta}_d = \frac{x_d}{n} \ ,$$

Let d = 1,.....,D., the index for the cells of the contingency table, and x_d = the number of sample units which fall into cell d.
(Agresti, 1990)

3.5.5 MLE for Loglinear Model

The log-likelihood for the saturated model is;

$$l(\theta|y) = \sum_{ijk} y_{ijk} \ln \theta_{ijk}$$

Generally ML estimates are difficult to obtain for some log-linear models, so a numerical method, iterative proportional fitting, is used (Schafer,1997).

3.5.6 Longitudinal

Longitudinal Ml estimates.

For the complete data case, the usual linear growth curve model is given as :

$Y_i = X_i \alpha + e_i$,

where e_i is assumed to be $N(0,\Sigma)$, α is the vector of parameters, and Σ is a $T \times T$ positive definite covariance matrix. Given that the MLE of Σ is $\hat{\Sigma}$, the MLE for α is

$$\hat{\alpha} = \left[\sum_{i=1}^{N} \mathbf{X}_i^T \hat{\Sigma}^{-1} \mathbf{X}_i \right]^{-1} \sum_{i=1}^{N} \mathbf{X}_i^T \hat{\Sigma}^{-1} Y_i .$$

For the ignorable case the X_i is replaced by A_i, and the Y_i becomes Y_{io}, the observed part of Y_i, and Σ becomes Σ_i (i here for ignorable) which gives a ML estimate for α of (Laird, 1988)

$$\hat{\alpha} = \left[\sum_{i=1}^{N} A_i^T \hat{\Sigma}_i^{-1} A_i \right]^{-1} \sum_{i=1}^{N} A_i^T \hat{\Sigma}_i^{-1} Y_{io} .$$

3.5.7 Repeated Binary outcomes

This subject is addressed by Baker (1995), and gives an account of an analysis similar to that of Diggle and Kenward (1994) except that this is for binary outcomes. Essentially it is based on modelling each different missingness pattern, so again modelling the missingness mechanism is all important to the accuracy of the analysis. Fitzmaurice, Laird and Zahner (1996) propose a method for modelling binary responses in terms of conditional log-odds ratios. However this means ignorable nonresponse mechanism models, unless the missingness model itself is specified.

3.5.8 Mixed models

Pattern mixture models

Little (1993 and 1994) has developed a class of pattern mixture models for normal incomplete data. The usual case for the multivariate normal is:

$$P(Y,W|\theta,\psi) = P(Y|\theta)\ P(W|Y,\psi)$$

where $P(Y|\theta)$ is the complete data model, and $P(W|Y,\psi)$ is the missing data mechanism. But for pattern mixture models:

$$P(Y,W|\varphi,\pi) = P(Y|W,\varphi)\ P(W|\pi)$$

where Y is conditioned on the missing data pattern W, and Y is in fact a marginal distribution, which is a mixture of distributions.

These two are the same if $\theta = \varphi$ and $\psi = \pi$.

If MAR, $P(W|Y,\psi) = P(W|Y_{obs},\psi)$, and models based on the likelihood are obtained by integrating out the Y_{mis} from the density $P(Y|\theta)$. So if θ and ψ are distinct, then ignorability holds.

When the pattern mixture models are used, the parameters of the marginal distribution of Y are expressed as functions of φ and π in these functions. If $\varphi = \theta$, and $\pi = \psi$, then the missing data mechanism is MCAR. Pattern mixture models provide a flexible class of models, not MAR. A two sample t-test is an adequate means of testing whether there is a difference between respondents and nonrespondents (y_{obs}, y_{mis}). However nonignorable techniques rely heavily on normal distribution assumptions (Little and Rubin,1987).

Bayesian techniques are often used using a Jeffrey's (non informative) prior. Little(1994) shows that 95% CI's are always wider for the Bayesian estimate than the maximum likelihood

estimate, and states that "when intervals deviate significantly, the Bayesian intervals are a more honest reflection of uncertainty than the Maximum likelihood". (This is borne out by the log-likelihood vs. log posterior estimates, (See chapter 9). The log posterior is always less than the log-likelihood, reflecting additional uncertainty).

Little generalises the ignorable model analysis for pattern mixture models. This is achieved by the addition of a parameter to describe the extent of missingness on X_2 as well as X_1. The value of the parameter will reflect the missingness and may vary along with other parameters in the model. With this, a fixed prior is not necessary and the standard ML estimate may be used. Park and Lee (1997) have developed a Weighted Least Squares GEE approach to longitudinal missing data, similar to Little's pattern mixture model in that it suggests additional parameters for each type of missingness pattern.

3.5.9 Likert-type scales

The literature available for dealing with missing data involving Likert scales is sparse to say the least. Likert scales are often used in attitude and opinion research.

One paper (Downey and King, 1998) deals with this problem comparing item mean substitution and person mean substitution. As both of these are essentially different versions of conditional mean substitution, there are problems not only with shrunk variance estimates, but the covariance structures are not preserved.

The problem with not imputing of Likert scales, is that the missing data are vitally important, especially when an index is formed, (summing over several variables). This is often is the case with this type of data, and if the individuals are not imputed or replaced, then the results will be biased and the index contain meaningless results unless complete case analysis is used. This may well be biased, as well as being wasteful of data.

Alternative methods for Likert scales are mode substitution (item), or median substitution, which is probably more robust. But again the variance reduction is problematic, along with artificially small p-values.

Downey and King describe Person mean substitution as:

$$Pms_i = \frac{\sum_{1}^{k} x_{ik}}{(k - m_i)}$$

and Item mean substitution as:

$$Ims_k = \frac{\sum_{1}^{i} x_{ik}}{(i - m_k)}$$

where x_{ik} = score for person i on item k

i = number of people

k = number of items

m_i = number of missing items for person i

m_k = number of missing people for item k

However the greater the number of items in the scale, the closer regression imputation is approximated. The PMS method inflates covariance and leads to incorrect results if the missingness is omitted, (for missingness >20%). Retaining the sample size and therefore the statistical power is what is important here. In his paper Raaijmakers (1999) notes that loss of statistical power, with listwise

deletion is as high as 35% for 10% missing, and a huge 98% for 30% missing.

The usual problem with listwise deletion is that missing values do not occur in a random fashion, as sometime various groups within a population will decline to answer certain questions. In the same way extreme values may be over (or under) represented by those with extreme views. if the pattern of missingness is non-random, listwise deletion is generally regarded as not acceptable. Replacement of missing values can be a good idea (rather than listwise deletion, or model based estimated procedures), as very often the scale score is only the start of a whole series of different statistical procedures.

Model based procedures such as Maximum likelihood estimation, do not assume that the missingness is random, and so are a vast improvement on list deletion. The main objection Raaijmakers has to these type of methods, is the fact that they are complicated, and that uncongeniality can be a problem. Hence he recommends a replacement option, which he calls relative mean substitution.

$$RMS = x_{ai} = \frac{\frac{\left(\sum_{k=1}^{n} x_{ik}\right)}{n}}{\frac{\left(\sum_{j=1}^{N}\sum_{k=1}^{n} x_{jk}\right)}{Nn}} \times \frac{\left(\sum_{j=1}^{N} x_{aj}\right)}{N}; (j \neq i)$$

where x_{ai} = the estimated value x for missing item a for person i,

k = the valid items (1 to n) of the ith respondent and

j = valid cases of the sample (1 to N) with no missing data, excluding person i.

This is effectively a form of conditional mean substitution, which performs reasonably well at low rates of missingness, but when homogeneity of scales is low (i.e. the variance is greater) this method does not work so well.

Bradlow and Zaslavsky (1999) describe survey non-response as nonignorable and have suggested an ordinal data structure for Likert type scores, modelled by an ordinal probit model with effects for persons and items. A Bayesian hierarchical model is developed, and a sensitivity analysis is shown which checks out parameter estimates of individual cases and distributional assumptions.

3.6 Non-Ignorable Missing.

Non-ignorability means that the missing data mechanism needs to be known and modelled to be able to impute without bias.

Dealing with non-ignorable missingness:
Most non-ignorable methods involve joint probability modelling of both the $P(Y|\theta)$ data and the missingness mechanism $P(W|Y,\psi)$, and jointly estimating θ and ψ from Y_{obs} and W. There are more parameters that can be estimated from Y_{obs} and W. Also used is stochastic censoring for continuous data, and non-ignorable contingency table approaches are used for categorical data.

3.6.1 Non-Random Missingness.

When the degree of nonresponse is non-random and therefore non-ignorable, the effect of the non-random missing data increases with the severity of non-randomness (Choi and Lu, 1995). If the missing data is ignored, the results of an analysis can be very biased, and the problem is made worse by heteroscedasticity. Choi and Lu give a test for the change in power function for random, and non-random missing data.

If the missing data mechanism is non-ignorable, then any analysis will be biased. The test is an asymptotic, one-sided test for binary data. H_o: $P_x = P_y$. (Being the two groups in a trial) $\lambda_x n$ is the expected number of observed x, and $\lambda_y n$ is the expected number of observed y. Let $\delta = E(\hat{p}_x) - E(\hat{p}_y)$, and $\Phi(.)$ denote the standard normal distribution, where $\pi(\delta) =$ the power of detecting a difference δ based on a one-sided test at the 5% level. Then:

$$\pi(\delta) = 1-\Phi\left[1.645c_\delta - \frac{E(\hat{\delta})}{(\lambda_x n)^{-1}E(\hat{p}_x)\{1-E(\hat{p}_x)\}+(\lambda_y n)^{-1}E(\hat{p}_y)\{1-E(\hat{p}_y)\}}\right]$$

and

$$c_\delta^2 = \frac{(\Gamma_x+\Gamma_y)\{1-(\Gamma_x+\Gamma_y)/(\lambda_x+\lambda_y)\}}{\lambda_y E(\hat{p}_x)\{1-E(\hat{p}_x)\}+\lambda_x E(\hat{p}_y)\{1-E(\hat{p}_y)\}},$$

Where $\Gamma_x = \lambda_x E(\hat{p}_x)$, and $\Gamma_y = \lambda_y E(\hat{p}_y)$.

For Normal Data: (Again one sided at the 5% Level)

$$\pi(\delta) = 1-\phi 1.645 - \left[\frac{E(\hat{\delta})}{\sqrt{(\tau_x^2/\lambda_x n_x + \tau_y^2/\lambda_y n_y)}}\right]$$

Where $\tau^2_\Xi = E(X^2)-\{E(X)\}^2$, similarly $\tau^2_y = E(Y^2)-\{E(Y)\}^2$.

The power is determined by comparing the results with MAR data with the same extent of missingness.

3.7 Data Models

Data models may be continuous, categorical, mixed, or mixture models, as in panel or clustered data, longitudinal data, or case control studies.

3.7.1 Multivariate Normal

Models for nonignorable missingness must include the missingness mechanism. This is often done by the use of an informative prior, applied as the normal inverted Wishart distribution, to the multivariate normal, to give a normal posterior distribution. Alternatively if the missingness mechanism is known, this can be incorporated into the imputation model.

3.7.2 Multinomial (Saturated)

This assumes that the categories are unordered or nominal. If the data is ordinal, then the multivariate normal may be a reasonable approximation. This is the case with 'coarsening' of the data. If the ordinal quality of the data is ignored, then multinomial methods may be used on the data, but this will mean throwing away information, something which is to be avoided at all costs in missing data analysis.

If Y_1, Y_2, Y_3,.....,Y_p are recorded for n units then the complete data can be expressed as an n×p data matrix Y. If iid then Y is a contingency table with D cells, where $D = \prod_{j=1}^{p} d_j$ is the number of combinations of y_1, y_2, y_3, ...,y_p (The missingness pattern). The cell probabilities must sum to 1, i.e. $\sum_{d=1}^{D} \theta_d = 1$, so the multinomial model has D-1 free parameters (see chapter 3.5.3). The missingness mechanism needs to be modelled here for this to be valid. This could be done via a constrained Dirichlet prior, or including the missingness mechanism in the imputation model.

3.7.3 Loglinear

A loglinear model has the same underlying distributional assumptions as the multinomial model, but it also imposes further constraints upon the elements of θ, Let $\eta_d = \log \theta_d$, d = 1,2,3.....D. We write $\eta = (\eta_1, \eta_2, \eta_3, ..., \eta_d)^T$ or more usually $\eta = M\lambda$, where M is a design matrix which is fixed and known, and λ is a parameter vector. We can test loglinear models by comparing their deviances as in:

$$G^2 = 2 \sum_{y \in Y^*} x_y \ln \frac{x_y}{n\hat{\theta}_y},$$

which is asymptotically equivalent to the Pearson [χ^2] Goodness of fit test

$$\chi^2 = \sum_{y \in Y^*} \frac{\left(x_y - n\hat{\theta}_y\right)^2}{n\hat{\theta}_y}.$$

(Agresti, 1990) Alternatively, for ordinal data, the method of chapter 8 may be applied using a constrained prior to model the missingness mechanism.

3.7.4 General Location Model

This is used when the data is mixed, i.e. partly categorical and partly continuous. The categorical part is divided into a contingency table, and the continuous part is assumed to be multivariate normal. The categorical cells are allotted numbers $r = (r_1, r_2, r_3, \ldots r_d)$ with frequencies $y = \{y_r : r \in R\}$. Let D = the number of the cells in the contingency table indexed by $d \in (1,2,\ldots.D)$, U = n×D matrix, of indicators for the position of cell 'd', μ_D is a D-vector of means corresponding to cell 'd', and Σ is the D×D covariance matrix. It is helpful if the proportion missing are arranged so that:

$$r_{mis} = (r_{1\,mis} < r_{2\,mis} < r_{3\,mis} < \ldots.. < r_{dmis}) .$$

The general location model, named by Olkin and Tate (1961), is easily defined in terms of marginal distribution (multinomial) R and conditional distribution Z given R. Given R, the continuous part Z is modelled as multivariate normal. When there are two categorical groups, this reduces to the model underlying Discriminant analysis. Again either a suitable prior needs to be applied, (probably the constrained Dirichlet, and the inverted Wishart) or the missingness mechanism included in the imputation model, for nonignorable inferences to be valid.

Parameters of the general location model are:

$\theta = (\pi, \mu, \Sigma)$, where $\mu = (\mu_1, \mu_2, \mu_3, \ldots.,\mu_d)^T$ is a matrix of means. The variance Σ in θ needs to be positive definite for Σ , and $\sum_{r \in R} \pi_r = 1$, and the model for Z given R is a multivariate regression $Z = U\mu + \varepsilon$, where ε is an n×D matrix of errors, whose rows are independently distributed as $N(0,\Sigma)$, and the matrix of regressors is used to predict

each column of Z. In practice this is only useful when the number of groups in the categorical part is small compared with n, the total number of cases. Also there need to be sufficient numbers in each cell, so as to estimate μ.

3.8 Likelihood theory

Likelihood based models are only valid under the ignorability assumption. Therefore if there is nonignorable nonresponse, then the missingness mechanism needs to be modelled, and incorporated into the analysis, in order that likelihood based models give correct inferences.

3.8.1 Coarsening

Coarsening is used when continuous data is discretised, and it is possible to impute to a band as opposed to an actual figure. This is a popular option particularly with income data which tends to be informatively missing. It may be possible to apply MAR techniques to coarsened data without creating too much bias. The trade off here however is that a certain amount of the precision in the data becomes lost. This is a case of throwing away some data in order to gain other data. (Heitjan, 1999)

3.8.2 Sensitivity to Normality

Sensitivity analysis is the only way to test for informative missingness, and in reality distinguishing between MAR and NMAR, is very difficult. Many imputation techniques depend heavily on the

assumption of normality, which Meng (1994), and Barnard and Meng (1999) show may be problematic.

3.8.3 Categorical

Categorical Nonignorable Missingness

Bonetti, Cole and Gelber (1999) have developed a method of moments (MM) estimation procedure for categorical data with nonignorable missingness. Essentially this is a methods of moments alternative to the usual maximum likelihood test. The MM estimates are simple compared to ML estimates. There is a test whether or not the missingness in the data is ignorable. The MM fitting of a model is a convenient sensitivity analysis for the MAR assumption. By allowing for nonignorable missingness, the bias may be reduced.

3.8.4 Bayesian Approach

The Bayesian approach for nonignorable missingness is to use a constrained prior, which will provide information about the underlying distribution, or alternatively use prior knowledge about the missingness mechanism. If the missingness mechanism is unknown, as it so frequently is in NMAR cases, then the only real option is to use an appropriate prior, apply it to the model in question, and then take draws from the resulting posterior distribution to impute the missing values.

3.9 Analysis of missing data

There are many and varied ways of analysing missing data. See chapter 5 for a discussion on types of imputation, chapter 6 for a review of software for handling missing data, and chapter 7 for rules for selecting an appropriate method of imputation. Where multiple imputation is used, a set of rules is given by Rubin (1987) for combining estimates.

3.9.1 Rubin's Rules for Recombining Estimates

Rubin's Rule (Rubin 1987)

After obtaining m imputations of Y_{miss} analyse the M completed datasets and combine the result: [Rubin's Rule]

$\hat{Q}$ = complete data point estimate U = complete-data variance estimate

$$\overline{Q} = m^{-1}\sum_{t=1}^{m}\hat{Q}^{(t)} \qquad B = (m-1)^{-1}\sum_{t=1}^{m}\left(\hat{Q}^{(t)} - \overline{Q}\right)^2$$

$$\overline{U} = m^{-1}\sum_{t=1}^{m}U^{(t)} \qquad T = \overline{U} + (1+m^{-1})B$$

The Interval Estimate is $\overline{Q} \pm t_v\sqrt{T}$ where $v = (m\text{-}1)\left[1+\frac{\overline{U}}{(1+m^{-1})B}\right]^2$

Let B = the between imputation variance, and T = the total variance estimate. Let $Q^{(t)}$ be the point estimate using the t^{th} set of imputed data, and $U^{(t)}$ be the variance using the t^{th} set of imputed data.
The variability provides a measure of uncertainty due to missingness. This is the advantage over EM which has one dataset, but no standard errors.

If the fraction of missingness is λ, then the efficiency of m imputations is: $(1+\lambda/m)^{-1}$ for the variance. (Rubin 1987, p114). If λ= 0.3, when m=5, the estimate is 95% efficient. For this to be so, the imputations need to be 'proper' (Rubin 1987). Iterations of Y_{miss} must be separated to ensure independence, otherwise they will be correlated.

3.9.2 Rules for Analysis: % missing categorical, mixed, and continuous.

Rules for analysis include determining the type of data, the type of missingness, the extent of missingness, the end use of the dataset, and whether a full analysis is required, and is the data set a large one. (see chapter 7, and the appendix.)

3.9.3 Longitudinal data

Longitudinal Nonresponse

Diggle and Kenward (1994) provide a method for dealing with Longitudinal Nonresponse that is nonignorable, essentially drop outs. Here a logit model for univariate missing data is extended. What is shown to be important here is that any model here dealing with nonignorable nonresponse needs to have the missingness mechanism fully specified. The sensitivity of the model to this non-ignorable nonresponse needs to be investigated. Supplemental information remains a good choice for modelling non-response, and if suitable explanatory covariate information is available, the non-ignorable case may be converted to ignorable non-response (MAR). Under this assumption, standard methods such as ML may be

applied. Alternatively Pattern Mixture modelling could be applied, where the sample is stratified by the missing data pattern and each stratum is modelled separately (See chapter 3.9).

3.9.4 Bayesian Methods (Multiple Imputation): as applied to Frequentist Ideas.

A number (M) of complete data sets are usually created sequentially.

In reality they often are iterates of Y_{miss}, particularly when Imputation is 'circular'. That is, $[Y_{miss}^{(t)}, Y_{miss}^{(2t)}, \ldots\ldots\ldots, Y_{miss}^{(mt)}]$ are collected by Data Augmentation, where t is large enough to ensure independence. Data Augmentation (DA) is similar to EM in that the E+M steps become the I(mputation)-steps, and P(osterior)-steps. This is a type of Markov Chain Monte Carlo technique for creating draws from probability distributions.

For the I-step draw a value of the missing data from Y_{miss}

$$Y_{miss}^{(t+1)} \sim P(Y_{miss}|Y_{obs}, \theta^{t})$$

and the P-step a new value of θ, from the complete data Posterior

$$P(\theta|Y_{obs}, Y_{miss})$$

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{miss}^{(t+1)})$$

The distribution only depends on the previous draw (Markov property). ACF plots assess the convergence. Traditional MI methods have a very Bayesian flavour and rely heavily on normality. In the case of binary data the results could be unreasonable.

3.9.5 Parameter Expansion for Data Augmentation

Liu, Wu (1999) provide an algorithm similar to that of PX-EM,(see chapter 3.5.1) but this time the parameter expansion is that of data augmentation. Auxiliary variables are used to speed up a Gibbs sampling algorithm. This is essentially a reparameterisation. Even if an improper prior is used, the resulting posterior distribution expansion parameter is always proper.

3.9.6 Nonparametric Method

Newton and Zhang (1999) provide an algorithm for Nonparametric Bayesian analysis based on a one step posterior predictive distribution. This is achieved via the Markov Chain Monte Carlo method. They describe this as a type of stochastic approximation algorithm derived by Nonparametric Bayesian considerations.

3.9.7 MCMC Algorithm.

Gelfand and Sahu (1999) State that applying priors in an MCMC situation need to be carefully considered. A prior which is too precise will influence the inference about the posterior distribution. If the prior is too diffuse, then the posterior surface is badly behaved. However the authors show that if a diffuse prior is used, often a proper posterior is obtained for lower dimensional parameters.

4 Motivation and Data Description

4.1 The problem:

1 To review missing data literature and current imputation methods, and to develop a strategy for imputing values into different types of missingness, and different endpoint requirements, and to compare different analyses.

2 To develop a likelihood based method for dealing with Likert type data. Currently the available methods are variations on class mean imputation, either person mean based or variable mean based, with very broad classes. In this day and age seventies' type methods are outdated and a more up to date method for this type of data needs to be developed.

4.2 Motivation for this study:

The writer was involved in the analysis of the Manawatu Nutrition data (Watson, 1996), and missing data was handled by case deletion (available case analysis). It soon became apparent that the data was not MCAR, and a better way of dealing with this problem needed to be used.

During this time the 'Nutrition in Pregnancy' study was expanded, and data was collected during 1998 and presumably 1999. Whilst working on the preliminary statistical analysis during 1998, it was clear that this data was in fact informatively missing. For example

data was collected on the mothers' weights, but mothers with greater weights were much more likely to not record their pre-pregnancy weights. In the case of multiple item missingness, this was found to be MAR, as ethnicity was a very good indicator for the missingness mechanism in many cases, as was age, and other socio-economic factors. So it was very clear to the writer that case deletion was not the best way to handle this problem, especially as the rate of missingness was over fifty percent for some variables.

4.3 The two data sets used here.

1 The Nutrition in Pregnancy Dataset - known as 'Nutrition' dataset.

2 The Genetic Foods Dataset. This is the result of an Australian survey into attitudes towards Genetic Engineering.

4.3.1 Nutrition Data set.

The Nutrition in Pregnancy Study is a study of the maternal food intakes and activities during pregnancy. The study was centred on Palmerston North, and included both rural and city women. For ethical reasons all mothers involved here are volunteers, and only those aged between 18 to 35 were supposed to have been included, although some older mothers were actually included. Excluded from the study were mothers with known serious medical and obstetric problems, as it was intended to be a study of normal healthy pregnancies. This could be the reason why the range of the babies' birthweights was higher than expected.

The aim of the study was to explore the relationship between the diet of the mothers during pregnancy, as well as their activities over a set period and maternal and babies' health (measured by growth). Data collected included how the mother felt during pregnancy, (and during the post partum period), also how tired she was. Her medical background (as recalled by her) is also collected, as was qualitative information on medical care during pregnancy, the birth experience and post partum. In addition, the relationship between morning sickness, carsickness, and oral contraception was also explored. The mother's height was recorded at the 4-month visit, and weight and skinfold measurements are recorded at 4 months and 7 months during pregnancy, and at 2, 6 and 12 months post partum. These form longitudinal variables - the 'how the mother is feeling' ones are categorical, and the weights and skinfolds continuous data.

At the initial interview demographic data was collected on age, income, socio-economic status, education, ethnicity, level of work, involvement in groups, support the mother received, smoking prior to and during pregnancy, alcohol consumption prior to and during pregnancy. The activity data was collected over two periods of 8 days, one in the 4^{th} month, and one in the 7^{th} month. This data was then averaged into lots of 1440 minutes (added together and then divided by number of days collected). There was some complete data, but also a number of the mothers completed less than the complete number of days. By averaging over the actual days collected, a representative (average) 'day' is given for the 4^{th} month and the 7^{th} month. For the dietary data again two lots of 8 days of data were collected, although not all of the mothers completed all of the days. Again the same strategy is applied, to give one representative (average) 'day' for the 4^{th} month and the 7^{th} month, for the nutrient intake data.

Also collected were the babies expected delivery date, actual date of birth, weight and head circumference at birth, as well as birth difficulties. Because of difficulties of getting measures at a particular time post partum; the records from babies' healthcare record book of measurements at 4 weeks, 6 weeks, and 8 weeks were recorded where present; this resulted in a lot of missing data. For some of the babies all three sets of measures were present, for others, just one or two (height 6 weeks, head circumference 6 weeks). Also the babies recorded measurements at one year are recorded by the study, as was the place of birth.

The Demographic categorical variables tended to be completely observed, with the exception of number of school children, number of pre-schoolers, and place of birth. These three variables each had one missing value that was imputed prior to many of the analyses for Normal data, using Discriminant analysis. One case had rather a lot of missing data - including the birthplace and the entire baby measures etc. The mother miscarried during pregnancy, and one wonders at the appropriateness of imputing this particular case.

That misgiving aside, all the other cases involved live births, (in two cases twins). The extent of item nonresponse in the data set is varying and there is probably good reason to assume all MAR. The nutrient data for the 4^{th} month was completely observed, for the 7^{th} month there were 4 missing values (2% missingness). Other continuous variables included sleeplie - the number of minutes in a day spent sleeping and lying down. This had a low missingness of 1.5%. Seven (3.5%) mothers did not record their height and three (1.5%) did not record their weight at four months. Age had a missingness rate of 1%, whereas the activity sleeplie at seven months was missing for thirteen (6.5%) of cases.

However, the real missing data began to appear with the babies' measurements post partum. Twenty-six (13%) of babies had no head circumference measurement recorded at birth. The six week measures (explained earlier) had a moderate rate of missingness for weight (twenty-two or 11% missing), and a high rate of missingness for head circumference (seventy-nine or 39%) and extreme missingness for height at six weeks (one hundred and twenty-four or 61%).

The baby measures at a year were not that great either: fifty-seven or 27.5% were missing for weight at one year; sixty-eight or 34% were missing for height at one year, and eighty-four or 41% missing for head circumference at one year. Length of breast-feeding (censored at twelve months) had a reasonably moderate rate of missingness, twenty-eight or 14% of cases.

The 'Nutrition Data Set' is an abridged (variable truncated) and randomised version of the data set from the Nutrition and Pregnancy Study. This is to ensure confidentiality.

If all of these baby measures are put into a regression analysis, only 29 out of 197 were complete cases. (Although most of these only had one missing observation).

4.3.2 Genetics Foods Data Set.

This was a dataset collected by Norton (1999). The study was to look at the attributes of Genetic Engineering (G.E.), and how widely genetic engineering is understood in the community. G.E. is a relatively new technology, which has only been developed over the

last 20 years. For the average Joe Citizen, not much has been disseminated about G.E. until very recently. This survey was conducted by the mail out method, a self-administered survey. Whilst face to face surveys are generally more successful, the cost of such a survey would have been prohibitive. Telephone surveys with random digit dialling were considered but the known nonresponse rate with this method was considered too high. The alternative was a mail survey. The advantage of this is the respondent does not have to share responses with the interviewer. Measurement Error is the difference between the answer on the response form, and the true answer. It can be a major problem if some respondents give an answer 'which the interviewer wants to hear'. So this form of measurement error is minimised in mailed questionnaires. As telephone directories are not a complete list, the electoral roll was chosen as the target population. In Australia 98% of citizens over the age of 18 are on the electoral roll (12 million).

Random number generation was used to provide a sample proportional to the size of each state or territory, giving a total sample size of 2291. The survey was administered as a 4-stage method. Firstly a postcard was sent out to the prospective respondents advising them they had been selected in a survey and would be receiving a package in the near future. Secondly the actual survey package was sent out, this contained the survey questionnaire, a letter to the respondent, a reply paid envelope and a magnet (this gave the Central Queensland University Crest, and a contact phone number for the researcher).

The third contact was a follow up letter, sent in late December 1996 - A simple reminder that the survey had been sent and please could they complete it. This was only sent to those who did not complete the survey - this included a contact phone number for

people who wanted more information. Finally a replacement survey was sent to those who still had not responded a month after the reminder letter.

Ultimately 1009 of the 2291 people contacted responded, with Northern Territory having the poorest response rate 35% and ACT having the best 54%. The overall total response rate was 45%. However, 132 were returned as unable to be completed due to language problems, death, age or illness. If these are discounted then the overall response rate rises to 52%.

If the unit nonresponse is considered in the original analysis, stratification by state enabled the possibility of dealing with unit nonresponse by weighting methods.

Likert Scales used:

For each issue, two opposing views were presented one on the left and one on the right. The questions asked were Likert-type questions, using a six-point scale.

1. Strongly agree with view on left.
2. Agree with view on left.
3. Mildly agree with view on left.
4. Mildly agree with view on right.
5. Agree with view on right.
6. Strongly agree with view on right.

Two opposing views are presented, on either side.
A six-point scale was chosen here to prevent fence sitters.

The survey was designed in four parts.

1. Questions relating to G.E.
2. Questions relating to Science and Technology.
3. Questions relating to Food and Attitudes.

4. Demographic data.

1. The G.E. questions

Firstly asked about G.E. in specific cases then attitudes to G.E. in general. The specifics consisted of discussing:

a. Tomatoes with altered gene structure.
b. Cheese.
c. Insect resistant wheat.
d. The blue rose.
e. The lean pig.
f. Blowfly resistant sheep.
g. Tomato with fish genes.

The idea being to determine whether there was concern at the technology, concern related to particular products or no concern at all.

Questions asked in each genetically engineered gene example that could include:

a. The type of G.E. was acceptable.
b. Release would cause environmental damage.
c. The product is important.
d. Eating the product would cause long term health effects.
e. Would the respondent buy the product?
f. Would they be worried about eating the product?
g. The product would be a good idea if properly labelled.
h. The benefits of the product outweighed the risks.

2. The Science and Technology questions.

The first part asked questions about the respondent's general knowledge of matters of technology, and attitudes to research in

general. The second part asks simple science questions. The third part ascertained where the bulk of the respondents knowledge came from (which media) and the fourth, how the respondents rate their own scientific knowledge.

3. Attitudes.

First part is attitudes to different foods and shopping habits. Second is, 'Are the labels studied?' Third is whether the respondent eats at fast food outlets regularly. The Fourth part is mainly questions to decide how conservative the respondent really is.

4. Demographic Questions.

a. Sex.
b. Age.
c. Church attendance, belief in God, Denomination.
d. Size of town where they have lived (7 size categories) and was it in Australia / Overseas. Also in childhood / Adult years / Now.
e. Postcode of residence.
f. Language Spoken.
g. Educational level attained.
h. Science level attained.
i. Total household income.
j. Children in household.
k. Does the respondent shop for food for the household.
l. Does the respondent prepare meals for the household?
m. Any members of the household having special dietary requirements.
n. How often the respondent interests themselves in science and technology matters.
o. Asks the respondent if they have been part of a special interest group, scientific, consumer, social, political group and so on.

p. Does anyone in the household have a science or technology related job.
q. Political affiliations.

Item nonresponse.

For the first question the nonresponse was 93 out of 1009 or 9%. For the rest of the first part all of the item nonresponse is below 5%. For the second part, Item nonresponse for the technology knowledge questions was around 2.5%, and the science questions up to 4%. The third part of this section had low (around 1%) item nonresponse. The fourth question 2% nonresponse.

In the third part of the questionnaire for the attitudes to food questions nonresponse was again low. Less than 5%.

For the Demographic questions, nonresponse to the gender question was higher than that of age, Gender being 3%. Age nonresponse was 1%. Gender may well be able to be imputed by look-up tables, as the respondents are selected from the electoral roll, but this is not always the case, as some names are without gender distinction.

The religion questions were well answered with only the 'Do you believe in God' question having a nonresponse rate of greater than 5% (6%).

The location of resident's questions was surprisingly badly answered with nonresponse rates of between 9 - 15%. The postcode question provided a non response of 3%, (Here look-up tables could have been used, as it was a mailout survey, so the researchers had access to this information) but the English language question is probably the most likely to be NMAR, as

probably the event of non response is quite likely is related to the respondent's inability to speak English.

The education question is above the 5% threshold but the science education and employment questions both lie at about 3%. The income question is above the 5% threshold as expected, as is the political affiliation question (both around 7%). Others here were less than the 5% level.

Due to time constraints, all that was done with this data base was to illustrate how different types of questions give rise to different extents of missing data, even within the same dataset of willing respondents.

5 Imputation

5.1 What is Imputation, and why Impute?

Imputation is a term to describe a variety of ways to generate (a) value(s) to replace a missing item, and therefore form complete records for each case.

Why Impute? Imputation is largely used to reduce non-response bias, and to make use of data that would otherwise be discarded. Standard statistical software will often discard without warning the user, especially in the case of multiple linear regression. What started out as 197 cases can rapidly become 30 cases! Weighting can be a form of imputation - although it is generally used for correcting unit nonresponse. However using the correct form of imputation is vital.

An inadequate form of imputation, whilst correcting for nonresponse bias, will distort the relationships between variables, particularly affecting the variances and covariances. If imputation is not carried out then the case mentioned above (multiple linear regression) is a problem, as is the case of missing values in ordinal categorical data, such as Likert scales that are aggregated into an index.

The effect of a few missing values becomes a major problem in this kind of analysis.

Classes of imputation:

External, or deductive or lookup methods: These are methods not based on data of complete cases, where the imputed values may

be gathered from known information, often other questions in the same survey. (A warning here, too many validation type questions in the same survey will lead to interviewer fatigue.) Using a look-up table where appropriate.

Deterministic methods:

These are methods that always give the same answer, given the same set of explanatory variables. For example using CART or Regression.

Stochastic methods:

These methods give a different answer, given the same set of characteristics. An example is a deterministic method with an added random component.

Multiple Imputation methods: May be frequentist, or Bayesian. The idea is to impute m values (m>1) for each missing value, thereby forming m complete data sets. Each dataset is analysed separately in the manner that the analyst chooses, as if complete data. The results from each of these analyses are combined according to 'Rubin's Rules' (see chapter 3.9.1) to give final estimates. The advantage here is that the variance is preserved (actually slightly inflated, and correctly so, as this reflects the uncertainty due to the imputed data).

The advantages of imputation are:

- Imputation provides complete data sets.
- Complete data sets allow the use of complete data analysis, that is no case deletion, or waste of data.

With an imputed data set, standard analysis is allowed to proceed.

Disadvantages of Imputation are:

- Can allow data collection to influence the type of imputation.
- Increases the overheads of a survey, although if this is weighed against the cost of repeated callbacks, this is not such a disadvantage.
- Imputed data will decrease the credibility of the data, any inferences drawn, should be kept to a minimum.
- The majority of the 'Quick and Dirty' methods will reduce the variance, giving the double jeopardy of bias and spuriously optimistic precision.

5.2 Complete Case Methods Overview[α]

Complete case methods essentially mean deleting all cases containing missing values. This may severely reduce a dataset, and this is valid ONLY when the data are MCAR. However if the assumption of MCAR is true, and the dataset is sufficiently large to allow wasting of data in this way, then the inferences for a population will be accurate when these methods are used, that is, the information matrix is valid. Complete case methods are the default used by all the major statistical packages. In reality however the MCAR constraint is too restrictive to be generally observed, and it is rare indeed that a dataset will achieve this ideal. (See Little's test for MCAR, chapter 3.4 - this test rejects when the data's missingness is not MCAR).

[α] Whilst it is recognised that the methods of this section involving Available Case, or Case Deletion, are NOT in fact Imputation Methods, they are included in this section for completeness, as they are still (unfortunately) a widely used method of dealing with Missing Data.

5.2.1 Case Deletion

This involves deletion of any case that has any elements missing, so as to provide a rectangular dataset of completely observed values (before any analysis is attempted). Case Deletion is used in ordinary regression analysis, as a diagnostic procedure- (for example, Cook's distance). Influential observations are noticed when the effect of deleting them is a large change in the resulting estimates.

Omission of observations has three possible effects.
1 The significant result becomes more significant.
2 The insignificant result becomes significant. (Or vice versa)
3 The insignificant result becomes even less significant.
The second case causes a problem, more so than the other two.

Tests of hypothesis are commonly used for selection of variables in multivariate regression, so the influence of observations on the test statistic can change the whole decision-making process.

Omitting cases, or variables, or both, may seriously bias the analysis if the missingness mechanism is nonignorable. (Little and Rubin, 1987: See chapter 3.6, Nonignorability). Researchers, due to their being unaware of its existence, very often ignore the bias introduced by case deletion. This means inference about populations can be very unreliable, if bias exists. (If the missingness mechanism is not MCAR). VERY FEW surveys have an MCAR missingness mechanism; this condition is too stringent.

5.2.2 Available case

In SPSS this is also known as the pairwise case, meaning that only the cases missing in the chosen analysis are deleted. The problem with available case deletion is illustrated in the case of multiple regression, where it may be impossible to compare one model with a preceding one as the number of cases changes with each regression model, as cases are included or dropped out (they are not nested). In fact the number of included cases can increase or decrease with the included dependent variables. So a particular disadvantage of pairwise deletion (deleting cases only as they are missing) is that it destroys the nesting procedures, and makes comparing of models virtually impossible.

5.2.3 Logical substitution and Look-up tables

These methods, when it is possible to use them, would possibly be one of the most accurate forms of imputation, but its use is limited to known external information. A good example is 'Are you married?' If say a previous question on age is answered, and the respondent is fourteen years old then, then to impute the answer 'No' would be reasonably safe.

An example of Look-up tables, is where the question is one of income, and the respondent is a beneficiary. Then the amounts are known and available in Look-up tables. If the questions are correctly answered, these methods have the greatest chance of giving 'true' values. Look-up tables are used particularly in 'Government type' surveys. However Look-up tables may not be as accurate as deductive methods, but may be slightly more widely applicable.

5.3 Mean Based Methods Overview

These essentially involve calculating the mean for a particular variable, and imputing this same estimate into each missing value space for that variable to provide complete data. The advantage of mean imputation is its simplicity and ease of implementation, and also that it is very easy to explain to a non-statistician. Any statistical software or spreadsheet can calculate the mean and replace the missing values. If the missing data are MCAR and only estimates of the mean, or column totals are required (n $\times$ the mean) then, and only then, may this method be used. The disadvantages with this family of imputations is that the distribution of the variable will become compressed, and increasingly so with increasing proportions of missing data. If this is used, sampling error estimates will be invalid, and so the relationship between variables will be distorted, and that the information matrix will no longer be valid.

5.3.1 Mean Substitution

The mean of a variable is imputed into each missing value for that variable. This is primarily used for continuous data, where the underlying distribution is known to be normal.

5.3.2 Mode Substitution (categorical)

This is used where the data are categorical, particularly in the case of nominal data. The mode is calculated, and this is then imputed. Again this method would be chosen for its ease of use. There are no underlying distributional assumptions here, only valid under the assumption of MCAR.

5.3.3 Median Substitution (robust)

Median substitution would be applied in two separate cases:

- Instead of mean substitution when the underlying distributional assumption of normality is violated (that is, the data are strongly skewed). The advantage of this method is that it is as easy to use as mean substitution.
- If the data are categorical, and ordinal, then median substitution may be appropriate for Likert type data that are later to be aggregated into indexes.

Again this type of imputation will affect the variances and relationships between variables. Median substitution is only valid when the MCAR assumption holds.

5.3.4 Discriminant Analysis

This method works well for categorical data, when there is fully observed covariate information. Initially a regression analysis is done on the fully observed covariate, using the categorical variable as an explanatory variable (one way ANOVA if more than two levels). If the regression is significant with a reasonably high R^2, or in the case of the ANOVA, a significant F-test, then this covariate may be used in standard Discriminant analysis, to reasonably accurately predict the 'true values', which are missing.

The categories which the Discriminant analysis places the missing values in may then be imputed into the datasets (M^{c}Donald, 2000, Personal Communication). This method requires the missingness mechanism to be MAR, (as it is based on other observed covariate information) and should maintain the variance estimates, provided

the degree of missingness is not to great. Relationships may also be maintained (as this is categorical). The disadvantage with this method is that it depends on having a fully observed continuous covariate that is strongly correlated with the variable of interest. This may not be available, but if it is available, the method should probably be the method of choice for nominal categorical data.

This idea doesn't appear to have received much coverage in the literature apart from an obscure reference about generalised discriminant analysis in mixture modelling (Barnard and Meng, 1999), and deserves further study.

5.3.5 Stochastic Mean Substitution.

This is mean substitution with a random residual component added in; that is $\bar{x}+\varepsilon$ where ε is assumed Normal iid. This **is** better than mean substitution as there will be less reduction in the variance estimates. However again the method should only be used under the assumption of MCAR. The method would be reasonably easy to implement, particularly if the variable of interest is standardised normal N (0,1). Stochastic mean substitution is very much based on the assumption that the original distribution is in fact normal.

5.3.6 Mean within category substitution (conditional)- class mean.

Essentially the data is stratified and each stratum has its own mean. If more than one variable is to be imputed, then the strata will probably change for each variable, suggesting a variable by variable approach. The use of AID (automatic interaction detection) for continuous variables and CHAID (χ^2 automatic interaction

detection) for categorical variables can be useful for deciding strata membership. These are found in SPSS. Then the class mean is imputed for each missing value within that class. This method is suitable when the missingness mechanism is known to be MAR, and the variable of interest is correlated with other (few) variables.

This could possibly be used in Likert type data, where the variables being imputed may later be aggregated into an index. This method is more useful than mean imputation for reducing nonresponse bias, but the problem of compressing the distribution still remains, and estimates of the variance or the other relationships between variables in the data not used in defining the strata, are distorted. Mean within category (conditional class mean) substitution is really only useful when means and totals are required.

5.4 Data Substitution Methods Overview

These include putting in actual values taken from responding units and placing them into the empty spaces of non-responding units. This is best done not at random, but using 1: stratification 2: sampling without replacement. This method assumes all cases are equally likely and MCAR.

5.4.1 Colddeck

This is rarely used today, but involves the imputing of information from an external source such as an earlier survey. This is not recommended, and only used when a researcher has valid reasons to believe the externally imputed value to be more valid than an internally imputed one, such as a known nonresponse bias.

5.4.2 Hotdeck- random

This is where a randomly selected case chosen from the responding units has its data values placed into those of a nonresponding unit. This is better done without replacement (Lessler, Kalsbeek, 1992). Real values are assigned to the nonrespondent from a similar responding unit. Hotdecking is in fairly common use today in industry. Government departments and survey firms will commonly employ this particular tactic. Its big advantage is that the random hot deck method does not have the variance reducing problems that beset mean substitution.

5.4.3 Hotdeck- next available case.

This is a random variation on the theme of random Hotdecking without replacement. The next available case is used to impute the missing value.

- Sequential: This implementation uses Imputation classes (strata). A donor value is applied to each class. Each class is processed sequentially. If a case is missing, this is replaced by the stored value for the class. If it is observed, then it replaces the stored value for its class, and may be used for imputing into subsequent cases within the class. The advantage of this is that often 'like groups' may be in sequential order. This is useful for census data, where there are many cases and few variables. Sequential hot decking is used in the 1991 UK Census. A disadvantage is that in forming the classes, continuous data needs to be discretised (coarsened) therefore losing detail (the

class indicator). Sequential Hotdecking can be used when the data are MAR. (Mills, Teague, 1991).

- Hierarchical: The hierarchical hot deck method is implemented by sorting the data into a heirachy of imputation classes. This method is similar to the sequential hot deck, except that the class boundaries are either rigid, say age; or not, say geographical area. This method has the advantage that it allows the use of other variables to be included, without building complicated models. Again it is useful if the missingness mechanism is MAR, but again continuous data needs to be converted to discrete groups. Hierarchical hot decking is used in the US current population survey. (David, Little, Samuhel, Triest, 1986)

5.4.4 Last value carried forward (Hot deck)

This is also a form of hot decking; the last recorded value is imputed in. This is very similar to sequential Hotdecking, although without classes or strata. This is useful only if MCAR.

5.5 Time Series Models Overview

Time series models are used in longitudinal surveys or panel surveys (repeated measurements). Essentially if a value is bounded missing [the value before (t-1), and the value after, (t+1) are present] then interpolation can be used to obtain the missing value and impute it. In this way a last value carried forwards in repeated measures can be used if the value at t-1 is observed, but not at t,

and this value can be imputed into the missing value. There is some justification for this method, as it is from the same case. However the model does not take into account the mechanism for 'what has changed between measurements' for example, mothers skinfold measurements at different stages of pregnancy, which are different because of physiological changes, and time series methods assume no underlying changes. This method could be useful if MCAR.

5.5.1 ARIMA models

ARIMA models are useful after appropriate differencing and seasonality components and data transformations are taken into account. Linear models may then be used. In the case of Normal data, the model giving the minimum sums of squares of residuals is the maximum likelihood fit. ARIMA models are useful for MAR data. Software used in forecasting (or backcasting) mode could predict the missing values. This type of modelling is used in Proc Expand, in SAS, to predict missing longitudinal values in time series. ARIMA modelling can be done in Minitab, S-PLUS, SAS, SPSS and ASTSA, as well as other major packages.

5.5.2 Kalman Filter models

The Kalman filter is used to model: First a system trend with noise; Second a system slope with noise at each instant of the first calculation; Third the previous instant's slope with noise to this instant's slope. (Essentially a refining process by differentiation). The Kalman filter gives the slope and trend at any particular instant, and uses ML to estimate the three noise estimates. The Kalman filter is useful for small datasets. The use of forecasting would give

fitted values, which are then incorporated into the model, which is refit until convergence (Box, Jenkins 1970; Box, Tiao, 1975).

5.5.3 Period on Period Movements Ratio.

The period on period movements ratio is based on taking the difference at t=1, and t=2 and calculating the average ratio between the two or more periods. This means as long as the first value is observed, the second value can be calculated by applying the ratio from one to another. Period on period movements ratio methods are suitable for short period panel surveys (monthly or quarterly) where the case by case variance is constant and assumed MAR. This method is relatively straightforward to apply, and is used in business surveys with economic data. Imputation is achieved using backward revision and forecasting. A disadvantage with this method is that if it is used for multiple variables, it will not maintain the relationship between variables.

5.5.4 Within Case Year on Year Movements Ratio.

Within case Year on Year Movements Ratio is based upon an algorithm of taking the movement from last year to the current year without using any imputed values. The average is calculated, (or possibly the median if the data spread is not normal). The value of the previous year is multiplied by the average of the movements to obtain an imputed value. This method is suitable when there is a small movement (year to year) variance, but a large case by case variance. This method is suitable only when MAR, but variances and the relationships between variables are distorted. It is generally used in economic surveys.

5.6 Regression Imputation Overview

This is used to predict the missing values based on regression models involving the other variables, and the fitted values are put into the model.

5.6.1 Predictive Regression Imputation

Predictive regression imputation is useful for correlated covariates, but not so good with categorical variables with many categories. This method can take into account many different variables, and so reduce the nonresponse bias. Predictive regression imputation is useful for MAR. However it compresses variance estimates, and distorts the relationship between variables. A different model is required for each variable containing missing values. That is, the composite data set is built up iteratively. (Lessler, Kasbeek, 1992; Kalton, Kasprzyk, 1986) Usually the variables are imputed sequentially, from the least missing to the greatest missing, and an imputed value may be used in a later prediction. Where the variable of interest is continuous, then linear regression is used, although if the data is skewed a transformation may be used to correct the skewness and possibly stabilise the variance. (David, Little, Samuhel, Triest, 1986)

5.6.2 Predictive Mean Matching

Predictive mean matching (by Regression Imputation) compares the values for complete cases. The case with the closest match gets imputed into Y. This is similar in a way to hot deck procedures in that it has a 'donor' element, but enables the use of predictor variables when these are available. This method is suitable for

continuous data. The disadvantages are that the variance is compressed, and relationships between variables are distorted. Also a different model is required for each variable. This method is used for economic data (Little, 1988).

5.6.3 Random (Stochastic) Regression Imputation

Random (Stochastic) Regression Imputation is really just regression imputation with an added random element, for greater accuracy of the variance estimate. This method works best if the variable of interest is correlated with other variables in the data. The random residual term may be added in a variety of ways. One could add a random N $(0,\sigma)$ from complete data. Or one could use residuals from the complete data and 'borrow' a case for imputation of the residual (this avoids the normal assumption). Or thirdly if the response is there, but the predictor is missing, match a complete case with a missing case that has an actual value of Y very similar to the predicted value of Y in the missing case. The residual is then 'borrowed' and 'given' to the imputed case. This avoids the (regression) assumption of no correlation between variables, and allows the variables to be correlated.

If logistic, then use a value of Y with a probability equal to the predicted probability. This time the variances of survey means are inflated. This is not a bad thing, as this reflects the uncertainty of the missing data. However the relationships between the variable of interest and the variables not used in the imputation model, will be distorted. A different model is required for each variable, but still this is an improvement on straight regression imputation.

5.6.4 Logistic Regression Imputation

If the data is binary data, then Binary Logistic Regression can be used, imputing the value of Y giving the highest predicted probability. It is possible to use nominal or ordinal logistic regression but in these cases (where there are more than two levels) it is better to use alternative methods (David, Little, Samuhel, Triest, 1986).

5.7 Other single imputation methods Overview

This is a new and emerging field, using methods based on the data, as opposed to methods based on distributional assumptions.

5.7.1 Nearest Neighbour Imputation

Nearest neighbour imputation relies on being able to identify the distance between two responses based on distance measure. The imputation is only possible with item nonresponse as some of the data must be complete to calculate the distances. The distances may be calculated on external information or other covariates, within the dataset itself. The data needs to be continuous. To use this method, first the data needs to be standardised, preventing unequal contributions, as in much the same way principal components analysis needs standardising. If extreme values are present ranks may be used.

The distance is then calculated from the sample unit with the missing value, to all other units with full information (or enough to calculate the difference). This may be done in various ways the most common of which are: 1 Euclidean distance, 2 Mahalanobis

distance, 3 A weighted distance using a prior to give some variables more prominence than others (a form of ridge regression).

There need to be several variables available on which to apply this method, to provide a close match. The proportion of missingness cannot be too high, as this will cause problems. Nearest neighbour methods are not suitable for categorical data.

This method requires considerable preparation of the data and can also distort distributional shapes (of categorical data). Nearest Neighbour Imputation maintains the relationships between the variables reasonably well, and requires data to be MAR (Vacek, Ashikaga, 1980).

For imputation the distance is calculated, using one of the methods above. Then the closest unit of interest is determined, which is then used to impute a value for the missing item. This is a classification technique. It is found in the software SNOB. (Wallace et al, 1998)

5.7.2 Neural Networks

Neural Networks are information systems that recognise patterns in the data, without any missing values, and then apply a solution to the incomplete data. Neural network methods can be applied where sufficient complete cases exist, and are a better method for categorical than numerical variables. Neural networking methods are not suitable for large numbers of missing values, as the shape of the distribution is not maintained.

Neural Networking is a three tiered system, with the lowest level being the inputs (of the variables). Next there is a hidden middle layer (with a pathway from each input to each part of the middle layer). Last there is a top layer (which has a pathway from each element of the hidden middle layer to each element of the top layer). In the case of missing data this layer consists of imputed outputs. Each pathway has an associated weight.

What actually happens is:

1 The dataset is divided into three parts. Two of these contain only complete cases and the third will contain the missing values (incomplete cases). The two complete case groups are used to develop the neural solution, which is then applied to the incomplete cases.
2 The data needs to be segmented (stratified) and then encoded into categorical data, and/or normalised.
3 Training and Validation is done involving complete cases, and develops an imputation strategy for each case. (The weights are altered as each case is considered). The validation set chooses the weighting for each case.
4 Cases are run through the final solution which have been developed, using the best weighting chosen by this procedure.

The number of cases required for training is large, if there are insufficient cases in each segment (stratum), then the solution is not worked out. For a training set there needs to be 10 (m+n) cases where m = number of inputs, and n = number of outputs. In the case of twenty inputs and one output, then 210 cases are required for training. However the greater the number of inputs per output the more accurate the solution is.

Neural networks have the added advantage of being able to work around missing data. If a node is missing, then the weighting on the surrounding nodes compensates and an alternative pathway(s) is (are) found. This is really only suitable for large quantities of data, and can get rather unwieldy rather quickly. In the field of missing data analysis, neural network methods are still unproven, as research is ongoing, and more work is needed here. SPSS will perform this type of analysis (Azuage et al 1999).

5.8 Model Based Imputation Methods Overview

This includes single imputation involving the EM algorithm, and all types of multiple imputation. In all cases a sensitivity analysis should be applied, to identify the type of missingness. Is the missingness mechanism MCAR? If cases of systematic differences exist, then the missingness mechanism is not truly MCAR, but may be MAR. If MAR on Z (often a demographic variable such as age, sex, ethnicity and so on), then this is commonly known as informatively missing and may be continuous or categorical. In this case the Model based methods may be used the 'Z' variable is included in the 'imputation' model. If NMAR, then the missingness mechanism needs to be modelled, and if this can be done, then again the problem is converted to one of MAR.

Multiple imputation imputes more than one value for each missing value. This gets around the problem of underestimating the variance of the variable, as a range of values are put in place of the missing value. Multiple Imputation (MI) can be achieved by a variety of methods, and can be analysed according to complete data

methods, and then the estimates combined using Rubin's Rules (see chapter 3.7.1).

Efficiency of Multiple Imputation: Rubin's efficiency of an estimator, based on m imputations, where γ is the fraction of missingness. E is given as a percentage. (Rubin, 1987)

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

Figure 5.1. Efficiency of Imputation Table

m \ γ	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

The goals of multiple imputation are to reflect the uncertainty due to missing data, preserve important aspects of the data distributions, and preserve the relationships in the data between the variables (Schafer, 1997).

5.8.1 EM Based Single Imputation.

EM Based Single Imputation is a method for obtaining Maximum Likelihood estimates from the observed data of the underlying model parameters that are efficient. This works by having a two step procedure, firstly an E(xpectation) step, which predicts $Y_{j\ mis}$ from $Y_{j\ obs}$ and θ_j (θ_j = the underlying model parameter). The second part is the M(aximisation) step, which estimates θ_{j+1} from $Y_{j+1\ obs}$ and $Y_{j+1\ mis}$. These two steps are iterated until convergence.

Draws are then made on the ML estimate given the other variables in the model, and then an estimate is imputed into Y_{mis}. This gives reasonably accurate estimates, but no standard errors. If the rate of missingness is high, then convergence is slow (Dempster, Laird, Rubin, 1977).

5.8.2 Multiple Imputation - Bayesian

The advantage of this method particularly is that m data sets are produced. Then the within dataset variance can be compared with the between dataset variance, and therefore give an estimate of the variability due to imputation. One method is to use an implicit method based on propensity scores (Rosenbaum, Rubin, 1983). An approximate Bayesian Bootstrap (Rubin 1981) is used to generate imputations. These multiple imputations are independent 'draws' from a posterior predictive distribution for missing data given the observed data. Propensity scores are the calculation that a subject would have a missing value in the variable in question, that is a conditional probability of missingness given the observed covariates.

5.8.3 Multiple Imputation MCMC based - Bayesian

- Normal: Data is augmented using a non-informative normal inverted Wishart prior. (Of course a proper prior may also be used.) A Markov chain is simulated, by randomly I(mputing) - (the I step) the missing data given the observed data and the current parameter value. Then there is a draw of the P(osterior) - (the P step) distribution of the parameter. This process is repeated to give m draws, to fill in m datasets, therefore multiply imputing.

- Categorical: This is a MCMC method of I-P steps (see above) for fitting a saturated multinomial model, which returns an array of cell probabilities. A Dirichlet prior may be used here, or a standard non-informative one. Also a conditional method fits a log-linear model which also returns an array of cell probabilities.

- Mixed: The general location model is fitted here using the multinomial model for the categorical part, and the multivariate normal for the continuous part. The appropriate Dirichlet, and Wishart priors may be used, although more commonly a non-informative prior is used. This method needs MANOVA type constraints on the cell means, and returns a parameter list, from which to make draws for imputation.

- Longitudinal: Here a multivariate linear mixed model is fitted, to give multiple imputation of missing covariates. This model allows for random effects to be modelled.

5.8.4 Multiple Imputation - Conditional

The rationale behind the development of this is that, very often in data sets with large numbers of variables, the missingness not only affects the response variable, but also the predictor variables. Current methods of Multiple Imputation presume that the predictor variables used to impute the missing data are all present. Of course the reality is that very often there is missingness in the predictor variables as well. It is well known that single imputation, (particularly regression imputation) will provide quite a reasonable estimate for a missing value) but if this is then taken as an observed value and

analysed by normal software, then the variance estimate will be severely under-represented if the missingness is not MCAR.

However one could ask how valid it is to use one variable to predict the imputed value in another, and then use that predicted value in predicting the imputed value in the original variable? that is circularities, Y_1 imputed given Y_2 and then Y_2 given Y_1. So the order in which imputation is done is highly crucial, and it is problematic how this would affect longitudinal surveys, as these must be treated as repeated measures. When imputed data is transformed, how reasonable is this? That is can log Y_i or discretised Y_i from continuous Y_i be used to predict X_i ? Especially if X_i was itself predicted by Y_i and was transformed.

The imputation model can be non-linear or there may be an interaction term present. The MICE (Multivariate Imputation by Chained Equation, see chapter 6.4.5) program specifies a conditional distribution given the other variables using polytomous regression for categorical data and linear regression for numerical data - this uses a Gibbs Sampler from these conditionals.

The method used is multiple imputation, multiple analysis and then pooling of final results - as per Rubin. Each missing value is replaced by (m) values, and the differences in these m data sets preserve the correct variance estimates. Whilst the procedure is simple for combining the estimates, the procedure for generating the multiple imputations in the first place is not simple, and if the imputation procedure is 'uncongenial' with the underlying structure of the survey, or the data for that matter, then problems can exist. In fact, a badly imputed data set can give more biased results than case deletion would have.

What is needed is an appropriate predictive distribution from which the imputations are taken, often requiring iterative methods. For the MICE method of imputation, all that is required is the specification of the conditional distribution for the missing data in each incomplete variable. This algorithm assumes that the data are in fact missing at random and conditions each time on the most recently drawn values of the other variables. Using this method, the conditional models can be specified directly.

There is no need to choose a multivariate model for the entire data set, although it is assumed to exist. The imputed values are generated by iteratively sampling from the conditional distributions, so making a series of univariate problems from a multivariate one (This approach is also known as regression switching or variable by variable Imputation). However if the two conditional distributions $P[X_1|X_2]$ and $P[X_2|X_1]$ are incompatible then $P[X_1X_2]$ does not exist, and the algorithm will not converge.

If one is going to the trouble of imputing a dataset, then one would like the resulting data to be able to be used for any purpose the analyst might chose. For this to be so, then all available data needs to be used in the imputation process. However if there are hundreds of variables in a data set, it is not always appropriate to use all potential predictors, as problems of multicollinearity, etc will arise. Often only the best fifteen predictors, say, may be used. All variables contained in the complete data model should be used, including those in the response model.

If factors are known that have contributed to the missing data, for example stratification, then these should be included. Then include variables where the distributions for nonresponse and response

groups are different. Also include variables which explain large amounts of the variance of the targeted variable of interest.

5.8.5 Multiple imputation for GEE (Generalised Estimating Equations)

Xie, Paik (1997) discuss methods for dealing with analysis of correlated outcomes, modelled using the first two moments of the outcomes, and the working correlation within individuals. When the missingness mechanism is MAR, case deletion, or imputing sample means is shown to produce incorrect GEE estimates. Also suggested is a weighted GEE model, where the weight is the inverse of the probability of the covariate being observed.

This is one method of dealing with this: an alternative suggested by Xie, Paik (1997) is a multiple imputation approach. The values are imputed into the data set, which is then analysed as complete data. The imputation step is done using Rubin's Bayesian bootstrap. The analysis phase uses standard GEE techniques. This yields unbiased parameter estimates and 'proper' variance estimates. The SAS GEE macro can easily do this analysis. However the data first needs to be sorted into a monotone pattern. Such a pattern is an advantage with most imputation algorithms.

5.8.6 MI for Case Control Studies

In clinical trials the missingness mechanism may well be intended, as some of the subjects are not given a particular treatment. Rubin (2000, Personal Communication) proposes that a type of multiple imputation be used which takes into account the structural missingness, as well as unobserved values (missing outcomes). Rubin defines the intention to treat estimand as the average causal

effect, and an indicator for the noncompliance. This is then modelled, and then draws are made from the posterior distribution, which are imputed into the dataset, using m draws for each missing value creating m datsets. These are analysed by standard methods for case control studies, and then the parameter estimates are combined, to give an overall estimate for each parameter of interest.

6 Software for Missing Data

This is a new and emerging field. The practicalities of dealing with missing data are some distance behind the methodology available in the literature.

6.1 Overview of Software Available

For the majority of the commercial software, the default method for dealing with unobserved values is case deletion. All packages will automatically use this for at least some of their analyses. In the case of some of the older software, some analyses require only complete case data: incomplete cases must be removed prior to running the analysis.

Then there is the specialist missing data software which are freely available for downloading. These include the 'gold' standard for Imputation, Joe Schafer's Norm, Cat, Mix, and Pan. NORM is available as a Windows stand-alone program, and all these programs are available as S-Plus subroutines. (see chapter 6.4) Other useful programs include Amelia (Gary King, Harvard), MICE (van Buuren, Leiden) IVEWARE, MDM (S-Plus), Transcan (S-Plus, Harrel), Multimix (Jorgensen and Hunt, Waikato). These are reviewed here. Chapter (9.2.1) gives a comparison of the methods involved.

Other specialised software exists, the most well known in missing data circles being Solas, as well as others such as BMDP (although this package actually contains other routines also) which has a missing data routine, and Microsoft's Dalsolution.

Structure of this chapter on software: In chapter 6.2 the general-purpose commercial statistical software is reviewed. In chapter 6.3 the commercial programs for handling missing data, and more general packages containing missing data routines. Chapter 6.4 contains a review of freely available software, and chapter 6.5 reviews other useful software.

6.2 Commercial Packages

Available in the major commercial packages is: Mean imputation, Median and Mode imputation is a reasonably simple procedure to do with any of the standard software. It is merely a matter of calculating the estimate, and then using a recode mechanism available in the software.

6.2.1 Minitab

Minitab mainly uses case deletion. However the user can self program mean, median or mode substitution. Also it would be possible to use Minitab for mean within category substitution, stochastic mean substitution, and certainly Discriminant analysis. Even time series methods (using forecasting) and Regression imputation, manually imputing fitted values, are possible. Minitab will fit Ordinal Logistic Regression, using M^{c}Cullagh's cutpoint model (M^{c}Cullagh, 1980). This is useful for ordinal data. Minitab also fits Nominal logistic regression, suitable for unordered categories. Binary Logistic Regression is also available in Minitab, as is it in the other major packages.

Certain routines within Minitab perform case deletion implicitly, that is you can pass columns with missing data to these routines, and

they will case delete and then run. Other parts of Minitab, especially multivariate methods, and letter value displays require the columns not to have any missing data. This is more trouble for the user.

6.2.2 SAS

SAS data step

In the SAS data step missing values are handled by various means.

1 Missover statement. This keeps in data with missing values, and keeps the data in the correct array. Effectively this means case deletion, and prevents the routine from reading that line (or part of that line).

2 Lostcard. This statement resynchronises data when SAS encounters a missing record in the data. Otherwise SAS only discovers a record is missing when it reaches the end of the dataset if lostcard is not there. Lostcard works iteratively by writing an error message to the log, and then rereading the next set of observations.

3 Missing: specifies an actual value or character to each missing value.

4 Stopover: stops processing when encountering missing values for that line so no new data are read, at all.

5 Flowover: specifies that the input statement uses the list that is input and reads past the end of present record until the end of the data. (default)

SAS will also do all that is available in Minitab.

Type III Sums of Squares (Proc GLM, Proc Mixed etc), will deal with unbalanced data, and designs. (Type I, II, III, IV sums of squares were, until recently, only produced as standard options in the SAS package). Type III sums of squares are able to deal with missing data, and give (more) accurate variance estimates, in the presence of missing data. They are the partial sums of squares, and cannot be calculated from comparing the model sums of squares from different models (full and reduced), and are particularly useful for unbalanced data. Some of the major procedures have additional options available. Specialised options within SAS include Proc Expand (for longitudinal data), which works by interpolation using an ARIMA model.

Proc Mixed can fit a mixture of models. Unbalanced longitudinal data may be analysed with Proc Mixed. Proc Mixed is a generalisation of Proc GLM. Proc GLM fits standard linear models, whilst Proc Mixed fits a wider class of mixed linear models. All the usual class, model, contrast, estimate and lsmeans statements are the same, but their random and repeated statements differ. Proc Mixed will only give Types I, II, and III sums of squares tests for fixed effects. In Proc GLM the effects of the random statement are still treated as fixed as far as the model fit is concerned, and are there only for the expected mean squares.

In the random statement in Proc Mixed, the random effects constituting the γ vector (refer to equation below) are in the fixed model. When the data structures contain repeated measures, the mixed model approach in Proc Mixed is more flexible, and more widely applicable than either univariate or multivariate tests. The Mixed Model Equations are:

$$\begin{bmatrix} X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\ Z'\hat{R}^{-1}X & Z'\hat{R}^{-1}Z+G^{-1} \end{bmatrix}\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix}=\begin{bmatrix} X'\hat{R}^{-1}y \\ Z'\hat{R}^{-1}y \end{bmatrix}$$

where $\hat{\beta}=\left(X'V^{-1}X\right)^{-}X'\hat{V}^{-1}y$ and $\hat{\gamma}=\hat{G}Z'\hat{V}^{-1}(y-X\beta)$

The mixed model approach provides a larger class of covariance structures, and a better mechanism for handling missing data (Wolfinger, Chang, 1995). Proc Mixed will estimate and test for linear combinations of fixed and random effects.

A Mixed Linear Model is a generalisation of standard GLM, generalised in that this will adequately model non-constant variance, and the data may be correlated. The assumptions underlying Proc Mixed analysis are that the data are normally distributed, the means are linear in terms of the parameters, and the variances and covariances are modelled in terms of a different set of means, exhibiting a structure matching one of those available in Proc Mixed. Proc Mixed uses Restricted Maximum Likelihood (also known as Residual Maximum Likelihood).

This is ideal in missing data analysis, as this method will accommodate data that are MAR. Proc Mixed constructs an objective function and maximises over all known parameters. For REML the maximum likelihood is given by:

$$\ell(G,R)=-\frac{1}{2}\ln|V|-\frac{1}{2}|X'V^{-1}X|-\frac{1}{2}r'V^{-1}-\frac{n-p}{2}\ln(2\pi)$$

where r = $y-X\left(X'V^{-1}X\right)^{-}X'V^{-1}y$ and p is the rank of X. Proc mixed actually minimises -2 times these functions using a ridge-stabilised Newton-Raphson algorithm. Lindstrom and Bates (1988) give a discussion on why this is to be preferred to the EM algorithm. Also Proc Mixed uses the 'Sweep' based algorithms which are also used in Norm, Cat, Mix and Pan. Proc Mixed does not delete missing

level combinations for random effects parameters, because linear combinations of the random effects are always estimable.

Proc Mixed uses profile likelihood, and will fit these where possible. The default for fixed effects is type III sums of squares (partial sums of squares). These are calculated by constructing an estimated hypothesis matrix L and computing the SS associated with the hypothesis: $L\beta = 0$. These are invariant to order. These sums of squares do not add up to the model SS.

SAS has a data mining procedure, which will deal with missing values by neural networking techniques.

6.2.3 S-PLUS

S-Plus generally defaults to case deletion, but this option can be unchecked.

A useful additional library (an add on to S-Plus) for dealing with missing values, directly, is the algorithms developed by the Joe Schafer, Dept. of Statistics, Pennsylvania State University. These are largely stochastic versions of the EM algorithm, called the I-P algorithm (See chapter 6.4.7). Currently there are 4 available: norm, cat, mix and pan. (See chapter 6.4).

S-Plus will in its standard procedures case delete, but will give the option of not deleting. Recently incorporated into S-Plus is the concept of type III sums of squares for unbalanced data.

6.2.4 Base SPSS (Data step)

Some of the procedures offer differentiation of the type of missingness, for example system missing, and up to three different types of user missing. This is an option in base SPSS. Base SPSS offers mean imputation, median imputation, mean of surrounding values and so on as part of its data manipulation prior to analysis. Generally this is for continuous variables.

SPSS offers a neural networking module 'Neural Connection' for Neural Network analysis of missing data. This is useful for very large data sets.

6.2.5 SPSS MVA

SPSS MVA handles missing data by

1: Describing the data, whether randomly missing, extreme value missing, or different types of cases.
2: Estimates the means, SD, covariances and correlation, using listwise, pairwise (displays counts of Pairwise complete data) regression or expectation-maximisation method
3: Imputes missing values with estimated values using EM or regression methods.

The description is different depending on the type of variable being considered.

For continuous data SPSS will display the standard description of the data, but for indicator variables, a further indicator is made for missingness.

- This is shown by Vach and Blettner (1991) to be biased.

- Listwise and Pairwise depend on the data being MCAR (Little 1988)
- If data are normal, mixed normal, or student's t then the EM algorithm is used to Impute missing values.

For Pairwise description, % mismatch is used, and cross tabs are used with categorical and indicator variables.

A table is displayed from each variable showing number and % missing for each category. SPSS determines its missing values as extreme high, extreme low, system missing, user missing.

User missing is in three separate categories; Even so if more than a certain number are missing (n: to be determined by user), then that variable is deleted.
- SPSS shows case by variable patterns of missing and extreme values.
- Regression Imputation is used, of which Mean Imputation is a special case, but Regression Imputation can add a random component to the regression estimates, residuals and normal/student variates or just mean Imputation.

6.2.6 Statistica

Statistica can fit models with missing data by using Type VI (this is NOT a dyslexic Type IV) Sums of Squares, an invention by Hocking, which averages over missing values. When using Anova, Statistica uses Type IV SS (same as SAS) as well as Type VI SS, and this deals with the problem as part of the analysis.

Type VI SS are used when a design is unbalanced with missing cells, This is Hocking's effective hypothesis decomposition. (Hocking 1985). Any effect is evaluated by testing its unique contribution to the prediction of the outcome, after all the other effects have been entered into the model, a bit like Type II SS, but allowing for unbalancedness in the way Type VI SS do.

Effective Hypothesis is given by the matrix equality $H_{oo}\ \mu_o = H_o$. If the rank H_{oo} is less than rank H, then only part of the original hypothesis is tested - so if these constraints are not satisfied, then the original hypothesis is not true, but the hypothesis of maximal rank satisfying that condition.

6.2.7 Systat

Systat fits missing values in time series, by using a Polynomial; and is now a part of SPSS.

6.2.8 Matlab

Matlab is not really a statistical package, however it is well known and used in mathematical circles. Data analysis is one of its features, but more particularly to the field of imputation, is its use in Data Mining (Neural networks)

6.3 Commercial Packages which are lesser known

6.3.1 BMDP:

BMDP has algorithms for Multiple Imputation.
The algorithm BMDP5V (Dixon, 1988) imputes the Maximum Likelihood estimate, based on the likelihood:

$$L(\theta | Y_{obs}) = \text{const} \times \prod_{i=1}^{n} \int p\left(y_{obs,i}, y_{mis,i}, \beta_i | X_i, \theta\right) dy_{mis,i} d\beta_i$$

which assumes an ignorable missing data mechanism. BMDP8D (Dixon,1988) contains an algorithm for testing the MCAR assumption using t-tests for location.

6.3.2 Dalsolution

Dalsolution is similar to the paired procedure in SPSS, where two sets are compared to see how different they are; i.e. do they obey the same 'rules', or more appropriately, what proportion of the events do? Dalsolution has two major tasks:
1: The Analysis of relationship between categories or values.
2: Construction of new categories or values from existing ones.
This is based on Determinancy Analysis (Chesnokov, 1982).

Determinancy Analysis finds rules hidden in data. It defines new categories and values to be used in the search for rules, finds and analyses qualitative factors, computes critical bounds of numeric factors, and computes relationships between values from data sets that aren't directly connected.

Determinancy analysis is based on the frequencies of joint occurrence, or non occurrence, and if a circle A is a subset of circle B , then it is completely enclosed by B and the rule =A→B .

There are two characteristics, Accuracy and Completeness:
If $A \subset B$, then $A \rightarrow B$ is accurate, and complete.
If $A \cap B$ then the completeness of the rule $A \rightarrow B$ is equal to the accuracy of $B \rightarrow A$. That is completeness is defined as occurrences of A among occurrences of B, and the completeness rule is the accuracy rule with the arrow switched. This conditional approach characterises Dalsolution.

It differs from Regression Analysis by using conditional probability, as opposed to dependent and independent variables. It would appear to be a valid option, particularly in the case of non-numeric variables. The makers claim that regression analysis provides an approximate solution to forecasting, whilst DA provides an accurate one (again based on conditional probability). For further details the reader is referred to their website www.dalsolution.com. Dalsolution is relatively expensive at around $ US 1700.

6.3.3 Solas

Solas is a specialist software package for imputation. Solas offers a choice of mean imputation, Hotdeck imputation, last value carried forward, and multiple imputation.

Mean imputation has the advantage that it allows complete case methods, is easy to use and is useful when the percentage missing is small (<1%), or only means and totals are required. The disadvantages are the variance is underestimated, and the relationships are not maintained between variables.

Hotdeck imputation allows for complete case data analysis, and is better than mean imputation for maintaining distributions and

associations. Its disadvantages are it assumes an ignorable nonresponse, and is 'improper' (it does not restore the sampling variability).

'Last value carried forward' is a longitudinal technique that allows for complete case data analysis, and is easy to understand and implement. LVCF is no good when trends are present, and is biased when there is a differential drop out rate.

Multiple Imputation allows for complete data case analysis, and can be use for repeated measures (via the construction of longitudinal variables), and singly observed data. MI is 'proper' in the sense of Rubin, as it restores sampling variability. The disadvantage is that it has the confusion of multiple data sheets. Solas multiply imputes by first assigning propensity scores, then filling in the missing values by a Bayesian bootstrap, using a logistic regression. If this fails to converge, then the propensity scores are used.

Solas works by using the categorical data as a predictor, but will not impute into categorical variables per se. A big advantage of the Solas approach to MI is that the gathering of variables into a longitudinal arrangement is encouraged, and Solas allows complete case data to be used as predictors for missing values.

A disadvantage is that the first version of Solas did not actually allow one to export the multiply imputed data into other statistical packages for analysis. This has been allowed in the second edition. Another improvement in version 2.0 is that propensity scores (the default used if the logistic model does not converge) is not just applied blindly as in the first edition, but as an option.

6.4 Specialist Freeware Missing Data Packages

6.4.1 Amelia

Gary King's software for missing data is along similar lines to 'NORM' in that it is a stand alone windows program, or may be used with the mathematical software Gauss. It claims to have faster more efficient algorithms to run it. Without Gauss software it is difficult to run.

6.4.2 Cat

Cat is for categorical data. This is based on the multinomial distribution and is suitable for nominal data - the Estimation Maximisation algorithm is used here. The EM algorithm works iteratively by first using the value of θ given by the observed data. The algorithm then checks for missingness patterns, and if the pattern has no missing values, then the observed counts are added in. Otherwise the expected values are calculated and added in, then these are divided by n to give a new value of θ. Random zeros in the data may produce an estimate on the boundary of θ, or make θ unable to be estimated, so therefore the ML estimate will not be unique, but rather maximised along a ridge, with the EM algorithm converging to different stationary values depending on the starting value (Fuchs, 1982). Where the cells are empty because of a structural event being logically impossible, these should be omitted.

However during the E-step, the predictive mean is estimated, based on the observed data and the M-step is where θ is re-estimated from the predictive mean of x. Draws are then made from the posterior distribution.

6.4.3 IVEWARE

IVEWARE is SAS callable software that is used for complex survey designs with stratification and clustering. This software is suitable for providing means, proportions, and differences between subgroups, linear contrasts and also Multiple Imputation for missing data. It is intended to be general-purpose software for analysing survey data. Variance estimates are by Jack-knife or Taylor series approaches. It requires SAS 6.12 or higher.

This software is freely available at: http://www.isr.umich.edu/src/smp/ive/. Multiple Imputation is performed using a multivariate sequential regression approach to imputing missing values. Also available is PSTABLE which estimates means, proportions, subgroup differences, contrasts, linear combinations, using a Taylor series approach to obtain variance estimates under specified complex sample designs. Another module is Jackknif for fitting Linear, Logistic, Poisson and Polytomous regression models for data (as a result of the complex survey design). The Jack-knife approach is used to estimate variances. As yet this is only a beta version. (Raghunathan, et al 2000).

6.4.4 MDM

This is in the S-plus library MDM (Missing Data Machine). It includes various missing data routines. MICE started off as part of MDM, but is now its own library. (Brand, 1999)

6.4.5 MICE

Available in the S-Plus Library MDM is MICE, which creates mids, a multiply imputed data set. Mids are in turn input to produce mira, multiply imputed repeated analyses. These are input and then pooled to give mipo, multiply imputed-pooled results. Mice works well when the overall % of missingness is not too large, when the assumption MAR holds true. It also works well when the underlying assumption is multivariate normal. Monitoring can be done (for each incomplete variable) by checking on the mean and variance of the imputations.

6.4.6 MIX

MIX is for mixed data methods and is based on the general location model.

Again the EM algorithm is used for data augmentation, converted to an I step - using a random draw of (T_1, T_2, T_3) from the predictive data, and assumed value of θ . This obtains the E-Step, which then cycles through looking for missing data in the observed variables, and then each is drawn in turn from their predictive distribution.

The accumulated data is stored in (T_1, T_2, T_3) and then the p-step (the equivalent of the M step in EM) proceeds to give a new value of θ, given $(T_1, T2, T_3)$.

6.4.7 NORM

Norm is for continuous data, this deals with imputing values into normal data. Here instead of the expectation maximisation algorithm being used, this is replaced by a stochastic I and P step. The I step simulates the actual missing values, (Imputes random vectors), and the P-Step is a simulation of the normal inverted Wishart (central) distribution which is based on a multivariate normal.

6.4.8 OSWALD

OSWALD is an S-PLUS library designed primarily for the analysis of longitudinal data. This has been achieved by programming many of the standard S-Plus functions in an appropriate manner for longitudinal data. Missing values are a very common problem with longitudinal data.

Missing values are coded as NA's as all S-PLUS Routines do, and the na.omit option is present, although not recommended by the author of this library. These routines are designed to accept missing values in most cases, often by ignoring them. PCMID and plot.ldframe are specifically designed to handle missing values. In time series plots OSWALD will show bounded missing values by a dashed line, and dropouts by an 'x'. This is a very good way of graphically showing whether or not the dropout is informative (NMAR). Specifically OSWALD will model informative dropout for continuous longitudinal data.

In the case of completely random dropout (MCAR), then the Logistic Linear model is used. (From the OSWALD help files).

$$\text{Logit}\,[P_k\,(Y_1, Y_2, \ldots\ldots Y_k)] = \theta_0 + \theta_1 Y_k + \sum_{j=2}^{q} \theta_j y_{k+1-j}$$

where P_k is the conditional probability of a subject dropping out at t_k given the previous history $Y_1, Y_2, \ldots Y_{k-1}$ and Y_k (the current measurement of Y, say at point k). θ_o is the baseline (logit) dropout at any time and $\theta = (\theta_1, \theta_2, \ldots . \theta_q)$ (a vector of parameters relating to the dropout process to the unobserved (θ_1)), and previously responses. If $\forall$ j, all $\theta_j = 0$, then the dropout mechanism is MCAR. For Random Dropout (MAR), $\theta_1 = 0$ (with θ_j non zero for some $2 \leq j \leq k$), and assuming $P_k = 0$, $\forall$ $k < q$.

$$P_k = \frac{e^{\theta_0}}{\left(1 + e^{\theta_0}\right)}$$

For NMAR (Informative Dropout), it is merely a case of adding a set of initial estimates for the dropout parameters θ_i :

PCMID (formula, c,(ν^2, τ^2, ϕ), c(θ_1, θ_2,...),θ_0).

If any initial estimates are zero, they will be fixed at zero during estimation procedure. If $\theta_1 = 0$, this is the same as fitting a random dropout model instead of informative dropout (Diggle and Kenward, 1994).

The Logistic model may include covariates in the constant term, thus extending the model. PCMID involves strong assumptions for the missingness model:

- It provides a method of testing for MCAR, MAR, or NMAR by fitting hierarchical models.
- The need to check the robustness of model estimates and inferences with respect to model assumptions. (Dobson, 1998)

6.4.9 PAN

Pan is software written to deal with clustering, and repeated measures (although not together), and is based on the Generalised Linear Mixed Model.

6.4.10 TRANSCAN

The S-plus Transcan function uses a different approach, that of cubic splines regression to impute each incomplete variable, given all other variables (Alzola and Harrell 1999). The procedure assumes a multivariate normal probability distribution, and whilst it is easy to use, the question is "does it stand up theoretically?" (What is the theoretical background?) "Is it 'proper' [Bayesian] in the sense of Rubin?" (van Buuren et al, 1999) [Proper Imputation occurs when the prior distribution (of missing and observed values) is known and bounded, whereas with improper imputation the distribution of missing values is unknown and infinite].

6.5 Other Packages which may be Useful

6.5.1 MULTIMIX

MULTIMIX: (Hunt, 1996; Hunt and Jorgensen, 1999) This is another package based on mixture models, and clustering. The missing values 'add-on' is currently under construction. MULTIMIX is essentially Cluster analysis or 'Unsupervised Learning'. The sample is partitioned into groups, so that members are as similar as possible.

It differs from discriminant analysis in that members are clustered from within a population, rather than assuming they are from separate populations, and classified according to the population they come from. All populations can be put into clusters, one extreme being to classify all into one group, and another is to allocate each into its own group.

MULTIMIX uses an hierarchical algorithm to achieve this, using agglomerative techniques. This is displayed as a dendogram. The user needs to decide where to cut the dendogram, and what form of linkage to use, and what form of proximity to use. This method works best with small datasets, can accommodate (a) categorical data (multinomial model), (b) continuous data (Normal model) or (c) mixed (categorical and continuous). The latter is called the general location model (multinomial and normal) (Krzanowski, 1988).

This is where a separate model is fitted for each of the cells of the categorical contingency table. MULTIMIX maximises the complete data loglikelihood for each subvector in the partition. This can be used to give estimates of parameters and ultimately multiple imputation.

6.5.2 SNOB

This is a mixture models approach, developed by Chris Wallace, and David Dowe (1998)- using Multinomial, or Bernoulli distributions. It uses the minimum length message principle to do mixture modelling. SNOB models a statistical distribution by a mixture (or weighted mixture) of other distributions. It deals with normal/Gaussian distributions or discrete (Bernoulli or categorical)

distributions, Poisson Distributions, Von Mises Circular Distributions, and Missing data. Bayesian methods and minimum message length are used to deal with missing data.

Snob is a program for unsupervised classification of multivariate data (clustering) using minimum message length to decide upon the best classification. It can be used for hundreds of cases, but only tens of attributes, which can be continuous (real valued) or discrete (categorical). Currently it is not commercialised.

SNOB uses Multi-state attributes. This is based on relative probabilities of occurrence in each class. Each attribute is labelled to specify the state of that attribute, with a value, which will nominate the appropriate state. Descriptions are needed for the state space, as well as the attribute itself. The number in a particular class t having a value m for a multistate attribute d is n[m, d, t] with a relative frequency of occurrence of this state being:

$$p[m, d, t] = (n[m, d, t] + 1/n\ [d, t] + m[d])$$

where m[d] is the number of states possible for attribute d, and n[d,t] is the number of members of class t having known values for attribute d, and n = frequency of observations within each class.

The total message length is minimised by choosing a hypothesis to maximise the probability distribution. Inference from minimum message length is used as a classification technique.

The original algorithm 'SNOB' gave inconsistent estimates of the parameters for discrete data, but this has been largely fixed in SNOB 2. Data is described as a message having two parts. The first describes the hypothesis being studied, the second states the

actual data, using a binary string to describe that which would be true if the hypothesis being tested were true.

If a prior distribution is specified over a set of possible theories, (and presumably non-redundant) then the part one of the code is efficient. Use of the non-redundant code for the expression of the underlying theory is the equivalent of the assumption of a prior distribution. The shortest explanation is the preferred one (Minimum Message Length). If there is no known prior probability, then Bayes rule is applied - choosing the theory which gives the shortest explanation, that is, that of the highest posterior probability.

The value of precision to which an estimated parameter vector is rounded is proportional to the reciprocal of the square root of the derivative of the log function. If the log likelihood has a narrow peak, then the estimate is stated with high precision. MML is a means of comparing different explanations of the same data, the difference in length being the log posterior odds ratio in discrete cases. The classification message is:

1: Part 1 (a) the number of cases, (b) the relative abundance of each class, (c) the distribution parameters of each attribute, (d) the class each individual is estimated to belong to.

2: Part 2 for each individual gives the values for the distribution of (c) in part 1 and for the class assigned in (d) in part 1. (Wallace, Dowe,1998)

7 Rules for Imputation

See the Appendix, for this information presented as a flowchart.

7.1 Imputation Strategies

Assuming imputation is chosen, questions to be asked are

1. Type of data? (That is, is the data continuous, categorical, mixed longitudinal, etc.)?
2. What is the missingness mechanism?
3. What is the extent of the missingness?
4. What is the required outcome, and is this variable essential to the analysis (i.e. the means totals, proportions or further analysis)(Quality of Imputation)
5. What is the size of the dataset?

Selecting a model based method for MAR techniques.
Firstly, the initial set of target imputation variables is selected. Then a set of predictor variables is chosen which are relevant to this analysis. Any covariate that is correlated with the target variable should be included in the imputation model.

The broad methods will be: (after deciding the type of data and missingness)

1. Can case deletion be used here? (Is the data Missing Completely at random) or can other MCAR methods be used. (Mean type imputation or hot decking)
2. If look up methods is available or appropriate, this is the most accurate form of imputation that there is.
3. Imputations can be derived by looking at the correlation between variables to decide on a good predictor model. (Missing at Random Methods)

4. Non Missing at Random Methods include modelling the missingness mechanism.

7.2 Type of Missingness: Is the missingness MCAR, MAR, NMAR?

7.2.1 Continuous Data, MCAR.

For MCAR, case deletion is always valid, but is it appropriate? If only means and totals are required, no need to case delete, but probably mean imputation is adequate here. If the rate of missingness is greater than 50%, deletion of the entire variable needs to be considered, unless there are compelling reasons to keep that variable in the analysis, if the integrity if the dataset is to be maintained. If the rate of missingness is less than 1%, then all methods are satisfactory.

If the level of missingness is less than 5%, but greater than 1% mean imputation, or its closely associated techniques are adequate. If the missingness is greater than 5%, but less than 20%, care needs to be taken with mean imputation and case deletion, especially if there is a possibility that the MCAR assumption is invalid.

Even small levels of missingness can seriously distort estimates if this assumption does not hold. Greater than 20% missing and less than 50% missing (level of missingness is high). Needs great care to use a suitable form of imputation, certainly the MCAR assumption needs to be valid, (for case deletion and mean Imputation) and if there is any doubt about this, then the data needs to be assumed the weaker MAR.

7.2.2 Continuous Data MAR

Random Hot deck is an option here if the data is MAR, If the level of missingness is less than 1%, then case deletion is fine at this level, as is mean imputation. Above 1% to 5% missing, mean imputation is adequate only if means and totals are to be considered. If the requirement is for further analysis, and the missingness is less than 5%, the mean imputation family with added stochastic residual is preferred. Conditional mean matching may well be the method of choice here.

If MAR, and the level is of missingness is between 5%and 20%; that is moderately high, and then regression imputation may be valid, particularly if means and totals only are required. If in this situation the dataset is to be 'on used' for further public use- (by analysis by others than those doing imputation), then regression imputation may be used with a stochastic residual component.

Predictive means matching and hierarchical hot deck can also be used under these conditions. When the extent of missingness is greater than 20%, these simple single imputation methods are not recommended. Imputation via the EM algorithm would be the only single imputation method recommended. Multiple imputation is recommended in these circumstances, as inferences about estimates will always be proper, the sampling variance will be maintained.

7.2.3 Continuous data NMAR

For this kind of data, if the missingness level is less that 1% all types of imputation except case deletion are acceptable. If the

missingness is less than 5%, and the end use is means and totals only, then first the missingness mechanism needs to be modelled. If the dataset is large, use nearest neighbour, if small use MI (after modelling the missingness mechanism. When further analysis is required, then the missingness mechanism needs to be modelled, and then use MAR techniques.

For a missingness level of between 5% and 50%, regardless of end use, the missingness mechanism needs to be modelled, and then the appropriate MAR technique applied. If possible variable deletion should be applied, but also nearest neighbour techniques for large data sets are possible.

7.3 Categorical data.

For the purposes of imputation, categorical data is split into two parts, ordinal data, and binary and nominal data.

7.3.1 Ordinal data, MCAR.

For a missingness level of less than 1%, all methods are acceptable. When the extent of missingness is less than 20%, and the dataset is small, then person mean imputation may be used, as could variable mean imputation, case deletion, and mode imputation, look up methods. Other imputation methods could include multiple imputation by EM, or single imputation by EM, also Data augmentation by CAT. Large datasets could include the entire previous list (bar for the CAT), as well as neural networks, and hot decking. If the extent of missingness within the dataset is greater, say up to 50%, these methods can still be used, however care must be taken with the mean/median based methods as the extent of

missingness increases, to be sure that the MCAR assumption is in fact valid. If the missingness rate is greater than 50% for an individual variable, then the variable should be deleted unless there are very good reasons for not doing so, otherwise median imputation could be valid.

7.3.2 Ordinal data, MAR

Here essentially the same methods as MCAR may be used, except for case deletion, which may not be used. Another difference is that instead of mode/median imputation ordinal logistic regression is recommended (M^{c}Callagh's cutpoint model), Data Augmentation (CAT, or Norm (rounded)), may also be used. If the extent of missingness is greater than 50%, then the variable should be deleted, but if there is a good reason for keeping it use ordinal logistic regression imputation.

7.3.3 Ordinal data NMAR.

Firstly the end use of the data has to be examined, and if there are only means and totals required, the missingness mechanism is modelled, then use ordinal logistic regression imputation. If the datset is large the MAR technique may be used if the missingness is adequately modelled. If an individual variable is essential to the analysis, then use MAR techniques with the missingness mechanism modelled, (preferably with a low rate of missingness), first.

7.3.4 Binary, Nominal data MCAR

Again all methods are acceptable, if the missingness is less than 1%. Binary and nominal data differ from ordinal data in that the order of the levels is not important. All the types of imputation used for binary imputation are equal to that of ordinal imputation, with the exception of the logistic regression, here binary and its generalisation nominal logistic regression are used. Discriminant analysis is valid with binary missing data, as in nominal missing data.

7.3.5 Binary, Nominal MAR data

Imputation of this kind of data is again similar to the corresponding case for the ordinal data, except case deletion. Another difference is that instead of mode/median imputation binary or nominal logistic regression is recommended. Data Augmentation (CAT, or Norm (rounded)), may also be used. If the extent of missingness is greater than 50%, then the variable should be deleted, but if there is a good reason for keeping it use binary/nominal logistic regression imputation. Discriminant analysis is particularly useful here.

7.3.6 Binary Nominal NMAR

Firstly the end use of the data has to be examined, and if there are only means and totals required, the missingness mechanism is modelled, then use binary/nominal logistic regression imputation. If the datset is large, the MAR technique may be used if the missingness is adequately modelled. If the variable is essential to the analysis, then use MAR techniques with the missingness modelled first. For the NMAR case it is better to delete the variable

if the missingness mechanism cannot be modelled. If it can be modelled and the variable is essential, it is better to use binary/ nominal logistic regression, to investigate the MAR assumption, or how far this is deviated from.

7.4 Mixed data

7.4.1 Mixed data MCAR.

For mixed data (continuous and categorical) which is MCAR and the missingness rate is less than 1%, all imputation types are valid, even case deletion. However, for a missingness rate greater than this, either it is better to use a two step process, or a missing data routine geared for mixed data. The two step process involves firstly designing a categorical variable with different levels in a contingency table, and then fitting a separate model for each cell of the table, and each associated continuous bit. An imputation routine such as D A Mix would handle this (for small datasets). Nearest neighbour methods will handle this.

7.4.2 Mixed data MAR

These can be handled by a 1-step procedure such as DA MIX, or MICE, or IVEWARE. More frequently, a 2-step procedure is used. Using a categorical procedure, for the first step and a fitting a separate continuous model for each cell of the contingency table. The same constraints apply to each part of the table as do their categorical and continuous counterpart. See appendix for a list of suitable types of imputation.

7.4.3 Mixed data NMAR

Again, the missingness mechanism must be modelled for, then the appropriate MAR technique applied. If the end use is for further analysis, and the missingness is greater than 50% then variable delete.

7.5 Time series data

7.5.1 Time Series MCAR

Simple methods include case deletion, Last value carried forward, Proc expand, ARIMA modelling. Other methods include Multiple Imputation (Bayesian bootstrap) Period on Period Ratio, Year on Year Movement. If the dataset is very large, then the Kalman filter may be appropriate (state space models). Neural networks may be appropriate, as would nearest neighbour modelling. If small, the S-Plus routine DA PAN would be appropriate.

7.5.2 Time Series MAR

If the data set is small, then multiply impute, using DA PAN. Otherwise, multiple imputation using the Bayesian bootstrap, Last Value Carried Forwards.

If the dataset is large, then again the multiple imputation using the Bayesian Bootstrap, the last value carried forward Kalman Filter

(state space model), Neural Network and Nearest Neighbour imputation.

7.5.3 Time series NMAR

First model the missingness mechanism, then apply the appropriate MAR technique.

7.6 Other longitudinal studies (Repeated measures)

7.6.1 Repeated measures MCAR

Case Deletion, Last value carried forward and other MAR techniques.

7.6.2 Repeated Measures MAR

Multiple imputation, D.A. Pan. Multiple imputation Bayesian Bootstrap (longitudinal). Oswald (S-Plus) Proc Mixed (SAS) Kalman Filter (state space model) Neural Networks (categorical) Nearest Neighbour (continuous)

7.6.3 Repeated measures NMAR

First model the missingness mechanism, and then use MAR techniques

7.7 Panel data, and Clustered data

If the data is NMAR, and then the missingness mechanism must be first modelled for, then if the data set is small use Multiple Imputation Data Augmentation PAN.

7.8 Case control studies.

Assuming the data is not NMAR, then the data is conditioned on the assumption of compliance, to intention to treat. The posterior distribution is then modelled, and random draws made to multiply impute the data.

8 Some Approaches to Ordinal Categorical Data Imputation: Likert Data in Particular (a conjecture)

Consider the more trivial case for binary data (Ignorable Methodology):

For cross classified contingency tables often missing data could be considered a type of latent class, for example if a two by two contingency table, say at variable (C) shows a significant p-value, it may be that there is really no association between these two variables (For example the other variable being (A)), rather this is completely described by a third variable (B), which is unobserved, and the two in a contingency table may well be conditionally independent if a third variable is known. B may well be the missing data indicator, in the presence of which, A and C may be conditionally independent. When data is MAR, the addition of a missing data indicator is useful to explain confounding.

This could also be generalised to Likert data, in that a missing data indicator may explain, something that is unexplained by the observed variables.

One method of dealing with Likert data is Ordinal Logistic regression which can make use of a conditional form of continuous imputation using M^{c}Cullagh's cutpoint model in missing data routines.

Another possibility is that, the data be looked at from a purely categorical point of view, as a contingency table. Again, the trivial binary case is examined first.

An example is a two by two contingency table with variable classes, and each having two levels for simplicity. Let $\theta = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$, and $\theta_{ij} = P(a = i, c = j)$

If a third variable B is the missingness indicator which has two levels, observed and unobserved, then if conditionally independent, θ is determined by:

α = P (B=1),

where B_{a1}= P (A = 1| B = 1),

and B_{a2}= P (A = 2| B = 2),

and B_{c1}= P (C = 1| B = 1),

and B_{c2}= P (C = 2| B = 2).

Under the homogeneous model $B_{a1} = B_{a2} = B_a$, and $B_{c1}=B_{c2}=B_c$.

The homogeneity of missingness assumption (MCAR) means that the event probabilities θ_{ij} can be estimated using the observed data only. Suppose the total sample size is n, and suppose we observe x_{ij1} cases of A=i, C=j, B=1 (as B is observed)

Then the estimate of α is:

$$\hat{\alpha} = \frac{x_{111} + x_{121} + x_{211} + x_{221}}{n}.$$

The probability $\beta_a = \theta_{i+}$ can be estimated from β_{i+} =

$$\frac{x_{1+1}}{x_{++1}} = (x_{121} + x_{111})\frac{\hat{\alpha}}{n}$$

and similarly $\beta_c = \theta_{+2}$

The EM algorithm is used to estimate x= $\hat{x}_{ijk} = n\,\hat{\alpha}\,.\,\theta_{ijk}$, where θ_{ijk} is calculated from the current $\hat{\alpha}$ and $\hat{\beta}_a$ and $\hat{\beta}_c$ (E step), and then for the M step, new $\hat{\alpha}$ and $\hat{\beta}$'s are created from the $\hat{x}_{ijk}$.

When there is a two by two contingency table, and both variables are ordinal, as in Likert scales, a good model to describe this is the linear by linear association model, (Agresti, 1990). In terms of contingency tables, the linear by linear association model fits between the independence model (no association) and the saturated model (all possible interactions and effects included). The underlying assumption is that the points on the Likert scale are equally distanced.

As with all missing data analysis, sorting the variables into monotone order of missingness, is advantageous and speeds up calculations. When the data are missing completely at random, the categorical model is the multinomial model:

$$x|\theta \sim m(n,\theta)$$

where $x = (x_1, x_2, x_3,..,x_D)$, and θ is the parameter of interest, with probabilities $\theta = (\theta_1,\theta_2,\theta_3,.....,\theta_D)$, and $\theta:\theta_d \geq 0 \forall d$, and $\sum_{d=1}^{D},\theta_d = 1$. This is the underlying categorical model. However when the parameters (θ) are constrained, then the Loglinear model is used. Let $n_d = \log\theta_d$, $d = (1,2,...., D)$, and $n = (n_1,n_2,n_3,...,n_D)^T$, where $n = \text{Log } \theta$, $n_{ijk} = \log \theta_{ijk}$. $x_{+jk} = \sum_{i=1}^{I} x_{ijk}$ and $\theta_{I++} = \sum_{j=1}^{J}\sum_{k=1}^{K}\theta_{ijk}$, and the total sample size is $n = x_{+++}$.

$$n_{ijk} = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC},$$

$$\sum_{i=1}^{I}\lambda^A{}_i = 0,\ \sum_{i=1}^{I}\lambda_{ij}^{AB} = \sum_{j=1}^{J}\lambda_{ij}^{AB} = 0$$

$$n = \begin{bmatrix} n_{111} \\ n_{211} \\ n_{121} \\ n_{221} \\ n_{112} \\ n_{212} \\ n_{122} \\ n_{222} \end{bmatrix} = \text{m (the design matrix) } \lambda$$

$$\lambda = \left[\lambda_0, \lambda_i^A, \lambda_j^B, \lambda_k^C, \lambda_{ij}^{AB}, \lambda_{ik}^{AC}, \lambda_{jk}^{BC}, \lambda_{ijk}^{ABC}\right]^T, \text{ where}$$

$$\lambda_0 = -\log\left\{\sum_{ijk} e\left(\lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}\right)\right\}$$

For the linear by linear association model, (an appropriate model for Likert scales), Log $n_{ij} = n + \lambda_i^A + \lambda_j^B + \beta u_i v_j + \delta i$ (i =j), where u_i are the ordered scores for the I rows, and v_j the ordered scores for the J columns, assigned to the levels on an ordered scale. The relationship between the linear by linear association model and the independence model is that the independence model is a special case of the linear by linear association model, when $\beta = 0$. The likelihood equations are $\hat{n}_i = \eta_{i+}$, $\hat{n}_j = \eta_{+j}$, i=1:I, j=1:J,

$$\sum\sum u_i v_j \hat{n}_{ij} = \sum\sum u_i v_j \eta_{ij} \sum \hat{n}_{ij} = \sum \eta_{ij}$$

and the local log odds are uniformly of one sign. It can be shown that the loglinear model and logit models are in fact equivalent, and an easier way of showing the log likelihood of the linear by linear association model is to look at the logit form:

$$\log\left(\frac{\pi_{(j+1)|i}}{\pi_{j|i}}\right) = \log\left(\frac{n_{i,(j+1)}}{n_{ij}}\right) = \left(\lambda^y_{(j+1)} - \lambda^y_j\right) + \beta\left(v_{(j+1)} - v_j\right)u_i .$$

In this case Y is a response, and $\pi_{j|i}$ is the probability that y is in column j for observations I. In the case of unit spaced Likert data, this becomes $\log\left(\frac{\pi_{(j+1)|i}}{\pi_{j|i}}\right) = \alpha_j + \beta u_i$, where $\alpha_j = \lambda^y_{(j+1)} - \lambda^y_j$. The odds

of being in column J+1, as opposed to column J are multiplied by e^{β} (Agresti, 1990).

The missingness mechanism for a log linear model is assumed here to be ignorable, as the missingness mechanism is not modelled for. The modelling of nonignorable missingness needs to be the subject of further work, not addressed here.

For Bayesian iterative proportional fitting, the use of a Dirichlet conjugate prior is recommended, modified as a conditional Distribution.
Here ideally the Dirichlet Hyperparameters would be decomposed into hyperparameters α, into main effects and associate effects, essentially placing a prior distribution on a prior, (A second stage prior distribution on α, resulting in a mixture of Dirichlet priors. This is a constrained Dirichlet: log θ = m λ (the design matrix)), but in fact a non-informative prior is used.

Let Θ_M = the set of all parameters $\theta = \{\theta_y : y \in Y\}$ which satisfy the previous equation for the design matrix. If the prior density is $\pi(\theta) \propto \prod_{y \in Y} \theta_y^{\alpha_y - 1 + \beta u_i v_j} \text{for} \theta \in \Theta_M$, 0 elsewhere, then the complete data posterior is $P(\theta|x) \propto \prod_{y \in Y} \theta_y^{x_y + \alpha_y - 1 + \beta u_i v_j}$.

Alternatively, so that the cell means μ do not have to sum to any particular value, as in θ, The Poisson/Gamma representation is used. The cell means are Poisson counts, with a prior that is gamma distributed. The loglinear model is $\log\mu = M\lambda^*$ where
$\lambda^* = [\lambda^*_0, \lambda^A_1, \lambda^B_1, \lambda^C_1]^T$ and $\lambda^*_0 = \lambda_0 + \log\mu_{+++}$
The cell means for loglikelihood models are Poisson $x_{ijk}|\mu$ ~ Poisson $(\mu_{ijk}) + \beta u_i v_j$, so μ_{ijk} are a priori gamma variate, μ_{ijk} ~ $G(\alpha_{ijk})$.

For Iterative proportional fitting (ipf) the steps would be

1. The complete data likelihood is factorised into the cell probabilities

   $P(C = k|A = i, B = j) = \frac{\theta_{ijk}\beta u_i v_j}{\theta_{ij+}}$, when A and B are the categorical variables (Likert scales), and C is the missingness indicator, and the marginal probabilities for A and B are $P(A=i, B=j) = \theta_{ij+}$. This fixes $\left(\frac{\theta_{ijk}\beta u_i v_j}{\theta_{ij+}}\right)$ at its previous value, and then replaces (θ_{ij+}) by its maximum likelihood estimate.

2. The second step maximises $\left(\frac{\theta_{ijk}\beta u_i v_j}{\theta_{i+k}}\right)$.

3. The third step maximises $\left(\frac{\theta_{ijk}\beta u_i v_j}{\theta_{+jk}}\right)$.

Draws are taken from the complete data posterior $P(\theta|x) \propto \prod_{y \in Y} \theta_y^{x_y + \alpha_y - 1 + \beta u_i v_j}$.

(The starting values for θ should be inside the parameter space, care should be taken to not be 'on the edge'. (Applying a uniform non informative prior does work).

9 Analysis and Imputation of Data

9.1 Preparation of the data.

The data was prepared by first using the incomplete data and performing a simple available case regression analysis.

To illustrate this procedure for the nutrition data set the babies' birthweight was selected as a response variable, for a simple multiple regression. Stepwise regression was used to select the model, and from this ten possible predictor variables were chosen.

These ten were fed into a best subsets regression model. The best parsimonious model was a model with seven predictor variables, this also had the smallest Mallow's C_p, (A procedure which penalises for the number of predictor variables). If the C_p value is close to P the number of parameters, then this predicted model is not overfitted. (In this case C_p is less than P, but if the C_p is much bigger than P, then the regression model may be biased due to omitted variables).

An alternative within this is to look at the highest adjusted R^2 value, and the minimum s, which is the minimum square root of the mean squared error (known as the standard error about the regression line). Fortunately, all these criteria suggested the same best subsets model with seven predictor variables. This had a Mallows C_p of less than seven (the number of predictor variables), the highest adjusted R^2 (this means that the greatest amount of the variability was explained), and the least standard error about the regression line.

The variables chosen were a mix of different types. The babies' weight at six weeks post partum (This would not normally be included in a model for birthweight, as it is known only after the birthweight is known. However, it suits the purposes of this illustration to include it here since it is a predictor that has a reasonable level of missing data).

Other variables are, whether the mother intended to breast-feed the child prior to delivery, the number of pre-schoolers in the house, the extent of the mother's smoking at four months. Also included is whether the mother's partner was in paid work, the number of minutes the mother spent sleeping and lying down at the fourth month of pregnancy, and also the mother's weight at the seventh month of pregnancy. Four of these predictor variables are fully observed in this data set, but the other three are using available cases. For the purposes of illustration this same regression model is applied to each of the imputed data sets, and then the estimates are compared using Rubin's rules to combine multiply imputed data sets.

The imputed data sets consisted of SPSS MVA analysis, using EM imputation, regression imputation, mean imputation, and then Multiple Imputation using CAT and NORM, and MIX, also Multiple Imputation using Solas, and hotdeck, and group means, also from Solas.

Each of these methods came with their own set of imputation diagnostics, and difficulties with imputation. Given the size of the dataset, and the different types of variables, what was actually done was to use different imputation methods on 'bits' of the dataset, and then to reassemble the dataset. The problem here was to compare very different analyses with very different diagnostics. The

completed data sets were then analysed using standard multiple regression analysis, and the estimates then compared.

9.1.1 SPSS MVA Imputation

The imputation methods used by SPSS MVA were straight forward, and easy to use. Having said that, there were also problems in the way it was carried out. SPSS MVA uses listwise deletion in its regression imputation model, that is, dropping every case that has any missing data on any variable, (then imputing all missing values). The EM Imputation would appear to use pairwise deletion, for example, only dropping out cases that weren't completely observed for the model in question, (also known as available case imputation), for its imputation model. The regression imputation model would appear to be wasteful of the data. These methods are completely frequentist in nature, but neither involves multiple imputation. So whilst they would appear to be a better method than say mean imputation, the MVA add on would appear to not go far enough.

Imputation diagnostics include Little's MCAR test. For this particular test (based on the EM imputation), imputing only the variables used in the 'test' regression, the value was 4.701, on 10 d.f., which is tested against a χ^2 distribution, giving a p-value of 0.910. Therefore, in this case, pairwise deletion would be valid, although wasteful of the data.

However, when the complete data set (the continuous variables only) were imputed, Little's test failed, P-value = 0.030. This illustrates that, when dealing with a subset of the data one may

think MCAR is correct, but with more data, it becomes clear that MCAR is wrong.

It could be argued that, if the imputer is only interested in a subset of the data (say up to six weeks postpartum); it is perhaps best not to include for Little's test for MCAR other variables that are not going to be used in the regression model, (say data from twelve months postpartum). This could be so that a false MCAR is accepted, thus allowing for something such as case deletion. The analyst of the data may well argue that you do not need to include in the MCAR test variables that are not to be used in the intended analysis. However if an archive dataset is to be created (for use by others), then allowing for all missing data patterns in the variables is important.

Another problem here was that the SPSS MVA routine could only be used for the continuous variables. To get around this problem discriminant analysis (this needed to be done separately), was used to impute the categorical variables. (In this case there were three variables with missing values, each with only a few missing values. The missingness rate was <5%). Regression analysis was used to test the validity of the relationship between the categorical variable being discriminated and the continuous predictor variables. The only draw back here appeared to be that the predictor variables needed to be completely observed (for valid discriminant analysis to provide fitted values for imputation).

For the regression imputation, again the discriminant analysis was used for the categorical variables, and MVA regression analysis was used for the continuous variable. Options available were an added stochastic component error term, either a randomly picked residual was added to each case (hot deck), or a normal variate

residual, or a t-distributed residual. The choice would largely depend on the imputation model. In this case, the randomly picked residual was used.

9.1.2 Solas

Solas methods of single imputation used were: group means, and random hotdeck. These present no problems, the actual imputation is easy. If one wishes to do a simple linear regression, this is also no problem (with the completed data set). However when a multiple linear regression is attempted, (the final model, within Solas), the outcome is that the 'Y' variable is allowed to be selected, and the package then tries to place all other variables into the model as predictor variables. That is, Solas does not allow adequately for selection of the predictor variables.

So to perform an adequate analysis, the completed data set was exported, as an Excel spreadsheet. However this could not be opened in Minitab directly, the workaround was to open it in Excel, save again, and then open in Minitab. This enabled a regression analysis to be performed.

The Solas multiple imputation was performed, and again the same problem arose of not being able to perform a multiple regression, adequately. Also Solas does not allow the direct export (for example, 'save as') of its multiply imputed datasheets. They have to be copied and then saved as 'save datasheet', into Excel and again this then has to be opened and resaved before it is usable. Both of these bugs have been fixed in Solas 2. The Solas roll up editor window cannot be printed, or saved, which is rather irritating as this is useful information (It gives means and standard deviations for individual datasheets). The roll-up statistics window is able to be

printed, and this gives the total variance, as well as the between and within datasheet variance, but only for the imputed continuous data. The multiple imputation depends on a logistic model, but if this fails to converge, the default of propensity scores are used. Possibly if the convergence criterion were changed, then convergence may occur.

However it is a little disconcerting that only in the case of one variable did the actual logistic regression converge, showing that the imputation model was not really adequate. For the other seven variables the default of propensity scores were used. The output consisted of m = 5 datasets (this default can be changed), which were then (with difficulty), exported to Excel, resaved there, and then finally imported into Minitab for regression analysis.

9.1.3 S-Plus

S-Plus has various libraries for imputation. Transcan, and mice were not able to be reviewed here, as they were released too late, but the 'gold standard' Norm, and the related Cat, Mix, and Pan were used. This routine consists of data augmentation for missing data. In the case of the nutrition dataset, the dataset was carved up into pieces, and one bit used Norm (continuous variables), another used Cat (categorical variables), another used Mix (a combination of categorical and continuous) and the longitudinal variables used Pan.

Norm was straightforward, the EM algorithm was used to find parameter estimates, and then data augmentation was applied using the MCMC algorithm. The initial log-likelihood of the first estimate of the parameter was always the maximum, but was not too fast at converging (at around twenty iterations for twenty-seven

variables). Then about fifty steps were taken to ensure independent draws from the posterior distribution. After that the log-likelihood would oscillate around two different values, possibly as either the surface of the posterior distribution was very flat, or in fact multi-modal.

Taking one hundred steps between draws helped this problem somewhat. Another interesting feature of this routine is that the log posterior is always less than the log likelihood, reflecting the additional uncertainty. Another interesting diagnostic, not done here but possible would be to observe the autocorrelation function to assess the convergence of the series. Here ten data sets were produced. This routine is limited to thirty variables.

For the categorical data, the EM algorithm was initially used, but more than five variables become problematic, and convergence becomes very difficult. Again there is a problem using the unconstrained EM algorithm (the saturated model, multinomial), the loglikelihood oscillates, and again the surface is very likely multimodal. The number of iterations to convergence is large, >1000, and so it is unreasonable to use this with any degree of confidence. If the loglinear version ECM is used (a constrained version of EM), not the saturated model, but one preventing anything more than second order interactions, then by use of a conditional model, this is more reasonable, and will fit a model for five variables (no more than fifteen levels) within thirty iterations.

Again, 100 steps were found to be appropriate here. Independent draws were taken, but still the log posterior would oscillate, because of a multimodal surface. Ten datasheets were imputed. This routine should be used with care by the imputer.

Mix, in an ideal world would be the best possible way to deal with real data. It can take into account categorical and continuous data, by using the general location model. However, the dataset needs to be arranged with the categorical variables on the left, and the continuous part on the right hand side, and preferably all in monotone order of missingness.

Mix works reasonably well for small datasets, provided the contingency table is small. A separate model is fitted for each cell of the contingency table. Convergence for anything too large is slow, but not totally unreasonable, at around seventy iterations if the categorical part is constrained. This is slow in terms of time, A pentium II, 350 MHz, with 128 MB of RAM took around 12 hours to converge, using three categorical variables, and 27 continuous variables. This would be about the limit of the routines capability. Again, 10 data sheets were produced.

Pan was a little disappointing, as it can only deal with continuous longitudinal variables, and so was of little use in this particular data set.

(The following is a 'handy hint' for using the S-Plus based programs for multiple imputation). When transferring data into S-Plus via a spreadsheet such as Microsoft Excel, convert the missing values to an unreasonable figure such as 999 for categorical data, and 99999 for continuous data. This provides an entirely numeric data sheet for the transfer from Excel to S-Plus. This can then be transferred back to NA's within S-Plus.

Sometimes data is stored by the researcher as Excel files, and transferred to S-Plus electronically for analysis. The reason for this is that a tendency has been observed, with S-Plus in a networked situation, for the data to become corrupted during the import data

stage, from the Excel files. Cells with the non-numeric (in this case NA's - alpha characters), cause problems within S-Plus. By making the entire dataset numeric this problem is worked around. Strangely, this does not occur when S-Plus is installed directly into a standalone computer.

9.2 Analysis of data using Minitab

All of the different imputed datasets are analysed using Minitab's multiple linear regression. The different types of imputation tend to give differing estimates. Here are two tables of estimates:

1. The estimates of the β coefficients in the regression analysis of the completed data.

2. The estimates of the standard deviations of those same β estimates (from the regression analysis)

9.2.1 Results

Table 9.1. Estimates of coefficients under different Imputation schemes

Estimate	Allvalue	Regression	EM	SolasMI	Norm	G.Mean	Hotdeck
Constant	508.900	408.138	338.560	761.000	462.400	426.000	824.100
Number Preschoolers	116.670	118.612	103.537	115.770	107.620	111.530	115.190
Number Smoked 4	-146.310	-158.276	-151.850	-176.030	-146.510	-152.180	-204.470
Partner Paid Work	198.200	177.814	180.746	172.220	170.740	181.910	169.610
Babyfeed	120.330	116.482	119.626	104.900	112.520	107.940	95.810
Sleeplie4	-0.615	-0.590	-0.557	-0.711	-0.595	-0.647	-0.733
Weight7	4.697	5.834	5.139	7.787	6.101	7.004	7.344
Babyweight6	0.542	0.549	0.572	0.467	0.537	0.536	0.468

Looking at the estimates of the β coefficients the best performer (Best being the most consistent estimators, and given the MCAR

test result, the closest to the available case regression), would be Norm, followed closely by SPSS's EM Imputation, these give similar estimates. Ideally this would be tested by deleting different datasets under different missingness schemes, and then testing how accurate the estimates are to the known estimates. Solas MI is worryingly different from these two, particularly for the all important continuous variables (these were the variables being imputed). Regression imputation (SPSS MVA) would appear to perform well. The trivial group means and random hotdeck, give reasonably similar estimates to Solas MI, particularly with the imputed variables.

Here the allvalue regression analysis is performed in Minitab, using available case regression. The more trivial Group mean imputation, gives variance estimates that are slightly compressed compared with those of the allvalue regression, whilst the hotdeck estimates remain constant for the variance. Norm, SPSS EM imputation, and SPSS

Table 9.2. Standard deviations under different Imputation schemes.

Standard Deviation	Allvalue	Regression	EM	SolasMI	Norm	G.Mean	Hotdeck
Constant	354.200	332.893	330.005	360.620	329.470	354.400	358.500
Number Preschoolers	43.230	39.398	39.040	43.346	40.005	41.510	43.530
Number Smoked 4	61.990	58.487	57.764	60.268	55.532	57.630	60.680
Partner Paid Work	86.140	77.642	76.620	84.944	78.462	81.330	85.420
Babyfeed	40.560	38.327	37.786	40.994	37.780	39.250	41.100
Sleeplie4	0.311	0.287	0.284	0.314	0.289	0.302	0.313
Weight7	2.734	2.486	2.531	2.779	2.557	2.671	2.734
Babyweight6	0.049	0.046	0.046	0.049	0.045	0.050	0.049

regression imputation, give very similar estimates to each other, and one would suppose they would tend to be fairly accurate. Solas variance estimates are somewhat larger, possibly accounting for the extra uncertainty due to imputation, but very similar to the case deletion.

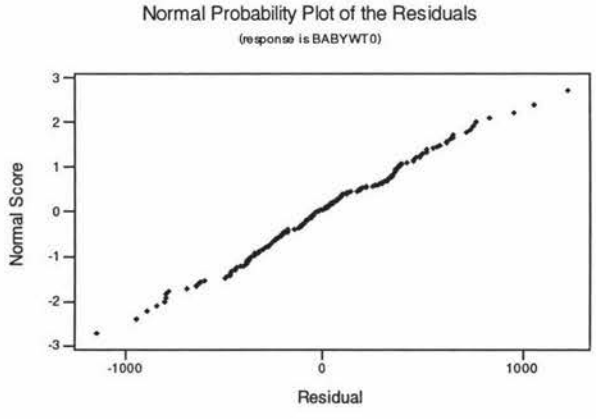
9.2.2 Validity of Imputations, and results.

The validity of some of these methods must be called into question. The methods requiring a MCAR assumption, appear to give very different results, casting doubt on the validity of MCAR for this dataset. Group means and Hotdeck appear very different to the available case analysis, but nearer to the Solas result.

The imputations requiring the more relaxed MAR, show that with the exception of Solas MI, the Norm, EM, and regression imputation schemes appear to be reasonably similar. Care is needed with the Solas MI imputation model, a better imputation model would surely improve the accuracy of the estimates.

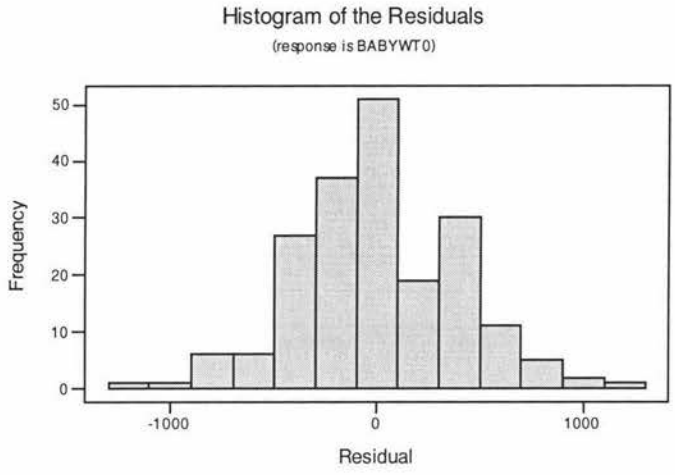
The regression analysis diagnostics showed that the assumptions the regressions were based on were valid, ie normality of the residuals, and constant variance, as shown by the next three diagnostic plots.

Figure 9.1. Normal probability plot of the residuals



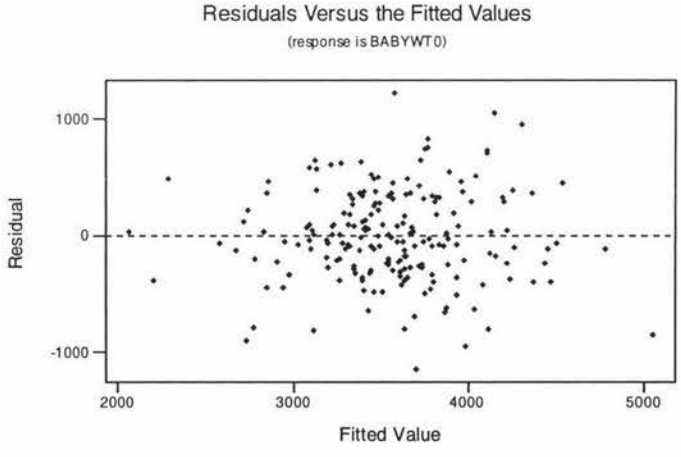
Here the normality plot shows that the residuals are normally distrbuted

Figure 9.2. Histogram of the residuals



The histogram of residuals is reasonably symmetric, not skewed.

Figure 9.3. Plot of residuals versus fitted values



The histogram of residuals versus fitted values.

9.3 Further Analysis

This same analysis needs to be run on other data, to be verified, as well as other analysis within the same dataset. A further improvement would appear to be the addition of an indicator for missingness in the final regression model. This would serve the purpose of showing whether the MAR assumption is valid. Time constraints do not allow for:

1. The MAR assumption to be investigated at this time,
2. Testing against known parameter estimates.
3. The investigation of between and within datasheet variance estimates. This is unfortunate as the ratio of these form an excellent indicator for just how efficient a (multiple) imputation is.

10 Conclusion

10.1 The Ethics of Imputation

Why Impute? Well the best reason for imputation has to be preservation of the data, that is not throwing away data, costly to collect, and biasing possible results into the bargain. Increasingly however, Survey (data) collection companies look into ways to make their businesses more effective, thereby reducing item nonresponse as much as possible. Unit nonresponse is already dealt to by means of stratification.

In certain European countries, the cost of collecting data is such that the researching company commissioning the survey, will demand a perfect dataset of the data Collection Company, and often will pay accordingly. More alarming is the fact that very often these companies are in fact government departments, collecting official statistics.

The experience in the Netherlands is that since the population census was made voluntary, participation by the general public has dropped to around fifty to sixty percent. Many older people are scared to reveal anything to officialdom. The net result is that with the demand for complete data, some very suspect methods are used. Voluntary business type surveys very often only have a response rate of around twenty percent. In the case of unit non-response, single response Hotdecking is very often the imputing option of choice. This will provide a nice clean rectangular dataset for the client to analyse in whatever way they like. Given sufficient predictor categories, this could well be valid, for the point estimates.

The variance estimates are another story altogether. It is well known that estimates of variance are very often subject to shrinkage if entirely deterministic methods are used. This has to be the major problem with 'Public use' datasets that are singly imputed. Multiple Imputation on the other hand is shied away from by those producing 'public use' datasets, as it is not popular with endusers, as the recombining of the estimates while straightforward, is perceived to be difficult.

To this end the businesses and government departments commissioning surveys are vocal about what they want, they want reliably imputed data, but they do not want the bother of more than one datasheet per dataset.

Many of the people making these kind of decisions are not the statistically minded. They tend to be managers and so on, who merely understand that they are paying for the data and so will call the shots about how they will receive it. Very often their tolerance to missing data is extremely low, and demand complete data.

How far does the statistical practitioner go with imputation of data? The commercial software gives warnings if the observed data is too low, SPSS MVA warns against the addition of residuals when the extent of missingness is greater than fifty percent. Solas warns against multiple imputation if the missingness rate is greater than forty per cent. The imputer needs to think very carefully about exactly what they are doing when considering the imputation of a variable which is less than fifty percent observed. Is the variable crucial to the analysis for some very good reason? If not then variable deletion needs to be considered for any further analysis.

A further concern is when the researcher needs to fill a quota (quota sampling) for say an ethnic minority. Maybe there would be a great temptation to impute values, for that ethnic group, especially if say the research grant depended upon filling a quota, and normal data collection methods proved impossible?

A further question posed in the nutrition dataset was that of imputing the data for the mother who miscarried during her pregnancy. The baby was never born (alive) so how appropriate is imputing the Baby's measures at say six weeks and twelve months?

Could all this perhaps be taken to extremes, and unscrupulous practitioners impute whole chunks of datasets?

As with all of these questions the integrity of the data is only as good as the integrity of the imputer, as the integrity of an analysis is only as good as the integrity of the analyst.

10.2 Conclusion

If a researcher has enough statistical knowledge to analyse their own data, and to interpret it, then they also have enough knowledge to impute their missing data. Packages such as SPSS MVA are very easy to use, and with the version 10 help files, even a sample interpretation is given.

The literature has exploded in this field in the past twelve months, with new and interesting developments taking place, many of which were presented at the conference on survey nonresponse, held in Portland, Oregon during October 1999. Here many ideas were

presented on how to prevent missing data, how to impute missing data, and how to produce valid variance estimates in the presence of imputation. Many of these are described in chapters two and three. Software for imputation has developed in the past year, largely spurred on by the release of Joe Schafer's Norm, Cat, Mix, and Pan, and catching up with the developments in the literature.

Having said that, in general it is better that the researcher have a little knowledge, on the subject. If imputing using one of the S-Plus libraries, or Solas, for multiple imputation, then great care needs to be taken with correctly specifying the imputation model.

The more straight forward imputations, Hotdeck, Groupmeans, regression and EM, still need some care to be taken with the model, that is selecting the type of residuals and so on. If all that is required is means and totals, then one of the more simple methods is appropriate. This is more fully discussed in chapter 7.

Chapter 7 guides the researcher through a series of questions designed to select an imputation method appropriate to that dataset.

It is unfortunate that time constraints did not allow the testing of MICE, or the assessment of one of the smaller packages, say MULTIMIX, or the analysis of complete data subjected to differing missing data regimes, and to explore the contribution of nearest neighbour techniques, and neural networks to this field. There is a great deal more which could have been done, particularly with this nutrition data. This is discussed in chapter 9.

Another lead to follow is that of the conjecture shown in chapter 8. The linear by linear association model is clearly the best log-linear

model for ordinal data, especially Likert scales. To put this in a missing data frame work and implement it properly (formally prove) with the genetic foods data set, would plug a hole in the literature. Much work is needed in this area. This is a very important field particularly in survey research (as every survey suffers from it to some degree), and there is no justification for ignoring it and case deleting as was done in the past.

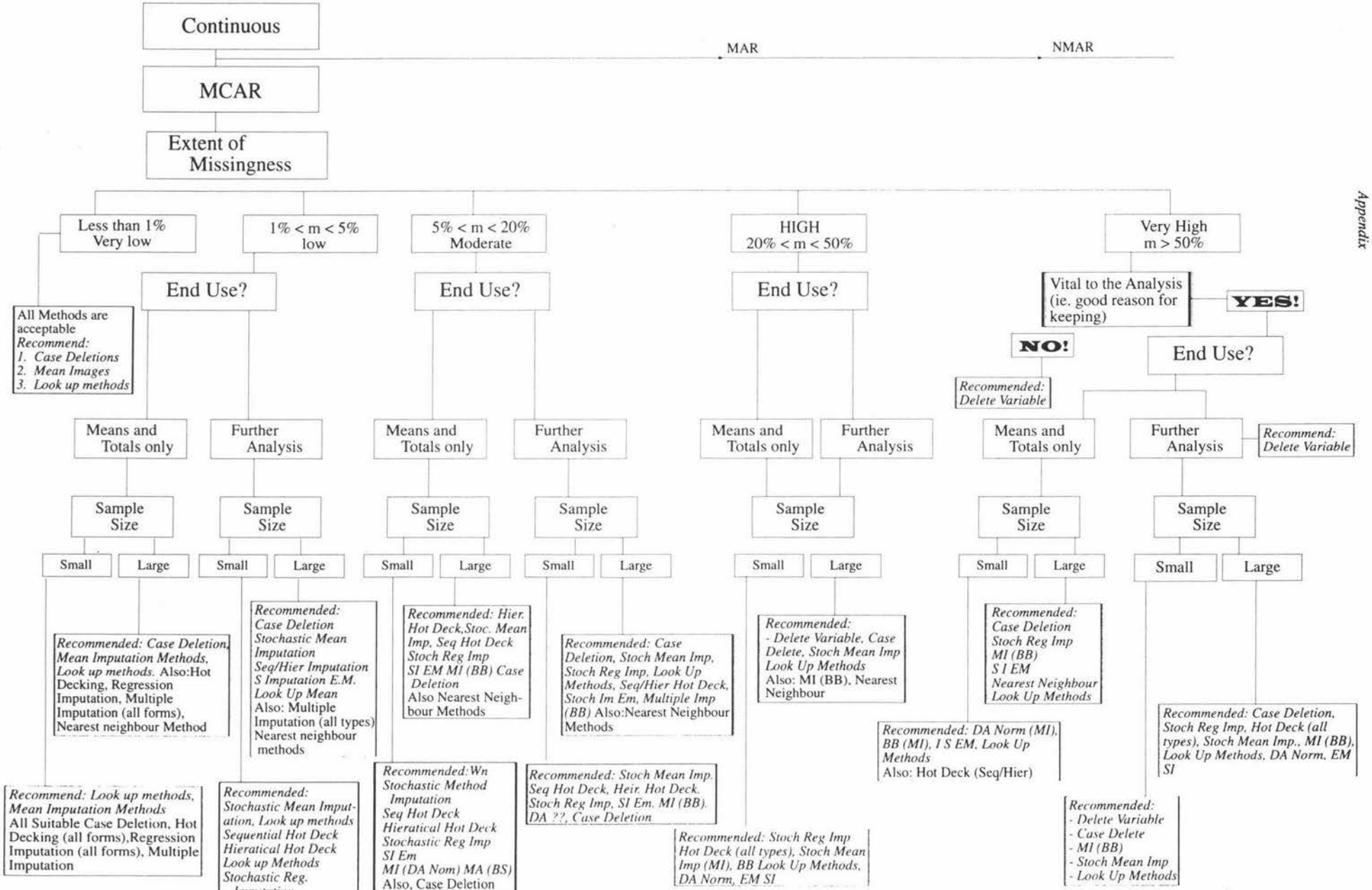
This has been a pleasure, to research, and a pity to let go such an interesting project.

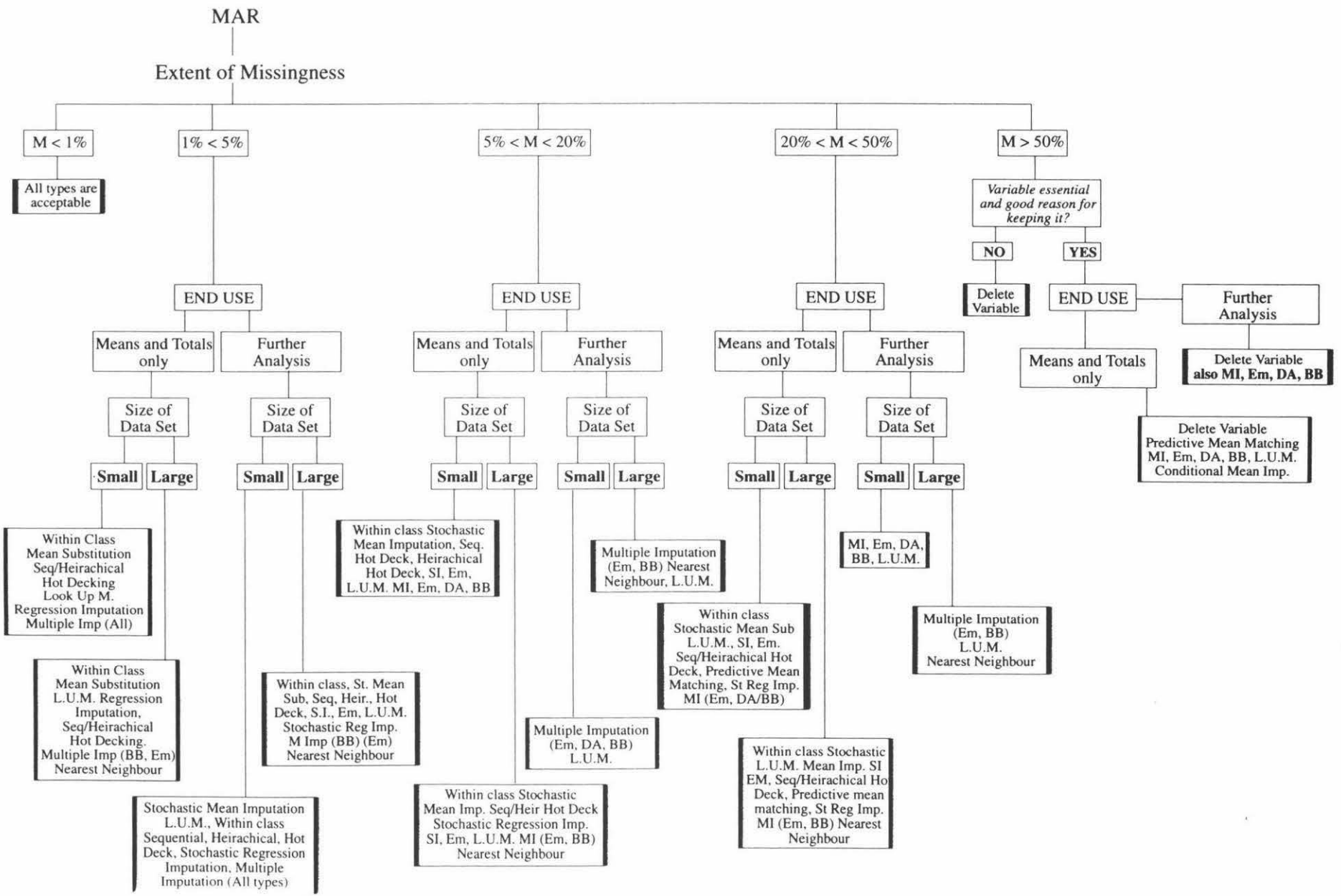
Appendix

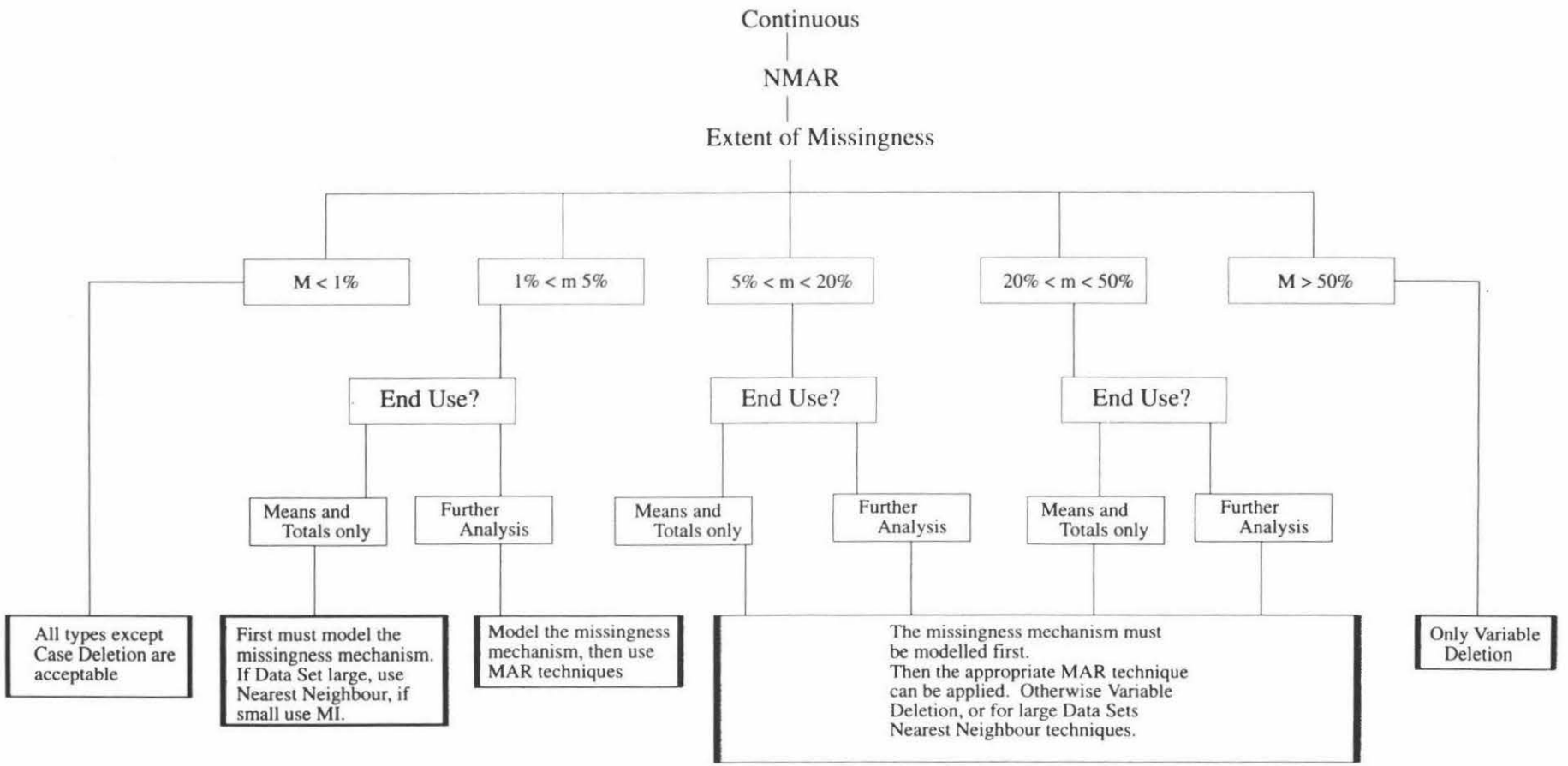
Flow Charts for

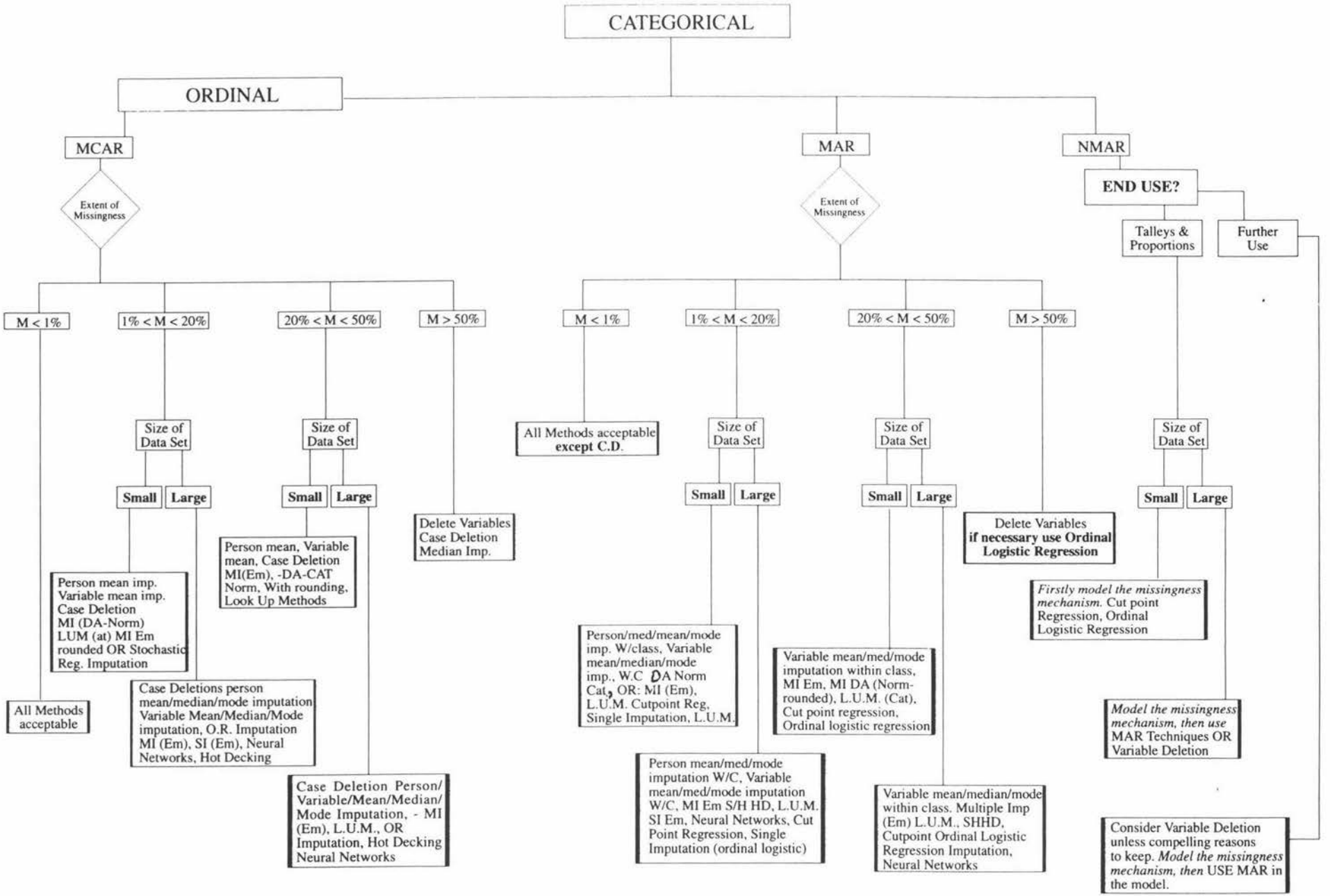
Imputation

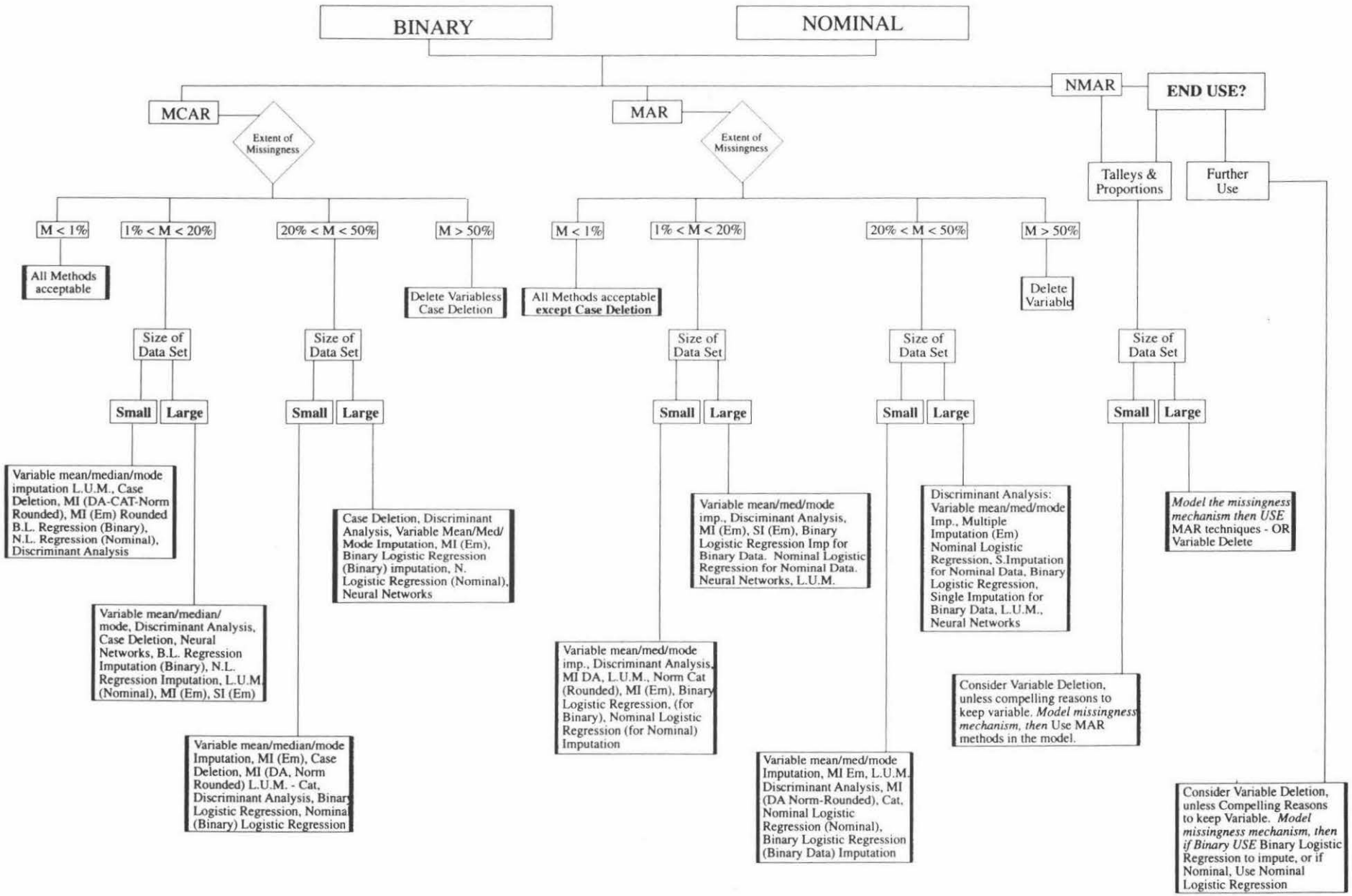
Decision Making

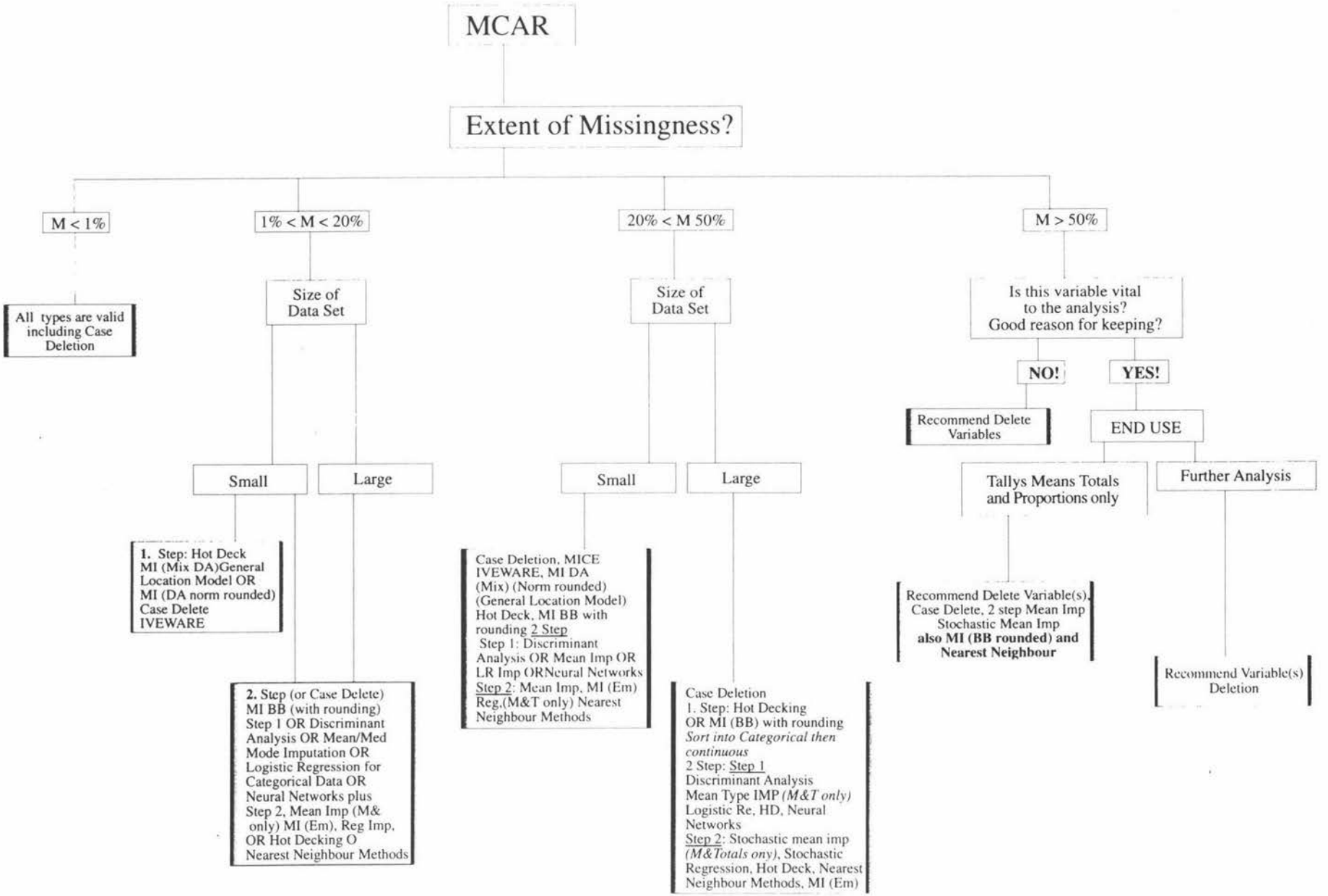

Continuous
MAR
NMAR
MCAR
Extent of Missingness
Less than 1% Very low
1% < m < 5% low
5% < m < 20% Moderate
HIGH 20% < m < 50%
Very High m > 50%
All Methods are acceptable
Recommend:
1. Case Deletions
2. Mean Images
3. Look up methods
End Use?
End Use?
End Use?
End Use?
Vital to the Analysis (ie. good reason for keeping)
YES!
NO!
Recommended: Delete Variable
End Use?
Means and Totals only
Further Analysis
Means and Totals only
Further Analysis
Means and Totals only
Further Analysis
Means and Totals only
Further Analysis
Recommend: Delete Variable
Sample Size
Sample Size
Sample Size
Sample Size
Sample Size
Sample Size
Small
Large
Small
Large
Small
Large
Small
Large
Small
Large
Small
Large
Recommended: Case Deletion, Mean Imputation Methods, Look up methods. Also:Hot Decking, Regression Imputation, Multiple Imputation (all forms), Nearest neighbour Method
Recommended: Case Deletion Stochastic Mean Imputation Seq/Hier Imputation S Imputation E.M. Look Up Mean Also: Multiple Imputation (all types) Nearest neighbour methods
Recommended: Hier. Hot Deck,Stoc. Mean Imp, Seq Hot Deck Stoch Reg Imp SI EM MI (BB) Case Deletion Also Nearest Neighbour Methods
Recommended: Case Deletion, Stoch Mean Imp, Stoch Reg Imp, Look Up Methods, Seq/Hier Hot Deck, Stoch Im Em, Multiple Imp (BB) Also:Nearest Neighbour Methods
Recommended: - Delete Variable, Case Delete, Stoch Mean Imp Look Up Methods Also: MI (BB), Nearest Neighbour
Recommended: Case Deletion Stoch Reg Imp MI (BB) S I EM Nearest Neighbour Look Up Methods
Recommended: Case Deletion, Stoch Reg Imp, Hot Deck (all types), Stoch Mean Imp., MI (BB), Look Up Methods, DA Norm, EM SI
Recommend: Look up methods, Mean Imputation Methods All Suitable Case Deletion, Hot Decking (all forms),Regression Imputation (all forms), Multiple Imputation
Recommended: Stochastic Mean Imputation, Look up methods Sequential Hot Deck Hieratical Hot Deck Look up Methods Stochastic Reg. Imputation Also CD, MI (all types)
Recommended:Wn Stochastic Method Imputation Seq Hot Deck Hieratical Hot Deck Stochastic Reg Imp SI Em MI (DA Nom) MA (BS) Also, Case Deletion
Recommended: Stoch Mean Imp. Seq Hot Deck, Heir. Hot Deck. Stoch Reg Imp, SI Em. MI (BB). DA ??, Case Deletion
Recommended: Stoch Reg Imp Hot Deck (all types), Stoch Mean Imp (MI), BB Look Up Methods, DA Norm, EM SI
Recommended: DA Norm (MI). BB (MI), I S EM, Look Up Methods Also: Hot Deck (Seq/Hier)
Recommended: - Delete Variable - Case Delete - MI (BB) - Stoch Mean Imp - Look Up Methods

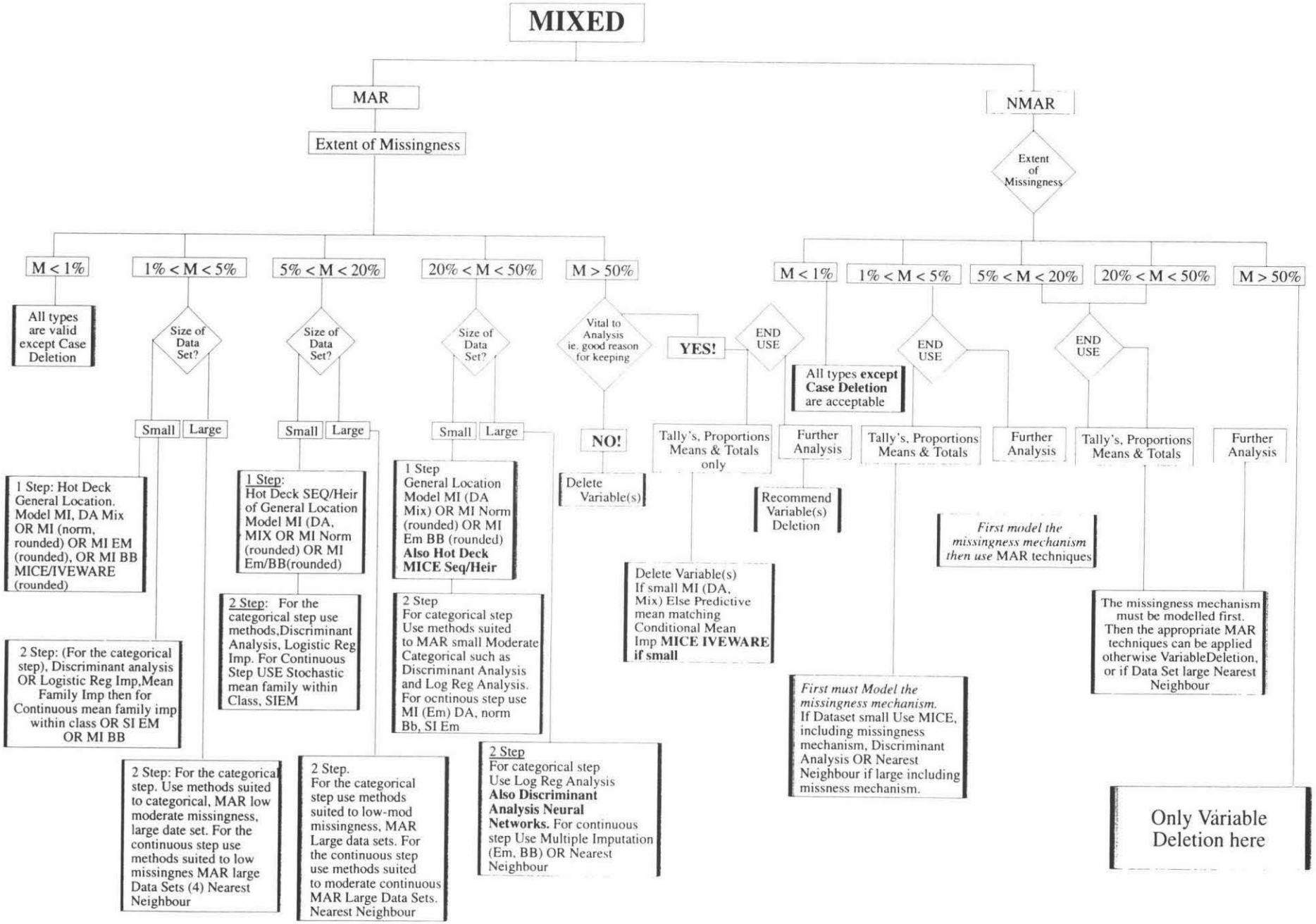

MAR
Extent of Missingness
M < 1%
All types are acceptable
1% < 5%
END USE
Means and Totals only
Further Analysis
Size of Data Set
Small
Large
Within Class Mean Substitution Seq/Heirachical Hot Decking Look Up M. Regression Imputation Multiple Imp (All)
Within Class Mean Substitution L.U.M. Regression Imputation, Seq/Heirachical Hot Decking. Multiple Imp (BB, Em) Nearest Neighbour
Size of Data Set
Small
Large
Within class, St. Mean Sub, Seq, Heir., Hot Deck, S.I., Em, L.U.M. Stochastic Reg Imp. M Imp (BB) (Em) Nearest Neighbour
Stochastic Mean Imputation L.U.M., Within class Sequential, Heirachical, Hot Deck, Stochastic Regression Imputation, Multiple Imputation (All types)
5% < M < 20%
END USE
Means and Totals only
Further Analysis
Size of Data Set
Small
Large
Within class Stochastic Mean Imputation, Seq. Hot Deck, Heirachical Hot Deck, SI, Em, L.U.M. MI, Em, DA, BB
Size of Data Set
Small
Large
Multiple Imputation (Em, BB) Nearest Neighbour, L.U.M.
Multiple Imputation (Em, DA, BB) L.U.M.
Within class Stochastic Mean Imp. Seq/Heir Hot Deck Stochastic Regression Imp. SI, Em, L.U.M. MI (Em, BB) Nearest Neighbour
20% < M < 50%
END USE
Means and Totals only
Further Analysis
Size of Data Set
Small
Large
Within class Stochastic Mean Sub L.U.M., SI, Em. Seq/Heirachical Hot Deck, Predictive Mean Matching, St Reg Imp. MI (Em, DA/BB)
Within class Stochastic L.U.M. Mean Imp. SI EM, Seq/Heirachical Ho Deck, Predictive mean matching, St Reg Imp. MI (Em, BB) Nearest Neighbour
Size of Data Set
Small
Large
MI, Em, DA, BB, L.U.M.
Multiple Imputation (Em, BB) L.U.M. Nearest Neighbour
M > 50%
Variable essential and good reason for keeping it?
NO
Delete Variable
YES
END USE
Means and Totals only
Delete Variable Predictive Mean Matching MI, Em, DA, BB, L.U.M. Conditional Mean Imp.
Further Analysis
Delete Variable also MI, Em, DA, BB

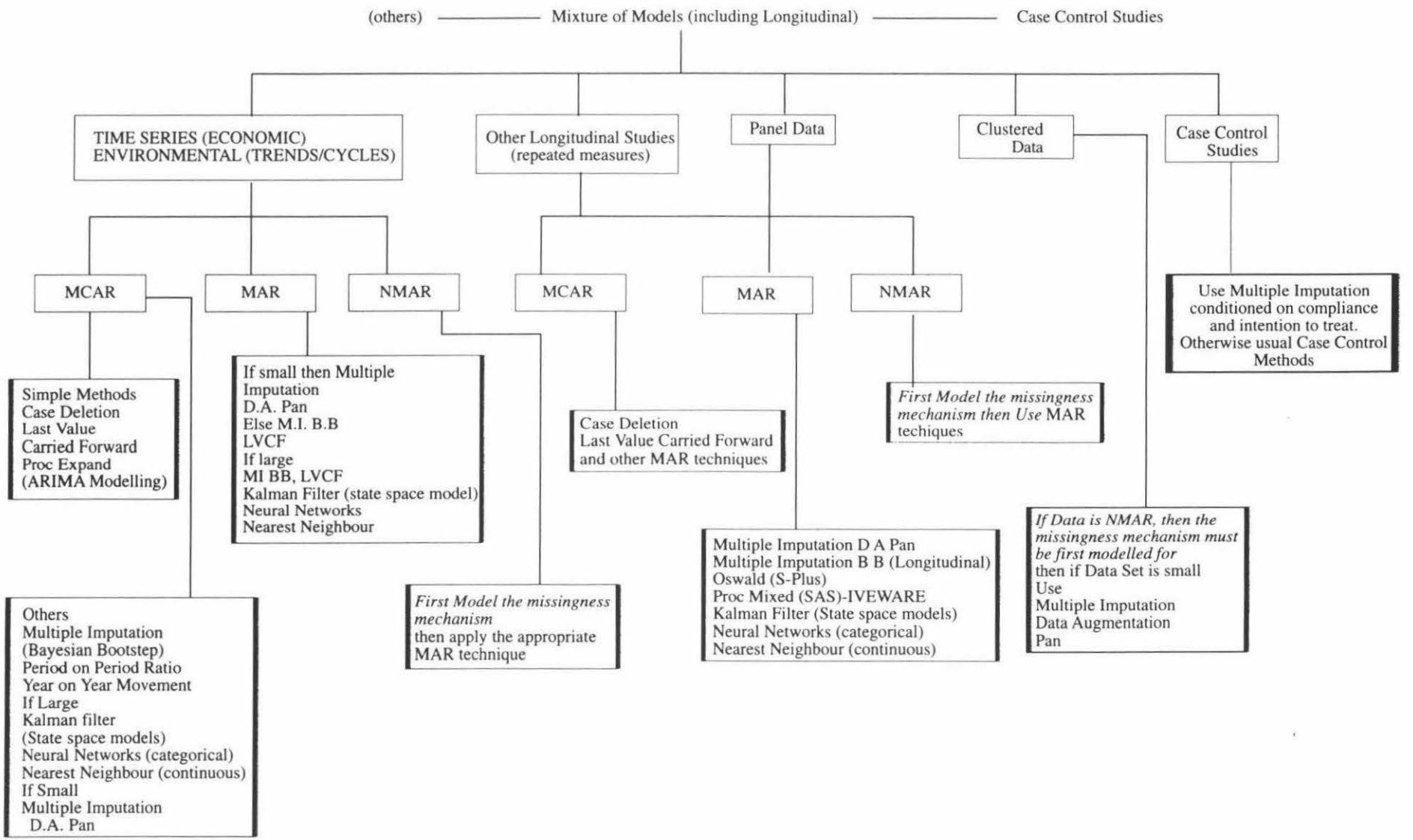

Continuous
NMAR
Extent of Missingness
M < 1%
1% < m 5%
5% < m < 20%
20% < m < 50%
M > 50%
End Use?
End Use?
End Use?
Means and Totals only
Further Analysis
Means and Totals only
Further Analysis
Means and Totals only
Further Analysis
All types except Case Deletion are acceptable
First must model the missingness mechanism. If Data Set large, use Nearest Neighbour, if small use MI.
Model the missingness mechanism, then use MAR techniques
The missingness mechanism must be modelled first. Then the appropriate MAR technique can be applied. Otherwise Variable Deletion, or for large Data Sets Nearest Neighbour techniques.
Only Variable Deletion


CATEGORICAL
ORDINAL
MCAR
Extent of Missingness
M < 1%
All Methods acceptable
1% < M < 20%
Size of Data Set
Small
Person mean imp. Variable mean imp. Case Deletion MI (DA-Norm) LUM (at) MI Em rounded OR Stochastic Reg. Imputation
Large
Case Deletions person mean/median/mode imputation Variable Mean/Median/Mode imputation, O.R. Imputation MI (Em), SI (Em), Neural Networks, Hot Decking
20% < M < 50%
Size of Data Set
Small
Person mean, Variable mean, Case Deletion MI(Em), -DA-CAT Norm, With rounding, Look Up Methods
Large
Case Deletion Person/ Variable/Mean/Median/ Mode Imputation, - MI (Em), L.U.M., OR Imputation, Hot Decking Neural Networks
M > 50%
Delete Variables Case Deletion Median Imp.
MAR
Extent of Missingness
M < 1%
All Methods acceptable **except C.D.**
1% < M < 20%
Size of Data Set
Small
Person/med/mean/mode imp. W/class, Variable mean/median/mode imp., W.C DA Norm Cat, OR: MI (Em), L.U.M. Cutpoint Reg, Single Imputation, L.U.M.
Large
Person mean/med/mode imputation W/C, Variable mean/med/mode imputation W/C, MI Em S/H HD, L.U.M. SI Em, Neural Networks, Cut Point Regression, Single Imputation (ordinal logistic)
20% < M < 50%
Size of Data Set
Small
Variable mean/med/mode imputation within class, MI Em, MI DA (Norm-rounded), L.U.M. (Cat), Cut point regression, Ordinal logistic regression
Large
Variable mean/median/mode within class. Multiple Imp (Em) L.U.M., SHHD, Cutpoint Ordinal Logistic Regression Imputation, Neural Networks
M > 50%
Delete Variables **if necessary use Ordinal Logistic Regression**
NMAR
END USE?
Talleys & Proportions
Size of Data Set
Small
Firstly model the missingness mechanism. Cut point Regression, Ordinal Logistic Regression
Large
Model the missingness mechanism, then use MAR Techniques OR Variable Deletion
Further Use
Consider Variable Deletion unless compelling reasons to keep. *Model the missingness mechanism, then* USE MAR in the model.


BINARY
NOMINAL
MCAR
Extent of Missingness
MAR
Extent of Missingness
NMAR
END USE?
Talleys & Proportions
Further Use
M < 1%
1% < M < 20%
20% < M < 50%
M > 50%
All Methods acceptable
Size of Data Set
Small
Large
Variable mean/median/mode imputation L.U.M., Case Deletion, MI (DA-CAT-Norm Rounded), MI (Em) Rounded B.L. Regression (Binary), N.L. Regression (Nominal), Discriminant Analysis
Variable mean/median/ mode, Discriminant Analysis, Case Deletion, Neural Networks, B.L. Regression Imputation (Binary), N.L. Regression Imputation, L.U.M. (Nominal), MI (Em), SI (Em)
Size of Data Set
Small
Large
Variable mean/median/mode Imputation, MI (Em), Case Deletion, MI (DA, Norm Rounded) L.U.M. - Cat, Discriminant Analysis, Binary Logistic Regression, Nominal (Binary) Logistic Regression
Case Deletion, Discriminant Analysis, Variable Mean/Med/ Mode Imputation, MI (Em), Binary Logistic Regression (Binary) imputation, N. Logistic Regression (Nominal), Neural Networks
Delete Variabless Case Deletion
M < 1%
1% < M < 20%
20% < M < 50%
M > 50%
All Methods acceptable **except Case Deletion**
Size of Data Set
Small
Large
Variable mean/med/mode imp., Discriminant Analysis, MI DA, L.U.M., Norm Cat (Rounded), MI (Em), Binary Logistic Regression, (for Binary), Nominal Logistic Regression (for Nominal) Imputation
Variable mean/med/mode imp., Disciminant Analysis, MI (Em), SI (Em), Binary Logistic Regression Imp for Binary Data. Nominal Logistic Regression for Nominal Data. Neural Networks, L.U.M.
Size of Data Set
Small
Large
Variable mean/med/mode Imputation, MI Em, L.U.M. Discriminant Analysis, MI (DA Norm-Rounded), Cat, Nominal Logistic Regression (Nominal), Binary Logistic Regression (Binary Data) Imputation
Discriminant Analysis: Variable mean/med/mode Imp., Multiple Imputation (Em) Nominal Logistic Regression, S.Imputation for Nominal Data, Binary Logistic Regression, Single Imputation for Binary Data, L.U.M., Neural Networks
Delete Variable
Size of Data Set
Small
Large
Consider Variable Deletion, unless compelling reasons to keep variable. *Model missingness mechanism, then* Use MAR methods in the model.
Model the missingness mechanism then USE MAR techniques - OR Variable Delete
Consider Variable Deletion, unless Compelling Reasons to keep Variable. *Model missingness mechanism, then if Binary USE* Binary Logistic Regression to impute, or if Nominal, Use Nominal Logistic Regression


MCAR
Extent of Missingness?
M < 1%
All types are valid including Case Deletion
1% < M < 20%
Size of Data Set
Small
1. Step: Hot Deck MI (Mix DA)General Location Model OR MI (DA norm rounded) Case Delete IVEWARE
Large
2. Step (or Case Delete) MI BB (with rounding) Step 1 OR Discriminant Analysis OR Mean/Med Mode Imputation OR Logistic Regression for Categorical Data OR Neural Networks plus Step 2, Mean Imp (M& only) MI (Em), Reg Imp, OR Hot Decking O Nearest Neighbour Methods
20% < M 50%
Size of Data Set
Small
Case Deletion, MICE IVEWARE, MI DA (Mix) (Norm rounded) (General Location Model) Hot Deck, MI BB with rounding 2 Step Step 1: Discriminant Analysis OR Mean Imp OR LR Imp ORNeural Networks Step 2: Mean Imp, MI (Em) Reg.(M&T only) Nearest Neighbour Methods
Large
Case Deletion 1. Step: Hot Decking OR MI (BB) with rounding *Sort into Categorical then continuous* 2 Step: Step 1 Discriminant Analysis Mean Type IMP *(M&T only)* Logistic Re, HD, Neural Networks Step 2: Stochastic mean imp *(M&Totals ony)*, Stochastic Regression, Hot Deck, Nearest Neighbour Methods, MI (Em)
M > 50%
Is this variable vital to the analysis? Good reason for keeping?
NO!
Recommend Delete Variables
YES!
END USE
Tallys Means Totals and Proportions only
Recommend Delete Variable(s), Case Delete, 2 step Mean Imp Stochastic Mean Imp **also MI (BB rounded) and Nearest Neighbour**
Further Analysis
Recommend Variable(s) Deletion


MIXED
MAR
Extent of Missingness
M < 1%
All types are valid except Case Deletion
1% < M < 5%
Size of Data Set?
Small
1 Step: Hot Deck General Location. Model MI, DA Mix OR MI (norm, rounded) OR MI EM (rounded), OR MI BB MICE/IVEWARE (rounded)
Large
2 Step: (For the categorical step), Discriminant analysis OR Logistic Reg Imp,Mean Family Imp then for Continuous mean family imp within class OR SI EM OR MI BB
2 Step: For the categorical step. Use methods suited to categorical, MAR low moderate missingness, large date set. For the continuous step use methods suited to low missingnes MAR large Data Sets (4) Nearest Neighbour
5% < M < 20%
Size of Data Set?
Small
1 Step: Hot Deck SEQ/Heir of General Location Model MI (DA, MIX OR MI Norm (rounded) OR MI Em/BB(rounded)
2 Step: For the categorical step use methods,Discriminant Analysis, Logistic Reg Imp. For Continuous Step USE Stochastic mean family within Class, SIEM
Large
2 Step. For the categorical step use methods suited to low-mod missingness, MAR Large data sets. For the continuous step use methods suited to moderate continuous MAR Large Data Sets. Nearest Neighbour
20% < M < 50%
Size of Data Set?
Small
1 Step General Location Model MI (DA Mix) OR MI Norm (rounded) OR MI Em BB (rounded) **Also Hot Deck MICE Seq/Heir**
2 Step For categorical step Use methods suited to MAR small Moderate Categorical such as Discriminant Analysis and Log Reg Analysis. For ocntinous step use MI (Em) DA, norm Bb, SI Em
Large
2 Step For categorical step Use Log Reg Analysis **Also Discriminant Analysis Neural Networks.** For continuous step Use Multiple Imputation (Em, BB) OR Nearest Neighbour
M > 50%
Vital to Analysis ie. good reason for keeping
NO!
Delete Variable(s)
YES!
END USE
Tally's, Proportions Means & Totals only
Delete Variable(s) If small MI (DA, Mix) Else Predictive mean matching Conditional Mean Imp **MICE IVEWARE if small**
Further Analysis
Recommend Variable(s) Deletion
NMAR
Extent of Missingness
M < 1%
All types **except Case Deletion** are acceptable
1% < M < 5%
END USE
Tally's, Proportions Means & Totals
First must Model the missingness mechanism. If Dataset small Use MICE, including missingness mechanism, Discriminant Analysis OR Nearest Neighbour if large including missness mechanism.
Further Analysis
First model the missingness mechanism then use MAR techniques
5% < M < 20%
END USE
Tally's, Proportions Means & Totals
The missingness mechanism must be modelled first. Then the appropriate MAR techniques can be applied otherwise VariableDeletion, or if Data Set large Nearest Neighbour
Further Analysis
20% < M < 50%
M > 50%
Only Variable Deletion here


(others)
Mixture of Models (including Longitudinal)
Case Control Studies
TIME SERIES (ECONOMIC) ENVIRONMENTAL (TRENDS/CYCLES)
Other Longitudinal Studies (repeated measures)
Panel Data
Clustered Data
Case Control Studies
MCAR
MAR
NMAR
MCAR
MAR
NMAR
Use Multiple Imputation conditioned on compliance and intention to treat. Otherwise usual Case Control Methods
Simple Methods
Case Deletion
Last Value
Carried Forward
Proc Expand
(ARIMA Modelling)
If small then Multiple
Imputation
D.A. Pan
Else M.I. B.B
LVCF
If large
MI BB, LVCF
Kalman Filter (state space model)
Neural Networks
Nearest Neighbour
Case Deletion
Last Value Carried Forward
and other MAR techniques
First Model the missingness mechanism then Use MAR techiques
Others
Multiple Imputation
(Bayesian Bootstep)
Period on Period Ratio
Year on Year Movement
If Large
Kalman filter
(State space models)
Neural Networks (categorical)
Nearest Neighbour (continuous)
If Small
Multiple Imputation
D.A. Pan
First Model the missingness mechanism
then apply the appropriate MAR technique
Multiple Imputation D A Pan
Multiple Imputation B B (Longitudinal)
Oswald (S-Plus)
Proc Mixed (SAS)-IVEWARE
Kalman Filter (State space models)
Neural Networks (categorical)
Nearest Neighbour (continuous)
If Data is NMAR, then the missingness mechanism must be first modelled for
then if Data Set is small
Use
Multiple Imputation
Data Augmentation
Pan

Bibliography

Afifi, A.A., Elashoff, R.M., (1966) Missing Observations and Multivariate Statistics: Review of the Literature. *Journal of the American Statistical Association.* **61** 595-604

Agresti, A., (1990) *Categorical Data Analysis.* J Wiley & Sons, New York.

Albright, V.A., Kitchell, D.A., Effectiveness of Monetary Incentives and Lottery-type Incentives in Mail Surveys. To appear in Proceedings; International Conference for Survey Nonresponse, Portland Oregon, 1999.

Anderson, T.W. (1957) Maximum Likelihood Estimates for the Multivariate Normal Distribution when some Observations are Missing. *Journal of the American Statistical Association* **52** 200-203.

Azen, S.P., Van Guilder, M., Hill, M.A., (1989) Estimation of Parameters and Missing Values Under a Regression Model with Non Normally Distributed and Non-Randomly Incomplete Data. *Statistics in Medicine.* **8** 217-228.

Alzola, C.F., Harrell, F.E., (1999) An Introduction to the S-Plus and the Hmisc and Design Libraries, 1999. http://hesweb1.med.virginia.edu/biostat/s/index.html

Azuage, F., Dubritsky, W., Lopes, P., Black, N., Adamson, K., Wu, X., White, J.A.,(1999) Predicting Coronary Disease Risk Based on Short Term RR Interval Measurements: A Neural Network Approach. *Artificial Intelligence in Medicine* **15** 275-297.

Baker, S.G., (1995) Marginal Regression for Repeated Binary Data with Outcome Subject to Non-Ignorable Non-Response. *Biometrics.* **51** 1042-1052

Bartlett, M.S. (1937) Some Examples of Statistical Methods of Research in Agriculture and Applied Botany, *Journal of the Royal Statistical Society, series B* **4**, 137-170.

Barnard, J., Meng X.L. (1999) Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES. *Statistical Methods in Medical Research* **8** 17-36.

Bonetti, M., Cole, B.F., Gelber, R.D., (1999) A Method of Moments Estimation Procedure for Categorical Quality of Life Data with Nonignorable Missingness. *Journal of the American Statistical Association* **94** (448) 1025-1034.

Borgers, N., Hox, J. (1999) Item Nonresponse in Questionnaire Research with Children and Young Adolescents. To appear in Proceedings; International Conference for Survey Nonresponse, Portland Oregon, 1999.

Box, G.E.P., Jenkins, G.M. (1970) *Time Series Analysis, Forecasting and Control.* San Francisco. Holden-Dey.

Box, G.E.P., Taio, G.C., (1975) Intervention Analysis with Application to Economic and Environmental Problems. *Journal of the American Statistical Association* **70** 70-79.

Box, M.J., Draper, N.R., Hunter, W.G., (1970) Missing Values in Multi Response Non-linear Fitting. *Technometrics.* **12** (3) 613-620.

Bradlow, E.T., Zaslavsky, A.M., (1999) A Hierarchical Latent Variable Model for Ordinal Data From a Customer Satisfaction Survey With "No Answer" Responses. *Journal of the American Statistical Association.* **94** (445) 43-52.

Brand, J.P.L. (1999) *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets.* PhD Thesis University of Rotterdam.

Cassel, C., Selen, J., Lundquist, P., (1999) Imputations Based on Box-Cox Transformations; Properties of Survey Estimates. To appear in Proceedings; International Conference for Survey Nonresponse. Portland Oregon (1999).

Chen, H.Y., Little, R.J.A., (1999) A test of Missing Completely at Random for Generalised Estimating Equations with Missing Data. *Biometrika.* **86** (1) 1-13.

Chen, R., Liu, J.S., (1996) Predictive Updating Methods with Application to Bayesian Classification. *Journal of the Royal Statistical Society. Series B.* **58** (2) 397-415.

Chesnokov, S.V.(1982) *Determinancy Analysis of Socio-economic Data.* Moscow. Nauka. (Russian).

Choi, S.C., Lu, I.L. (1995) Effect of Non-Random Missing Data Mechanisms in Clinical Trials. *Statistics in Medicine.* **14** 2675-2684.

Clayton, D., Spieglehalter, D., Dunn, G., Pickles, A. (1998) Analysis of Longitudinal Binary Data from Multi-Phase Sampling. (with discussion) *Journal of the Royal Statistical Society. Series B.* **60** (1) 71-102.

Cochran, W.G., (1977) *Sampling Techniques.* Wiley.

Cochran, W.G., Rubin, D.B., (1973) Controlling Bias in Observational Studies, a Review. *Sankhya: series A.* **35** 417-446.

Conaway, M.R.,(1992)The Analysis of Repeated Categorical Measurements Subject to Nonignorable Nonresponse. *Journal of the American Statistical Association.* **87** (419) 817-824.

Conaway, M.R.,(1993) Non-ignorable Non-response Models for Time Ordered Categorical Variables. *Applied Statistics.* **42** (1) 105-115.

Conaway, M.R.,(1994) Causal Nonresponse Models for Repeated Categorical Measurements. *Biometrics.* **50** 1102-1116.

David, M.H., Little, R.J.A., Samuhel, M.E., Triest,R.K. (1986) Alternative methods based for CPS income imputation. *Journal of the American Statistical Association* **81** 29-41.

Dempster, A.P., Laird, N.M., and Rubin, D.B., (1977) Maximum likelihood Estimation from Incomplete Data Via the EM Algorithm (with discussion) *Journal of the Royal Statistical Society, series B*, **39**, 1-38.

De Groot, M.H., (1970) *Optimal Statistical Decisions*. M^{c}Graw Hill.

DeVille, J.C., Sarndal, C-E., (1994) Variable Estimation for the Regression Imputed Hortvitz-Thompson Estimator. *Journal of Official Statistics* **10** 381-394

Diggle, Kenward, (1994) Informative Dropout in Longitudinal Analysis. (with discussion) *Applied Statistics* **43** 49-93.

Dixon, W.J. (Ed) (1988) *BMDP Statistical Software*. Berekley, University of California Press.

Dobson, A., (1998) Course notes for Biostatistics workshop on 'Analysis of Longitudinal Data Using Repeated Measures', The University of Auckland. July 1998.

Downey, R. G., King, C. V., (1998) Missing Data in Likert Ratings: A Comparison of Replacement Methods. *The Journal of General Psychology* **125** (2) 175-191.

Evans, Hastings and Peacock (1993) *Statistical Distributions*. Wiley

Fitzmaurice, G.M., Laird, N.M., Zahner, E.P., (1996) Multivariate Logistic Models for Incomplete Binary Responses. *Journal of the American Statistical Association* **91** (433) 99-108.

Fuchs, C., (1982) Maximum Likelihood Estimation and Model Selection in Contingency Tables with Missing Data. *Journal of the American Statistical Association* **77** 270-278.

Ganesalingham, S., (1999) 61.723 Multivariate Statistical Analysis: course notes. Massey University.

Gelfand, A.E., Sahu, S.K.,(1999) Identifiability, Improper Priors, and Gibbs Sampling for Generalised Linear Models. *Journal of the American Statistical Association* **94** (445) 247-253.

Gelman, A., Rubin, D.B., (1992) Inference from Iterative Simulation using Multiple Sequences (with discussion) *Statistical Science* **7** 457-511.

Gelman, A., Rubin, D.B., Carlin, J., Stern, H., (1995) Bayesian Data Analysis. Chapman and Hall, London.

Gray, R., Campanelli, P., Deepchand, K., Prescott-Clarke, P.,(1996) Exploring Survey Non-response: The Effect of Attrition on a Follow-up of the 1984-85 Health and Life Style Survey. *The Statistician* **45** (2) 163-183.

Groves, R., (1989) *Nonresponse in Sample Surveys and Survey Costs.*

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C., (1995) *Multivariate Data Analysis with Readings.* 4th Ed. Prentice and Hall.

Heeringa, S., Little R.J.A., Ranghunathan, T.E. (1999) Multivariate Imputation of Coarsened Survey Data on Household Wealth. To appear in Proceedings; International Conference for Survey Nonresponse, Portland Oregon, 1999.

Heitjan, D.F., and Basu, S.(1996) Distinguishing "Missing at Random", and "Missing Completely at Random" *The American Statistician* **50** (3) 207-213.

Heitjan, D.F. (1997) What Can Be Done About Missing Data? Approaches to Imputation. *American Journal of Public Health.* **87** (4) 548-550

Heitjan, D.F. (1999) Ignorability and Bias in Clinical Trials. *Statistics in Medicine.* **18** 2421-2434.

Hedderley, D., Wakeling, I., (1995) A Comparison of Imputation Techniques for Internal Preference Mapping, using Monte Carlo Simulation. *Food Quality and Preference.* **6** 281-297.

Hocking, R.R. (1985) *The Analysis of Linear Models.* Monterrey, California. Brookes-Cole.

Hox, J., (1999) A Review of Current Software for Handling Missing Data *Kwantitatieve Methoden: Special Issue on Missing Data* **20** (62) 123-138.

Huang, S., Brown, M.B.,(1999) A Markov Chain Model for Longitudinal Categorical Data When There May Be Non-ignorable Non-response. *Journal of Applied Statistics.* **26** (1) 5-18.

Hughes, J.A., Colley Gilbert, B.J. Item Nonresponse in a Survey with Sensitive Topics. To appear, in Proceedings: International Conference for Survey Nonresponse, Portland Oregon, 1999.

Hunt, L.A., (1996) *Clustering Using Finite Mixture Models.* PhD Thesis, University of Waikato.

Hunt, L.A., Jorgensen, M.A., (1999) Mixture Model Clustering using the Multimix Program. *The Australia and New Zealand Journal of Statistics.* **41** (2) 153-171

Ibrahim, J., (1990) Incomplete Data in Generalised Linear Models. *Journal of the American Statistical Association.* **85** 765-769

Jamshidan, M., (1997), An EM algorithm for ML Factor Analysis with Missing Data. In Berkane, M.(ed), Latent Variable Modelling and Applications to Causality. 247-258 Springer (New York)

Kalton, G., Kasprzyk, D., (1986) The treatment of missing survey data. *Survey Methodology* **12** 1-16.

Knaub, J., (1999) Using Prediction-Oriented Software for Estimation in the Presence of Non-response. To appear in Proceedings; International Conference for Survey Nonresponse, Portland Oregon, 1999.

Kuusela, V., Notkola, V., Survey Quality and Mobile Phones. To appear in Proceedings; International Conference for Survey Nonresponse, Portland Oregon, 1999.

Laird, N.M., (1988) Missing Data in Longitudinal Studies. *Statistics in Medicine* **7** 305-315.

Larsen, M.D., (1999) An Analysis of Survey Data on Smoking Using Propensity Scores. *Sankhya: The Indian Journal of Statistics: Special Issue on sample surveys. Series B.* **61** (1) 91-105.

Lee, H., Rancourt, E., Sarndal, C.E., (1999) Variance Estimation under Single Imputation. To appear in Proceedings: International Conference for Survey Nonresponse, Portland Oregon, 1999.

Lessler, J.T., Kalsbeek, W.D., (1992) *Nonsampling Error in Surveys.* Wiley.

Likert, R., (1932) A Technique for the Measurement of Attitude Scales. *Archives of Psychology.* **140** 44-53.

Lindsrom, M.J., Bates, D.M., (1988) Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated Measures Data. *Journal of the American Statistical Society.* **83** 1014-1022.

Lipsitz, S.R., Zhao, L.P., Molenberghs, G., (1998) A Semiparametric Method of Multiple Imputation. *Journal of the Royal Statistical Society. Series B.* **60** (1) 127-144.

Little, R.J.A., (1985) A Note about Models for Selectivity Bias. *Econometrica* **53** 1469-1474.

Little, R.J.A., (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* **83** (404) 1198-1202.

Little, R.J.A., (1992) Regression with missing X's, A review. *Journal of the American Statistical Association* **87** 1227-1237.

Little, R.J.A., (1993) Pattern Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association* **88** 125-134.

Little, R.J.A., (1994) A Class of Pattern Mixture Models for Normal Missing Data. *Biometrika.* **81** 471-483.

Little, R.J.A., (1995) Modelling the Drop-out Mechanism in Repeated Measures Studies. *Journal of the American Statistical Association* **90** 1112-1121.

Little, R.J.A., Rubin D.B. (1987) *Statistical Analysis with Missing Data.* Wiley.

Liu, T.-P., Rancourt, E., (1999) Categorical Constraints Guided Imputation for Nonresponse in Surveys. To appear in Proceedings; International conference for Survey Nonresponse, Portland Oregon.

Liu, C., Rubin, D.B., Wu, Y.N., (1998) Parameter Expansion to Accelerate EM: The PX-EM algorithm. *Biometrika.* **85** (4) 755-770.

Liu, J. S., Wu, Y.N., (1999) Parameter Expansion for Data Augmentation. *Journal of the American Statistical Association* **94** (448) 1264-1274.

Loomis, L., (1999) Nonresponse to Income Questions in Person-Based and Topic-Based Questionnaire Forms. To appear in Proceedings; International Conference for Survey Nonresponse, Portland Oregon.

M^{c}Cullagh, P., (1980) Regression Models for Ordinal Data. *Journal of the Royal Society Of Statistics, Series B* **42** 109-142.

M^{c}Donald, B.M., (2000) Personal Communication.

M^{c}Ilheny, C., Fleming, A., (1999) Religion and Nonresponse: Polling the Peace Process in Northern Ireland. To appear in Proceedings; International Conference for Survey Nonresponse, Portland Oregon.

Mariani, M.N., Olsen, A.R., Rubin, D.B., (1980) *Maximum Likelihood Estimation in Panel Studies with Missing Data in Sociological Methodology*. San Francisco. Jersey Boss.

Mason, R., Lessler, V.M., Traugott., (1999) Impact of Item Nonresponse on Nonsampling Error. To appear in Proceedings; International Conference for Survey Nonresponse, Portland Oregon.

Meng, X.L., (1994), Multiple Imputation Inferences with Uncongenial Sources of Input. *Statistical Science* **9** 538- 573

Michiels, B., Molenberghs, G., Lipsitz, S.R., (1999) A Pattern Mixture Model for Incomplete Categorical Data. *Communications in Statistics: Theory and Methodology.* **28** (12) 2843-2869.

Mills, I., Teague, A., (1991) Editing and Imputing Data for the 1991 Census. *Population Trends* **64** 30-37.

Newton, M.A., Zhang,Y., (1999) A Recursive Algorithm for Nonparametric Analysis with Missing Data. *Biometrika.* **86** (1) 15-26.

Norhold, E.S., (1998) Imputation: Methods, Simulation, Experiments and Practical Examples. *International Statistical Review.* **66** (2) 157-180

Norton, J., (1999) *Science, Technology and the Risk Society: Australian Consumers' Attitudes to Genetically-Engineered Foods.* Unpublished Doctoral Thesis, Central Queensland University.

Olkin, I., Tate, R.F., (1961) Multivariate Correlation Models with Mixed Discrete and Continuous Variables. *Annals of Mathematical Statistics* **32** 448-465.

Oudshoorn, K., van Buuren, S., van Rijckevorsel, J., (1999) Flexible Multiple Imputation by Chained Equations of the AVO-95 Survey. TNO report PG/VGZ/99.045 Contributed paper, International Conference for Survey Nonresponse, Portland Oregon.

Park, T., Brown, M.B.,(1994) Models for Categorical Data with Non-Ignorable Non-Response. *Journal of the American Statistical Association.* **89** 44-52

Park, T., Davis, C.S., (1993) A Test of the Missing Data Mechanism for Repeated Categorical Data. *Biometrics.* **49** 631-638.

Park, T., Lee, S-Y. (1997) A Test of Missing Completely at Random for Longitudinal Data with Missing Observations. *Statistics in Medicine* **16** 1859-1871.

Pauline, G., Ferrari, D. (1996) Do Expenditures Explain Income? A Study of Variables for Income Imputation. *Journal of Economic and Social measurement.* **22** 103-128.

Raaijmakers, Q. A. W. (1999) Effectiveness of Different Missing Data Treatments in Surveys with Likert-Type Data: Introducing the Relative Mean Substitution Approach. *Educational and Psychology measurement.* **59** (5) 725-748.

Raessler, S., (1999) An Evaluation of Imputation Techniques when the Missing-data Mechanism is Nonignorable. To appear in Proceedings; International Conference for Survey Nonresponse, Portland Oregon.

Raghuanthan, T.E., Solenberger, P.W., Van Hoewyk, J. (2000) IVEware: Imputation and Variance Estimation Software. Draft Report, Survey Research Center, Institute of Social Research, University of Michigan, July 2000.

Rao, J.N.K., Sitter, R.R., (1995) Variance Estimation under 2-phase Sampling with Applications to Imputation for Missing Data. *Biometika* **82** 453-460.

Raymond, M.R., (1986) Missing data in Evaluation Research. *Evaluation and the Health Professions.* **9** (4) 395-420

Raymond, M.R., Roberts, D.M., (1987) A Comparison of Methods for Treating Incomplete Data in Research. *Educational and Psychological Research.* **47** 13-26

Rosenbaum, P.R., (1995) Discussion of Causal Diagrams for Empirical Research. *Biometrika* **82** 698-699

Rosenbaum, P.R., Rubin, D.B., (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika.* **70** 41-55.

Rotnitsky, A., Wypij, D., A note on the Bias of Estimators with Missing Data. *Biometrics.* **50** 1163-1170.

Rubin, D.B., (1972) A Non Iterative Algorithm for Least Squares Estimation of Missing Values in Any Analysis of Variance Design. *Applied Statistics.* **21** 136-141.

Rubin, D.B., (1974) Characterising the Estimation of Parameters in Incomplete Data Problems. *Journal of the American Statistical Association.* **69** (346) 467-474.

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

Rubin, D.B., (1976b) Comparing Regressions when Some Predictor Variables are Missing. *Technometrics.* **18** 201-206

Rubin, D.B., (1981) The Bayesian Bootstrap. *The Annals of Statistics* **9** 130-134.

Rubin, D.B., (1987) *Multiple Imputation for Nonresponse in Surveys.* Wiley.

Rubin, D.B., (1996) Multiple Imputation after 18+ years. *Journal of the American Statistical Association.* **91** 473-489.

Rubin, D.B., (2000) Personal Communication.

Rubin, D.B., Schafer, J.L., (1990) Efficiently Creating Multiple Imputations for Incomplete Multivariate Normal Data. in Proceedings of the Statistical Computing Chapter 83-88 Alexandria, MD. American Statistical Association 1990.

Rubin, D.B., Schafer, J.L., Schenker, N. (1988) Imputation Strategies for Missing Values in Post Enumeration Surveys. *Survey Methodology*, **14**, 209-221.

Rubin, D.B., Schenker, N., (1991)Multiple Imputation in Healthcare Databases; An Overview and Some Applications. *Statistics in Medicine* **10** 585-598.

Rubin, D.B., Stern, H.S., Vehovar, V (1995) Handling "Don't Know" Survey Responses: The Case of the Slovenian Plebiscite. *Journal of the American Statistical Association.* **90** (431) 822-828.

Rubin, D.B., Thayer, D.T., (1982) EM Algorithms for ML Factor Analysis.
Psychometrika **47** 69-76

Rubin, D.B., Thayer, D.T., (1983) More on EM Algorithms for ML Factor Analysis. *Psychometrika* **48** 253-257

Sakia, R.N., (1990) Retransformed Bias. A Look at the Box-Cox Transformation to Linear Balanced Model Mixed ANOVA Models. *Metrika* **90** (6) 345-351.

Sarndal, C-E., (1992) Methods for Estimating the Precision of Survey Estimates when Imputation has been Used. *Survey Methodology* **18** 241-252.

Schafer, J.L., (1994) Comments on Multiple Imputation Inferences with Uncongenial Sources of Input. (Meng, X.L.) *Statistical Science* **9** 560-561

Schafer, J.L., (1997) *The Analysis of Incomplete Multivariate Data.* Chapman and Hall.

Schafer, J.L., (1998) The Practice of Multiple Imputation. Lecture notes prepared for Centre for Applied Social Surveys, University of South Hampton, U.K.

Shao, J., Chen, Y., (1999) Approximate Balanced Half Sample and Related Replication Methods for Imputed Survey Data. *Sankhya: The Indian Journal of Statistics: Special Issue on sample surveys. Series B.* **61** (1) 187-201.

Skinner, C. (1999) Developing an Imputation Strategy for a Self-Completion Survey of Local Authorities. Notes for Address to the Royal Statistical Society, London. March 17th 1999.

Steel, D., Vella, J., Harrington., (1996) Quality Issues in Telephone Surveys: Coverage, Non-response and Quota sampling. *Australian Journal of Statistics.* **38** (1) 15-34.

Tang, Fung (1997) Case Deletion Diagnostics for Test Statistics in Multivariate Regression. *Australian Journal of Statistics.* 345-353

Troxel, A.B., Harrington,D.P., Lipsitz, S.R., (1998) Analysis of Longitudinal Data with Non-ignorable Non-monotone Missing Values. *Applied Statistics.* **47** (3) 425-438.

Vacek, P.M., Ashikaga, T., (1980) An Examination of the Nearest Neighbour Rule for Imputing Missing Values. *Proceedings of the Statistical Computing Section, 1980, American Statistical Association.* 326-331.

Vach, N., and Blettner, M., (1991) Biased Estimation of the Odds Ratio in Case Control Studies due to the use of Ad Hoc Methods of Correcting for Missing Values for Confounding Variables. *American Journal of Epidemiology* **134** (8) 895-907.

Vach, N., Blettner, M., (1995) Logistic Regression with Incompletely Observed Categorical Covariates, Investigating the Sensitivity Against Violation of the Missing at Random Assumption. *Statistics in Medicine.* **14** 1315-1329.

Van Buuren, S., Oudshoorn, K., (1999) *Flexible Multivariate Imputation by MICE* TNO Report PG/VGZ/99.054. Contributed paper; International conference for Survey Nonresponse, Portland Oregon.

Wallace, C.S., Dowe, D.L.,(1998) Minimum Message Length Mixture Modelling of Multistate, Poisson, Von Mises Circular and Gaussian Distibutions. *In Proceedings 28th Symposium on the Interface,* eds L. Billard, N.I. Fisher, Computing Science and Statistics, **28** 608-613. Fairfax Station, VA. Interface Foundation of North America.

Watson, P.E., (1996) *Maternal Nutrition and Infant Outcomes*: Report to the Ministry of Health, 31 October 1996.

Wiggins, R.D., Lynch, K., Gleave, S., and Bynner, J., (1999) Teaching Applied Multivariate Analysis in the Context of Missing Data: a Comparative Evaluation of Current Software Remedies. To Appear in Proceedings; International Conference for Survey Nonresponse, Portland Oregon.

Wilks, S.S., (1932) Moments and Distribution of Population Parameters from Fragmenting Samples. *Annals of Mathematical Statistics.* **3** 163-195.

Wolfinger, R.D., Chang, M., (1995) 'Comparing the SAS GLM and MIXED Procedures for Repeated Measures', *Proceedings of the Twentieth Annual SAS Users Group Conference.*

Wolter, K.M., (1985) *Introduction to Variance Estimation.* Springer-Verlag.

Woolston, R.F., Clarke, W.R., (1984) Analysis of Categorical Incomplete Longitudinal Data. *Journal of the Royal Statistical Society, Series A.* **147** (1) 87-99.

Xie, F., Paik, M.C. (1997) Multiple Imputation for the Missing Covariates in Generalised Estimating Equations. *Biometrics* **53** 1538-1546.