Handling missing data in analyses of the UK Women's Cohort Study

By

Ula Ali Mohamed Nur

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

The Nuffield Institute for Health

November, 2004

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others

Acknowledgments

First, I wish to express my appreciation to the Economic and Social Research Council for the studentship, which was the major financial support for the research described in this thesis.

Completing a PhD is truly a lengthy event, and I would not have been able to complete this journey without the aid and support of countless people over the past three years. From the formative stages of this thesis, to the final draft, I owe an immense debt of gratitude to my supervisors, Mr. Darren Greenwood and Dr. Nick Longford, who not only served as my supervisors, but also challenged and encouraged me throughout my programme. I am also gratefully indebted to Dr. Janet Cade who listened to my stories of panic and guided me along my way.

And finally to my family who suffered through each paragraph along with me, I acknowledge a debt, an appreciation that extends beyond any words at my command.

In addition to my supervisors and family, I need to thank some people who I met during the thesis journey, Mr. David Bowers for his great support and funny jokes, at times when I really needed it and Dr. Mark Gilthorpe who although I met only in the final year of my research, never accepted less than my best efforts.

Abstract

Missing values are a problem in large-scale surveys with extensive questionnaires. The analysis of the complete records may yield inferences substantially different from those that would be obtained had no data been missing.

The aim of this dissertation is to critically examine ways of handling missing data in the UK Women Cohort Study (UKWCS). This is a large dataset with continuous, categorical and binary variables with missing values in almost every variable.

A number of simple imputation techniques, as well as multiple imputation developed by Rubin (1987), and multiple imputation by chained equations using the Gibbs sampling (Van Buuren, 1999), were explored in a number of illustrative analyses associated with the UKWCS.

Three approaches of handling missing dietary information on alcohol consumption were compared. The comparison shows that ignoring missingness by analysing only complete cases produces bias (lower means). Imputing an extreme value zero as is customary at present, underestimates the actual alcohol consumption, it also incorrectly increases the apparent precision of estimation (i.e. inappropriately small standard errors).

A published study, Pollard *et al,* (2001) which based its conclusion on one third of the records was replicated after handing missing data by multiple imputation. Multiple imputation by chained equations, an iterative technique, which deals with missing values when every variable is incomplete, was applied. This method greatly improved the results by utilizing most of the information in the incomplete

records. The method has the advantage that the algorithm intended for analysing the complete data is applied several times, without any alterations.

The implications of missing data were also studied in a survival analysis, investigating the link between incidence of breast cancer and a number of prognostic factors. The thesis recommends multiple imputation for handling missing data, by which most of the information in the dataset is exploited, and helps in efficient inferences to be made from subsequent analyses.

Table of contents

Acknowledgments **2**

Abstract **3**

List of tables **9**

List of figures **13**

Chapter 1 ..... **16**

Introduction **16**

1.1 Missing data: the problem 16

1.2 Diet and diseases 21

1.3 The Food Frequency Questionnaire 22

1.4 The U.K. Women's Cohort Study 25

1.5 Missing data in the UKWCS 29

1.6 Data cleaning 30

1.7 Data dictionary 31

1.8 Motivation 32

1.9 Outline of the thesis 32

Chapter 2 ..... **36**

Handling missing data **36**

2.1 Introduction 36

2.2 Missing data mechanism 38

2.2.1 Missing completely at random (MCAR) 40

2.2.2 Missing at random (MAR) 41

2.2.3 Missing not at random (MNAR) 42

2.3 Missing by design 43

2.4 Pattern of missing data 44

2.5 Handling missing data in the literature 45

2.6 Handling missing data in nutrition.....46
2.7 Existing methods of handling missing data.....51
2.7.1 complete case analysis.....52
2.7.2 Imputation.....53
2.7.3 Model based methods.....64
2.8 Multiple imputation.....69
2.8.1 Multiple imputation with categorical data.....76
2.8.2 Multiple imputation with continuous data.....80
2.9 The motivation for multiple imputation.....82
2.10 Bayesian approach to multiple imputation.....83
2.10.1 Gibbs sampling.....85
2.11 Applications of multiple imputation in the literature.....87
2.12 Software for handling missing data.....92
2.13 Discussion.....94

Chapter 3.....96

Alcohol consumption.....96

3.1 Introduction.....96
3.2 Missing data in alcohol consumption (FFQ).....101
3.3 Missing data in alcohol consumption (long questions).....103
3.3.1 Row borrowing.....104
3.3.2 Column borrowing.....105
3.3.3 Imputing means for the cells.....106
3.3.4 Regression imputation.....109
3.4 Alcohol nutrients.....113
3.4.1 Complete case analysis.....113
3.4.2 Imputing a default value.....114
3.4.3 Multiple imputation.....115
3.4.4 Multiple imputation by MCMC.....121
3.5 Discussion.....126

Chapter 4.....129

Imputation of missing predictors in a regression model: a comparison of methods.....129

4.1 Introduction.....129

4.2 The variables 131
4.3 Missing data in the covariates 134
4.4 Complete case analysis 136
4.5 Single variable analyses 137
4.6 Multiple imputation conditioning on one variable 140
4.6.1 Plausible values 140
4.6.2 Categorical variables 141
4.6.3 Regression methods for continuous predictors 146
4.6.4 Multiple imputation analysis 148
4.6.5 Complete case analysis and multiple imputation (conditioning on one variable) 149
4.7 Multiple imputation by chained equations 153
4.7.1 The quality of the imputed values 156
4.7.2 Monitoring convergence 157
4.7.3 Sensitivity analysis 161
4.7.4 Comparing complete case analysis and multiple imputation by chained equations 164
4.8 Discussion 166

Chapter 5 169

Handling missing data in survival analysis of cancer incidence in the UKWCS 169
5.1 Introduction 169
5.2 Cancer in women 171
5.3 Survival analysis (Complete case analysis) 173
5.3.1 Prognostic factors 174
5.3.2 Single variable survival analyses 175
5.3.2 The proportional hazards analysis 177
5.4 Missing data 181
5.4.1 Investigating missing data 182
5.4.2. Evidence of missing at random 183
5.5 Handling missing data by multiple imputation 187
5.6 Multiple imputation analysis 190
5.7 Sensitivity analysis 191
5.8 Hotdeck using STATA 195
5.10 Discussion 200

Chapter 6 ..... **203**

Assessing the missingness mechanism using a repeated FFQ **203**

6.1 Introduction 203

6.2 Block I: Bread/Savoury Biscuits 205

6.3 Plotting data 212

6.4 Measuring agreement in bread block of FFQ versus the repeated FFQ.... 215

6.5 Missing meant to be zero or 'never' 217

6.6 Alcohol consumption in the original and repeated questionnaire 219

6.7 Discussion 224

Chapter 7 ..... **227**

Summary **227**

7.1 Discussion 227

7.2 Future work 236

References **240**

APPENDIX A 260

APPENDIX B 281

APPENDIX C 291

APPENDIX D 294

List of tables

Table 1.1: Strengths and weaknesses of the FFQ: adapted from *Present Knowledge in Nutrition* 6th Edition ILSI (Brown, 1990).....24

Table 2.1: Illustration of hotdeck imputation in an incomplete data matrix.....60

Table 2.2: The proportion of missing information λ for the quantity estimated, number of imputations m and the efficiency of the multiple imputation procedure 73

Table 2.3: Two frequency questions each with 3 categories 76

Table 2.4: Number of publications of multiple imputation by year 87

Table 3.1: Alcohol nutrient in a pint of beer and a glass of cider, spirit and sherry 100

Table 3.2: Observed distributions of responses to Alcohol consumption block of FFQ 102

Table 3.3: Imputing zero for missing values in alcohol consumption of the FFQ 102

Table 3.4: Response frequencies and percentages to question B 104

Table 3.5: Column borrowing for alcohol items if response was "never" in the same alcohol item of the FFQ.....106

Table 3.6: Frequencies of responses by code to alcohol FFQ. 107

Table 3.7: Frequencies and percentages of missing items to question B and C after row borrowing, column borrowing and imputing means to the cells.....108

Table 3.8: Correlations between identical pairs of alcohol consumption.....109

Table 3.9: Correlation coefficients (all available data), of each alcohol type of question B and its counterparts from question C and the FFQ.....118

Table 3.10: The impact of handling missing data by the complete case analysis, imputing zeros and multiple imputation, on alcohol nutrient intake. 120

Table 3.11: Alcohol nutrients in the four types of alcohol and total alcohol nutrients in g/day, in the observed and the five completed datasets.....125

Table 4.1: Cross-tabulation of the responses to being vegetarian or vegan 132

Table 4.2: Numbers and percentages of missing values in variables included in the logistic regression model. 135

Table 4.3: Single variable logistic regression on all available data, with missing values in categorical covariates replaced by a dummy category for each variable. 138

Table 4.4: * Pearson's correlation coefficients of categorical variables (all available data) included in the logistic regression model 142

Table 4.5: Cross-tabulation of the two variables socio-economic class and highest educational qualification.....143

Table 4.6: Joint distribution of socio-economic class given highest educational qualification 144

Table 4.7: Cross-tabulation of the variable vitamin and vegan 145

Table 4.8: Conditional distribution of vitamin and vegan 145

Table 4.9: Results from the logistic regression model describing the relative probabilities of being a high fruit and vegetable consumer.....151

Table 4.10: Correlations between variables with the strongest association, in observed dataset and the five completed datasets.....157

Table 4.11: Correlations between variables with the strongest association, in observed dataset and the five completed datasets, with the lowest category set as starting value.....164

Table 5.1: Single variable Cox regression, of the potential risk factors on incidence of breast cancer, on all available data.....176

Table 5.2: Cox proportional hazard model fitted with 28,166 (81%) observations as the result of missing data in all prognostic factors. 179

Table 5.3: Test of the proportional hazard assumption based on Schoenfeld residuals, performed on time scale. 180

Table 5.4: Number observed and missing cases in variables included in the survival analysis model.....182

Table 5.5: Associations between missing values and prognostic factors used in the survival analysis model, auxiliary variables and survival time. 185

Table 5.6: Frequencies and their percentages of prognostic factors in the original and the five imputed datasets.....189

Table 6.1: Distribution of responses to the first block - White bread and rolls: Phase 1 against the repeated FFQ. 207

Table 6.2: Difference in responses to 'white bread' in original vs. repeated questionnaire. The median intake in the original questionnaire=3 'Once a week'. The median intake in the repeated questionnaire= 2 '1-3 times per month'. 207

Table 6.3: Distribution of responses to the first block - Brown bread and rolls: Phase 1 against the repeated FFQ. 209

Table 6.4: Difference in responses to 'brown bread' in original vs. repeated questionnaire. Median intake in the original questionnaire=3 'Once a week'. Median intake in the repeated questionnaire= 2 '1-3 times per month' 211

Table 6.5: Test of significant difference of consumption of bread using Wilcoxon signed rank test. 211

Table 6.6: Upper and lower limits of agreement for each bread type, D= mean difference and S=standard deviation of differences, between original and repeated questionnaire.....217

Table 6.7: Missing, zero and positive values of the responses to FFQ item about 'white bread' in original and repeated questionnaire.....218

Table 6.8: Missing, zero and positive values of the responses to FFQ item about Brown bread at original and repeated questionnaire. 218

Table 6.9: Distribution of responses to first question of alcohol consumption in original questionnaire against repeated questionnaire. 221

Table 6.10: Assessing the agreement of alcohol, smoking of Phase 1 and the...222

Table 6.11: Distribution of responses to smoking in original questionnaire against repeated questionnaire 223

List of figures

Figure 1.1: Bread/Savoury Biscuits block of the FFQ.....28

Figure 2.1: Patterns of non-response in rectangular datasets.....44

Figure 2.2: In a sample of n subjects the mean $\overline{Y}_{obs}$ of the observed is imputed for $Y_{a+1},\ldots,Y_n$ missing. 55

Figure 2.3: Illustration of how mean substitution corrupts covariances and.....56

Figure 2.4: Illustration of how mean substitution corrupts marginal distribution of Y.....57

Figure 2.5: Imputing using regression method 59

Figure 2.6: Multiple imputation replaces each missing value in the dataset by m imputed values. 71

Figure 2.7: An example where Y values can be imputed from X values by 81

Figure 3.1: Alcoholic beverages block of the FFQ.....97

Figure 3.2: Histograms of weekly consumption of Beer, Wine, Sherry and Spirits 5 years ago and now.....111

Figure 3.3: estimates (standard errors) and iterations of alcohol nutrients in the five completed variables Wine, Beer, Sherry and Spirits.....124

Figure 3.4: Thirteen records from the observed data and its relevant five imputed values generated by *MLwiN*.....124

Figure 4.1: 'Physical exercise', the observed values and the imputed values in the five completed datasets.....158

Figure 4.2: Age, the observed values and the imputed values in the five completed datasets. 159

Figure 4.3: 'Physical exercise' (mean) within ten iterations in the five imputed datasets.....160

Figure 4.4: 'Physical exercise' (Std deviation) within the ten iterations in the five imputed datasets.....160

Figure 4.5: 'Physical exercise' (mean) within the ten iterations in the five imputed datasets, with starting value set to zero. 163

Figure 4.6: 'Physical exercise' (Std) within the ten iterations in the five imputed datasets, with starting value set to zero.....163

Figure 5.1: Plot of the raw and smoothed scaled Schoenfeld residuals for drink alcohol once a week.....180

Figure 5.2: Difference in rates of getting cancer with and without recorded prognostic factors, alcohol, BMI, smoking and having children, tested using log rank method. 186

Figure 6.1: Scatter plot of White bread in original questionnaire vs repeated questionnaire.....213

Figure 6.2: Scatter plot of Brown bread in original questionnaire vs repeated questionnaire.....213

Figure 6.3: Scatter plot of Wholemeal bread in original questionnaire vs repeated questionnaire.....213

Figure 6.4: Scatter plot of Chapatis in original questionnaire vs repeated questionnaire.....213

Figure 6.5: Scatter plot of Papudum in original questionnaire vs repeated questionnaire.....214

Figure 6.6: Scatter plot of Tortilla in original questionnaire vs repeated questionnaire.....214

Figure 6.7: Scatter plot of Pitta bread in original questionnaire vs repeated questionnaire.....214

Figure 6.8: Scatter plot of Crispbread in original questionnaire vs repeated questionnaire.....214

Figure 6.9: Scatter plot of 'white bread' in original questionnaire vs repeated questionnaire, and the line of equality.....215

Figure 6.10: Difference against mean of frequencies of 'white bread' in original and repeated questionnaire.....216

Figure 6.11: Scatter plot of alcohol consumption in original questionnaire against repeated questionnaire.....221

Figure 6.12: Scatter plot of smoking in original questionnaire against repeated questionnaire.....223

Chapter 1
Introduction

1.1 Missing data: the problem

Standard statistical methods employed in epidemiological studies are valid only when applied to a representative sample of the population of interest. There are several known methods to select a representative sample of the population. The most widely used are:-

- Simple random sampling, in which each subject from the population has an equal chance of being selected. This can be achieved by generating random numbers using a computer, or by the use of random number tables. For example from a list of all patients in one population (e.g. registered cancer patients in a specific hospital) a 50% random sample can be obtained by preparing a list of randomly generated numbers, one for each patient, and selecting the patient if their random number is even.
- Stratified sampling, in which a population is stratified before simple random samples are selected from each stratum. This type of sampling can be helpful in studying a disease, which varies with respect to age, sex or family history. In this case a framework can be laid down initially based for example on sex and then patients can be divided into four age groups for males and females separately.

There are other ways; but I think these are the most common for postal surveys.

However, even if great effort has been exerted for the selected sample to be representative of the population, the validity of the applied statistical methods are eroded by incompleteness in the dataset.

In many surveys, respondents may be unwilling to answer some questions, or they can skip part of the questions by accident (item non-response). Respondents in some other cases refuse to participate, cannot be allocated or have died (unit non-response), or become unavailable for some other reason. In longitudinal studies, participants can sometimes drop out from follow-ups. Such incompleteness of the dataset that was intended to be as representative as possible of the population of interest is associated with three major difficulties: -

- Loss of information, efficiency or power due to loss of data
- Problems in data handling, computation and analysis due to irregularities in the data patterns and non-applicability of standard software
- Serious bias if there are systematic differences between the observed and the unobserved data, (Barnard *et al.*, 1999)

For example suppose that a continuous variable X1 is missing for a fraction of subjects in a study, and let us assume that the main goal of the study was to estimate the effect of factors X1, X2 and possibly other variables on a binary outcome Y. If there were no missing data a reasonable analysis would be to predict the outcome Y from X1 and X2 plus the other variables in the study using logistic regression. If we exclude X1 the incomplete variable, we can use information from all the subjects, with a possible risk of introducing bias to the odds ratios of X2. On the other hand, if we exclude all subjects with missing

values for X_1, we can use all the variables but have to reduce the sample size. This strategy risks both bias and loss of power.

The impact of the missing data on the results depends on: -

- The amount of missing data
- The mechanism that caused the data to be missing (Little and Rubin, 1987)
- The procedure the statistician or data analyst will use to deal with these missing data (Musil *et al.*, 2002; Streiner, 2002).

The method of data collection also plays an important role in the amount of missing data for a specific study. The major methods of data collection are self-administered questionnaires, face-to-face interview and the telephone interview. Each of the three methods has its advantages and disadvantages. For example, personal issues are very difficult to discuss in face-to-face interviews. A self-administered questionnaire is regarded as more confidential and respondents can give honest answers to sensitive questions. However, in self-administered questionnaires, respondents can make mistakes and skip questions, for example, they can skip a question with the lead-in passage 'How often do you drink alcohol? ' Trying to reveal that they don't drink alcohol where they were supposed to mark the response option 'never'. Telephone interviews are often less expensive since travel expenses are eliminated. However people tend to get impatient and are more likely to refuse telephone interview than a face-to-face interview. Therefore, the first consideration in planning a study should be to minimize missing data with a well-designed questionnaire, avoiding complicated and long questions, the right choice of data collection method and if possible arranging callbacks and follow-up letters to follow up non-respondents.

Unfortunately, missing data cannot be avoided in medical, epidemiological as well as research in other fields of study, even if great effort was put into planning and data collection. These difficulties lead researchers to think about developing methods to handle missing data when faced with this problem (Dempster *et al.*, 1977; Little and Rubin, 1987; Rubin 1976 and 1987; and Schafer, 1997).

The best method of dealing with missing data is to avoid the problem by careful design of the research and data collection, however when the problem exists a solution has to be found. Nevertheless, knowing how missing data occur can help in using the most appropriate method to deal with it. The impact of missing data in epidemiological and biomedical literature was discussed by Rubin and Schenker (1991); Roth (1994); Greenland and Finkle (1995) and Longford et al. (2000).

The simplest solution for the researcher when faced with item non-response is the complete-case analysis. With this approach all incomplete records are discarded, so as to force the data to a rectangular form, which can easily be analysed by most statistical packages. Although this method is simple, a large fraction of the sample may be excluded, as a result of excluding all records with missing items. Further, the complete cases may not be a representative sample, even if the original sample would have been. The obvious concern that arises is that the subjects with incomplete records may in some way be systematically different from those with complete records.

The ultimate goal of the researcher is to make inferences for the population rather than the subset of the population that would respond to all questions whose records would be complete. However, there are many problems with the complete

case analysis. First, the subset of these subjects may no longer be a representative sample from the target population, bringing about bias in the inference made. Second, the incomplete records are not used in the analysis; so a lot of useful information is discarded. Little and Rubin (1987) documented the deficiencies of the analyses based on the complete records of subjects who responded to all questionnaire items. Many methods were developed to account for this loss of information in surveys; unfortunately most of the traditional methods could not handle the problem of missing data effectively.

In medical research, the purpose of summarizing the behaviour of a sample is to draw inferences on the population from which the sample was drawn. An efficient estimator will be capable to draw these inferences by using all available information in the sample, even if the data is not complete. Another important criterion is that the precision is estimated without much bias.

The two exceptional approaches of handling missing data that give outstanding results are the EM algorithm (Dempster *et al.*, 1977) and multiple imputation (Little and Rubin, 1987; Rubin, 1987; and Schafer, 1997). Handling missing data by these two methods leads to reasonable results. The major drawback of these two methods is that they depend on assumptions that can easily be violated and their validity cannot be tested easily. But these assumptions are certainly violated in the simpler methods. These two methods together with their advantages and disadvantages, as well as the complete case analysis and other methods found in the literature will be discussed in detail in Chapter 2.

1.2 Diet and diseases

Chronic diseases such as heart disease, cancer and diabetes are the leading causes of death both in developed and underdeveloped countries. Diet is believed to have an important role in the development of chronic diseases, and during the last decade extensive research was focused on the nature and strength of the link between diet and diseases (Haroon, 2003; Gunnell *et al.*, 2003; Wilson *et al.*, 1998). A comprehensive report, published recently by the WHO/FAO, (2003) provides scientific evidence on the relationship between diet and chronic diseases. The report states that the burden of chronic diseases is rapidly increasing worldwide, both in developed and developing countries and that by the year 2020 chronic diseases will account for three quarters of the deaths world wide. The report also showed that the major known risk factors for chronic diseases in adulthood are high cholesterol (diet), and heavy or binge drinking.

A number of epidemiological studies have also found evidence that vegetarian diet is associated with lower all-cause mortality (Key *et al.*, 1999), and that the consumption of fruit and vegetables have protective effect against cardiovascular diseases as well as some types of cancer (Block *et al.*, 1992; Ness and Powels, 1997).

There has also been an extensive debate on alcohol and its effect on the body. Moderate consumption of wine was found to have a beneficial effect against coronary heart disease and cancer (Gronbaek *et al.*, 2000), but studies have also found an association between alcohol consumption and breast cancer (Enger *et al.*, 1999; Boughton, 2001; Key *et al.*, 2001). Furthermore, population studies

have shown that 80% of chronic heart diseases, up to 90% of cases of type 2 diabetes can be avoided by changing lifestyle factors, and that one third of cancers could be avoided by eating healthily and maintaining normal weight (Stampfer *et al.*, 2000; Hu *et al.*, 2001; Key *et al.*, 2002).

All this evidence, as well as increasing figures of cancers among women, led researchers to work more on links between diet and chronic diseases. To assess the implications of diet on people's health, information has to be gathered on what people eat and drink. There are a number of methods to assess dietary intake:

- Food diaries require the subject to report all food consumed in a specified period, commonly 1 to 7 days, often accurately weighing all food-stuffs consumed as well as leftovers;
- 24-hour recall consists of a list of food and beverages consumed the previous day or during 24 hours before the interview
- Food frequency questionnaire in which the respondent is asked to estimate the frequency of consumption often by ticking a frequency category, which indicates the number of times the food, is consumed per day, week, month or sometimes year.

1.3 The Food Frequency Questionnaire

The Food Frequency Questionnaire (FFQ) was defined by Margetts *et al.*, (1997) as "*A questionnaire in which the respondent is presented with a list of foods and is required to say how often each is eaten in broad terms such as X times per*

day/per week/per month, etc. Foods selected are usually chosen for the specific purposes of a study and may not assess total diet".

The FFQ is one of the least expensive and simplest methods for measuring diet, as it is self-administered and requires minimal instructions. The main strengths and weaknesses of the FFQ are shown in Table 1.1. An appropriate nutrient database should be constructed to convert frequency estimates of food intake to nutrient values. The limitations of food tables/databases need to be taken into consideration, particularly the extent to which missing values interfere with the aspects of diet that are to be assessed and if and how the limitations can be addressed (Cowin *et al.*, 1999).

Administering the FFQ twice to the same group of people can assess the reliability of the responses. A test of association or correlation coefficients can then be used to test the reproducibility of the questionnaires (Pietinen *et al.*, 1988; Bueno *et al.*, 1992; Engle *et al.*, 1990). A paper by Bland and Altman (1986) demonstrated that correlation coefficients can measure association but do not measure agreement. Therefore, assessing the agreement is preferable to the use of correlation coefficients. By this method, one can determine if there is any systematic difference between the administrations of the questionnaire (bias), and to what extent the two agree. However, very few FFQ's have been properly validated using this approach, (Cade *et al.*, 2002).

Crosscheck questions can also be used to check the reporting of certain foods. It has been found that people tend to over-report the consumption of fruit and vegetables, especially if each fruit and vegetable is listed singly in a long list. The number of servings of fruit and vegetables per week can be asked in the crosscheck question (Calvert *et al.*, 1997). Over-reporting may then be corrected

by a weighting factor selected as the number of servings per week from the crosscheck question.

Strengths	Weaknesses
An indication of usual dietary intake may be obtained	Memory of food patterns in the past is required
Intakes of both foods and nutrients can be assessed	Recall period may be imprecise
Highly trained interviewers are not required	Quantification of food intake may be imprecise because of poor judgment of recall of portions or use of standard sizes
The method can be interviewer administered or self-administered	Respondent burden is governed by number and complexity of foods listed and qualification procedure
Administration can be simple	Recall of past diet may be biased by current diet
Customary eating patterns are not affected	Heterogeneity of population influences the reliability of the methods
Individuals may be ranked or classified by food intake	Suitability is questionable for certain segments of the population such as individuals consuming atypical diet or foods not on the lists
Response rates are high	Questionnaires with a long list of foods tend to overestimate and those with a short list underestimate intake
Respondent burden is usually light	Validation of the method is difficult and often expensive
Relationship between diet and disease may be examined in epidemiological studies	Gives no information on meal patterns through out the day
Can be optically scanned to reduce data entry costs	Considerable programming time and expertise may be required to convert food frequencies into nutrients

Table 1.1: Strengths and weaknesses of the FFQ: adapted from *Present Knowledge in Nutrition* 6th Edition ILSI (Brown, 1990).

The crosscheck may however lead to underestimation of intake. For example respondents will tend not to report fruit juices when asked about consumption of fruits. This can be corrected by an easier to understand question such as "Not counting juices, how often do you eat fruit?" Although the crosscheck questions have been successfully used to assess over-reporting of fruit and vegetable intake (Calvert *et al.*, 1997), it was found not as effective in assessing the reporting of other food items (Wolk *et al.*, 1998).

The true long-term diet intake can be determined by information collected on what people eat, but this information tends to be biased as people generally either underestimate or overestimate their food intake. Bias can also arise from misjudgement, poor recall or simply because respondents skip part of the questionnaire either accidentally or intentionally, not wanting to reveal some aspects of their dietary habits.

Of probable interest in investigating chronic disease is long-term diet, so an FFQ relies on consistency of the subjects' diet. This is probably appropriate for middle-aged women who have a set life style, although long-term secular trends in diet may exert some influence.

1.4 The U.K. Women's Cohort Study

The UK Women's Cohort Study (UKWCS) aims to explore the relationship between diet and cancer incidence and mortality (from selected causes) in a group

of middle-aged women in the UK. The study also aims to detect the protective effect of vegetarian diet compared to fish eaters and non-meat eaters.

The original survey was targeted towards middle-aged women, living in England, Wales and Scotland. A 217-item food frequency questionnaire was sent to 65,000 women who were supporters of the World Cancer Research Fund (WCRF). The questionnaire was adapted from the FFQ used in the European Prospective Investigation into Cancer (EPIC) study (Riboli, 1992). All women aged between 35 and 69 years and who described themselves, as vegetarians in the original survey were included in the cohort. These were then matched for age to the nearest meat eater within the same 10-year age band; all fish eaters were also included. (Pollard *et al.*, 2001; Greenwood *et al.*, 2000). Women were then contacted by post to assess, in detail, their diet and lifestyle characteristics at baseline.

The final sample size was around 35,000 women this is referred to as Phase 1 of the cohort. Approximately a third of this sample describe themselves as being vegetarian, a third as red-meat eaters and a third as fish eaters. It should be noted that the UKWCS is not designed to be representative of the general population, but to maximize power for investigating vegetarian and fish-based diet, and to maximize the range of intake levels and variety of fruit, vegetables and fish consumed. The cohort will be followed up for ten years with repeat FFQ and other questionnaires at various time points throughout the period.

The FFQ of the UKWCS consists of 24 blocks of questions about classes of food and beverages, which total to 211 items (Appendix A). The respondents were asked to select the most appropriate of the ten response options for each item.

Each question has the lead-in passage:

'How often have you eaten these foods in the last 12 months?'

and the response options range from 'Never'(coded as 0) to 'Six or more times per day' (coded 9), see Figure 1.1 for an extract of the questionnaire. These FFQ questions are followed by 71 questions in a different format, which inquire about how subjects prepare food, what types of specific food products they prefer (e.g., full cream, semi-skimmed, skimmed, dried or sterilized milk), what food supplements they use, whether they are on a special diet, how much alcohol (and of what kind) they consume, how much they smoke and what physical activities they pursue (gardening, walking, DIY, etc.). A Section of the questionnaire has items about body size (weight, height, waist and hip sizes), history of serious illnesses, family history of cancer, level of education, type of employment, family circumstances (children's ages and birth weights), whether the subject is vegetarian or not, number of pregnancies and other socio-economic factors, and so on. Since the questionnaire is quite extensive, many subjects fail to respond to every single questionnaire item.

FOODS AND AMOUNTS	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS?									
	NEVER	Less than once a month	1-3 per month	once a week	2-4 per week	5-6 per week	once per day	2-3 per day	4-5 per day	6+ per day
BREAD/SAVOURY BISCUITS										
White bread & rolls	0	1	2	3	4	5	6	7	8	9
Brown bread & rolls	0	1	2	3	4	5	6	7	8	9
Wholemeal bread & rolls	0	1	2	3	4	5	6	7	8	9
Chapatis, Nan, Paratha	0	1	2	3	4	5	6	7	8	9
Papadums	0	1	2	3	4	5	6	7	8	9
Tortillas	0	1	2	3	4	5	6	7	8	9
Pitta Bread	0	1	2	3	4	5	6	7	8	9
Crispbread e.g. Ryvita	0	1	2	3	4	5	6	7	8	9
Cream crackers, cheese biscuits	0	1	2	3	4	5	6	7	8	9

Figure 1.1: Bread/Savoury Biscuits block of the FFQ

The UKWCS relies on volunteer participation. In fact, the subjects tend to have healthier diets and life-styles and are better educated than would be seen in a random sample of UK women in the same age group. This is partly because responders to studies like this tend to be of higher social class, and partly because the cohort from the WCRF study was recruited from potential charity supporters, and includes a large proportion of vegetarians (Greenwood *et al.,* 2000). Nevertheless, this sub-population is of particular interest. In any case, logistic difficulties, substantial non-response and high cost are a barrier to obtaining a large random sample of women.

In Phase 2 a stratified random sample of 400 subjects was selected, using a computer generated randomisation process, from the baseline participants, based on their reported dietary habits and use of food supplements such as vitamins, minerals and fish oils, mainly to get more information on the likely results of using supplements, and factors which might facilitate or inhibit supplement use.

This sample was sent a questionnaire using a methodology based on the *Theory of Planned Behaviour*, a social cognition model that shows the relationship between food choice and attitudes, (Conner et al., 2001; Ajzen, 1988, 1999).

For Phase 3, subjects who had returned the questionnaire in Phase 1 where mailed a food diary to be completed over four days. Each type of food consumed in these four days was to be weighted and recorded.

1.5 Missing data in the UKWCS

In epidemiological databases, records (rows) usually represent subjects, cases or observations; the columns represent variables measured for each subject. In the context of the UKWCS, rows represent subjects or women, and columns represent variables that are coded answers to the questionnaire items. Some long-format questions have several parts. In the UKWCS all unit non-response, that is, questionnaires received with just the name of the respondent and no responses to the questions, were excluded from the data file; there were around 400 such questionnaire forms.

A telephone company was used to remind respondents by a telephone call, if they failed to return the completed questionnaire, but it was considered too expensive to re-contact participants who returned questionnaires with gaps in them or which were not completed properly. If only complete cases were to be considered, a huge amount of information would be lost, and the dataset of 35,000 records would have only 12,000 complete records for analysis. For example, in Chapter 3 the average weekly consumption of alcohol is found by adding up the alcohol nutrients from four types of drinks beer, wine, spirits and sherry. The response

rate for these types of alcohol were 48% for beer, 82% for wine, 48% for sherry and 78% for spirits. A complete case analysis in which only complete records of the four types alcohol are to be included will discard 64% of the original sample. A great loss of information and waste of resources expended in collecting the data. The observed sample may not be representative of the original sample, if there were no missing data. The missing values or non-responses can be the result of subjects who are binge drinkers, who do not want to disclose their actual alcohol consumption, or simply did not answer questions they thought were not relevant to them.

1.6 Data cleaning

Every dataset contains some errors. Wrong conclusions may be drawn based on data in which many errors have been left undetected. Errors can be caused by data entry or as a result of inconsistent values reported in the questionnaire. The UKWCS was cleaned for possible inconsistency and errors. As expected in large-scale surveys, many extreme and impossible values were found. Impossible answers were also checked and were re-coded as inconsistent values if found, extreme values were checked and re-coded. For example, if the participant answered 'no' to being vegetarian and then entered number of years being vegetarian, the number of years being vegetarian was re-coded as an inconsistent value. Mistakes originating in the question from measurement or from incorrect questionnaire responses are difficult to detect unless they are out of range. For example as the data was double entered by a professional data entry company, data entry errors were not expected. Values that were clearly impossible, were re-

coded as inconsistent (for example, if the responder claimed to be vegetarian and the number of years being vegetarian were found to be more than the responder's age, the number of years being vegetarian was recoded as an inconsistent value). Syntax was written using the statistical package SPSS, for all the data cleaning stages, and documented in the data dictionary (Section 1.7) for any future use, (see Appendix B for a sample of the data dictionary).

1.7 Data dictionary

Before I could start dealing with missing values in the dataset, I needed to assess the content and quality of the data in the cohort. A data dictionary was a vital first step in developing my thesis.

The data dictionary contained documentation for all the variables, all the re-coding developed in the data cleaning, and a descriptive analysis for all the variables. A scoring system has been developed for the variables to give guidance to the researcher as to the quality of the data. All impossible values for variables have been considered and re-coded as necessary. This guide will hopefully also provide some indication as to how the subjects interpreted the questions. Appendix B includes 10 pages of the data dictionary, which shows the scoring system and definition for a sample of variables selected from the FFQ questionnaire as well as the long questions of Section 2. It was impractical to include the whole document of 156 pages as an appendix in this thesis.

The data dictionary is now an essential tool for the whole team working on the UKWCS, as it is a reference listing variables names, labels, and basic descriptive statistics.

1.8 Motivation

The principal methodological contribution of this thesis is to develop and extend methods that make use of the incomplete records in the UKWCS questionnaire, and to compare the properties of these methods. The questionnaire is long, therefore subjects lose concentration and motivation while completing it, or they can inexplicably omit responding to isolated questionnaire items or to whole sections. As a result records are incomplete and standard statistical analysis may lead to invalid inferences. The default solution is to omit all incomplete records from the analysis. In our case this amounts to substantial loss of information, which cannot be afforded. This thesis explores the development and implementation of methods for handling missing data, making comparison of its impact on results and on conclusions relating diet to cancer in the UKWCS.

The principal tool will be the method of multiple imputation. This method which was developed by Rubin (1987), will be explored. Although the mathematical methods are rather complex (discussed in Chapter 2), implementing the method is feasible in modern computing environments in which large-scale databases are stored and maintained.

1.9 Outline of the thesis

The aim of this project is to handle incomplete data in the UKWCS to make inferences about the components of the diet of middle-aged women in the UK,

and to prepare the ground for the (future) analyses of the association of diet with various chronic diseases, and cancer in particular.

A particular challenge of the project is to adapt these methods to the setting of the UKWCS, which consists of a large dataset (35,000 records), together with a large number of variables (around 600). The number of variables exceeds the 211 items and the 71 questions of the questionnaire, as new variables were computed, from others. For example a new variable *'age'* was computed from the date of birth, and the date when the form was received. Some of the variables are continuous while others are categorical, with as many as 10 categories each. Another difficulty in this dataset is that the pattern of missing values varies from one block of variables to another.

The UKWCS database will be used for a wide range of studies, for example:-

- Why do women use 'health foods' and supplements? psychological and social influences
- Factors affecting fruit and vegetable consumption: baseline analysis of women participating in the UKWCS

Devising a different type of adjustment for each of them, to deal with missing data, is impractical. Also, in most analyses standard statistical software or other complex programs will be applied. It would be very difficult to make any changes in the computational algorithms of these programs. Therefore, it would be especially useful, if the missing data were handled in such a way that these programs could be applied without any alterations. The method of multiple imputation satisfies this condition. A detailed description of multiple imputation will be given in Section 2.8, and its limitations and advantages are discussed.

In Chapter 2 the most commonly used methods of handling missing data will be reviewed.

In Chapter 3

- The effect of dietary assessment on the quality and quantity of data will be assessed.
- The two dietary assessment methods, which were used to collect information on alcohol consumption, and the impact they have on missing data will be described.
- The practical and conceptual issue of choosing the best method of handling missing data, where no complicated imputations would be needed, will be explored.
- Two types of multiple imputation (standard multiple imputation and multiple imputation by MCMC (Markov chain Monte Carlo) method) will be compared to other traditional methods such as mean substitution, complete-case analysis and imputing the most likely or most frequent values, in calculating alcohol nutrients from different types of alcohol intake variables.

In Chapter 4 multiple imputation by chained equation, its implementation advantages and drawbacks will be explored. This type of multiple imputation uses Gibbs sampling and will be applied to a logistic regression analysis comparing the highest and lowest consumers of fruit and vegetables to a number of socio-economic factors. Results will be compared to the same model using complete case analysis (Pollard et al., 2001).

The effect of different approaches to handling missing data on the results from a survival analysis, relating incidence of cancer and a number of life style and socio-economic factors, will be compared in Chapter 5. Results using multiple imputation will be compared to the results from the same analysis using the complete cases as well as results from data imputed using the hotdeck method implemented in STATA 8. The difference between the three sets of results will be reviewed.

In Chapter 6 response to repeated questionnaires sent to subjects five years after the start of the cohort will be compared to the original responses. This comparison will be used to study the consistency in responses, and through it, the mechanism of missing data. This will allow the validity of underlying assumptions to be partially assessed.

Chapter 7 will describe the final conclusions and outline future work.

Chapter 2
Handling missing data

2.1 Introduction

In most surveys, and the UKWCS is no exception, the problem of incomplete data is unavoidable. Large-scale questionnaire surveys rarely achieve a full 100% completion of all the questions. This problem is even worse in self-administered surveys when respondents receive questionnaires by post, and have to fill in the questionnaire without the help of an interviewer.

The main goal of a statistical procedure should always be to make valid and efficient statistical inferences about the population of interest. This goal has to be fulfilled with or without the presence of missing data. The basic criteria of evaluating a statistical procedure were established by Neyman and Pearson (1933). We assume that θ is the population quantity to be estimated, and that $\hat{\theta}$ is an estimate of θ based on sample data. In the presence of missing data the method of handling this missing data should be part of the procedure of calculating $\hat{\theta}$. If the calculated estimate $\hat{\theta}$ both on average over repeated samples is close to the population quantity θ then we can say that the method followed was acceptable. In other words the bias or the difference between the estimate $\hat{\theta}$ and the population quantity θ is small. The method should also provide small variance and standard deviation of $\hat{\theta}$.

The bias is defined as the difference between the true population quantity θ and the average of all possible estimates, this can denoted as Bias= $b(\hat{\theta}) = \theta - E(\hat{\theta})$.

The sampling variance is denoted by $\sigma_\theta^2 = E([\hat{\theta} - E(\hat{\theta})]^2)$.

If the population quantity θ was estimated by two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, we would choose the one with the smaller sampling variance. The variance and bias are combined in one measure called the mean squared error (M.S.E.), this is equal to the average of squared differences of $(\hat{\theta} - \theta)^2$ over repeated samples.

$$\text{M.S.E.} = E[(\hat{\theta} - \theta)^2]$$

The variance and bias are related to the M.S.E. by the identity

$$\text{M.S.E.} = \text{Variance} + (\text{bias})^2$$

A good estimator is one with a small M.S.E. The standard error of the estimate ($S.E.(\hat{\theta})$) can be calculated as the root of the M.S.E. A small standard error reduces the probability of Type II error (failure to reject the null hypothesis when it is true) and increases power. The 95% confidence interval, which is calculated as, $\hat{\theta} \pm 2S.E.(\hat{\theta})$ should cover the true population quantity of θ in 95/100 samples. If the method of calculating the 95% confidence interval is accurate this reduces the probability of Type I error i.e. the probability of wrongly rejecting a null hypothesis when it is true. In the presence of missing values the confidence intervals are valid only when the S.E. is estimated without bias.

The mechanism of missing data plays a major role in the selection of the appropriate method to handle the missing data. This chapter describes the mechanisms of missing data and the relationship between the missing data mechanism and the missing and observed data. The advantages and drawbacks of

various methods used for handling missing data, as well as applications of these methods in the literature, will also be explored. The chapter finally reviews available software for handling missing data.

2.2 Missing data mechanism

The most appropriate way to handle missing or incomplete data will depend upon how data points became missing (De Leeuw *et al.*, 2001). Little and Rubin (1987) define three types of missing data mechanisms, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

The complete data Y are defined as the dataset that would have been obtained if there had been no missing data. The collected data, which have a part missing, are called the incomplete data. Therefore

Complete data = incomplete data + missing data

To clarify the mechanism of missing data, I will introduce the response indicator R_{ij}, (i= 1, 2, 3,…,n), (j=1,2,3,…..,p), such that R_{ij} =1 if Y_{ij} is observed and R_{ij} = 0 if Y_{ij} is missing, (n is the number of rows in the dataset and p is the number of columns). The observed values of Y_{ij} can be denoted by Y_{obs}, the missing values by Y_{mis} .

The complete dataset Y can be defined as

$$Y= (Y_{obs} , Y_{mis})$$

The missing data mechanism can then be specified by a model for the response probabilities, which treats R as a random variable and defines the joint distribution of R and Y as

$$P(R_{ij}=1 \mid Y)=P(R_{ij}=1 \mid Y_{obs}, Y_{mis})$$

$$R_{ij}=\begin{cases} 1, & Y_{ij} \text{ is in } Y_{obs} \\ 0, & Y_{ij} \text{ is in } Y_{mis} \end{cases}$$

When the cause of missingness is known the missing data mechanism is said to be accessible, for example if data is missing in the outcome variable Y, and other variable X in the dataset hold information about why data was missing in the outcome, then the mechanism is accessible. When these causes are considered in the analysis the amount of bias can be reduced (Graham & Donaldson 1993). On the other hand inaccessible missing data mechanisms can be analysed by either guessing what one considers the true mechanism (Little and Rubin 1987; Rubin 1987), or by the collection of additional data to make the mechanism accessible (Graham and Donaldson, 1993). For example Huisman et al. (1998) reapproached respondents who did not answer all questions of questionnaire in a study on patients in a waiting list of orthopaedic practices (Krol, 1996). Of the 1,891 patients who took part in the study 1,330 (71%) of the patients responded, but 1,237 (93%) of the respondents had one or more missing values in the questionnaire. A sample of 435 (33%) of the non-respondents was selected to be re-approached either by telephone or mail. The method of re-approaching, which was not successful with every responder contacted, helped to obtain more

information on the missing data mechanism. The mean age of non-respondents was found to be higher than of respondents, the educational level of 46% of non-respondents was low and only 2% of them were highly educated. The 340 who were re-approached, 61% were females and 39% males. This information collected by reapproaching non respondents can later help in selecting the best method to handle missing data.

2.2.1 Missing completely at random (MCAR)

Data is said to be missing completely at random (MCAR) when there are no systematic differences between complete and incomplete records. Missing completely at random can be regarded as the simplest mechanism, as missing values are not related to observed values or missing values. Heitjan (1997) provides an example of MCAR: when a research associate shuffles raw data sheets and arbitrarily discards some of them. Another example of MCAR arises when investigators randomly assign research participants to complete two-thirds of a survey instrument. Graham et al. (1996) illustrate the use of planned missing data patterns of this type to gather responses to more survey items from fewer research participants than one ordinarily obtains from the standard survey completion paradigm in which every research participant receives and answers each survey question. Another example arises when weight and age are variables of interest in a particular survey if the probability that respondents provide their

weight regardless of their weight or age then the missing data is MCAR. MCAR can be described in notation form as

$$P(R_{ij} \mid Y_{obs}, Y_{mis}) = P(R_{ij})$$

Data are said to be MCAR when response probabilities are unrelated to any variables, or, in other words, missing data is a random sample of Y. Little (1988) developed a test for whether data are MCAR. The test compares the conditional distribution of $(Y_{obs} \mid R_{ij} = 1)$ to $(Y_{obs} \mid R_{ij} = 0)$. If there are differences between the two distributions, the assumption of MCAR is violated. This MCAR test is implemented in SPSS as a missing values analysis module. In the contest of our cohort this test can be implemented by performing a logistic regression on the outcome (missing/not missing). Any significant predictors of missingness suggests data are not MCAR, see Section 5.4.1. This mechanism of missing data is very unlikely to hold unless data are arranged to be missing by design. Definition and description of missing by design will be given in Section 2.3.

2.2.2 Missing at random (MAR)

Missing cases are said to be missing at random when incomplete data differ from cases with complete data, but the pattern of data missingness is traceable or predictable from other observed variables in the database rather than being due to the specific variable on which the data are missing. For example, if research participants with low income are less likely to return for follow-up sessions in a study that examines marital status over time, as a function of income, and the researcher measures income at the initial session, income can then be used to

predict the missingness pattern of the incomplete data. Another example is that older people are more likely to miss appointments in a study that assesses level of depression. In both of these examples, the actual variables where data are missing are not the cause of the incomplete data. The cause of the missing data is due to some other external influence. MAR means that the probability in Y_{mis} may be dependent on Y_{obs} but not Y_{mis}. In conditional probability terms

$$P(R_{ij} \mid Y_{obs}, Y_{mis}) = P(R_{ij} \mid Y_{obs})$$

Rubin (1987) defined the MCAR and MAR mechanism as ignorable missingness.

2.2.3 Missing not at random (MNAR)

Missing not at random (MNAR) which can be referred to as non-ignorable missing, applies when the pattern of data missingness is non-random, and it is not predictable from other variables in the database. If a participant in a weight-loss study does not attend a weigh-in due to informed concerns about lack of his weight loss, his data are missing due to non-ignorable factors. In contrast to the MAR situation outlined above, where data missingness is explainable by other measured variables in a study, non-ignorable missing data arise due to the data missingness pattern being explainable only by the very variable(s) on which the data are missing. In conditional probability terms

$$P(R_{ij} \mid Y_{obs}, Y_{mis}) \neq P(R_{ij} \mid Y_{obs})$$

2.3 Missing by design

Missing by design is the form of missing data, which is controlled by the researcher. The researcher can decide that part of the questionnaire is not applicable to some respondents; in this case this subset of respondents are asked to skip questions which are not relevant to them. In other cases of missing by design, the researcher decides to send the questionnaire to a subset of the sample, for example, to two thirds of the selected sample for a reason related to the research, such as aiming to reduce the overall response burden and cost. Graham *et al.* (1994) planned missing data in the design of the study using three questionnaire forms, which is a typical form of missing data by design. In this study a set of 130 questions was to be completed by child respondents. Three sets of questionaires were prepared each containing 100 questions out of the 130. The 130 questions were divided into a core block Q and alternative blocks A,B, and C, so that random subsamples of approximately equal size received one of three questionnaires as follows:-

- Questionnaire 1: Q A B
- Questionnaire 2: Q A C
- Questionnaire 3: Q B C

Using this technique, each block of questions from sets A, B, and C was presented to 2/3 of the respondents. This method maximized the total number of questions asked while maintaining a manageable number of questions for each child.

2.4 Pattern of missing data

A dataset with i rows (subjects) i=1,2…,n and j columns (variables) j=1,2…,p is said to have a univariate missing pattern if the data are missing in Y_{ij} for one variable while the rest of the variables are fully observed. A simple example of univariate missing would be a dataset of 1000 subjects. Each subject has five variables, sex (Y_{i1}), age (Y_{i2}), weight (Y_{i3}), height(Y_{i4}), and income(Y_{i5}) in this order, and missing values occur only on the height variable, see Figure 2.1(a). A monotone missing pattern, see Figure 2.1 (b), arises when Y_{im} is missing for a particular subject i then all subsequent variables, Y_{ik}, $k > m$ are also missing for that subject. If we consider the same dataset described above as an example that implies for a particular subject height and income will definitely be missing if weight was missing. Missing data are said to have arbitrary pattern when any sets of variables are missing for any subject, see Figure 2.1(c) for an example.

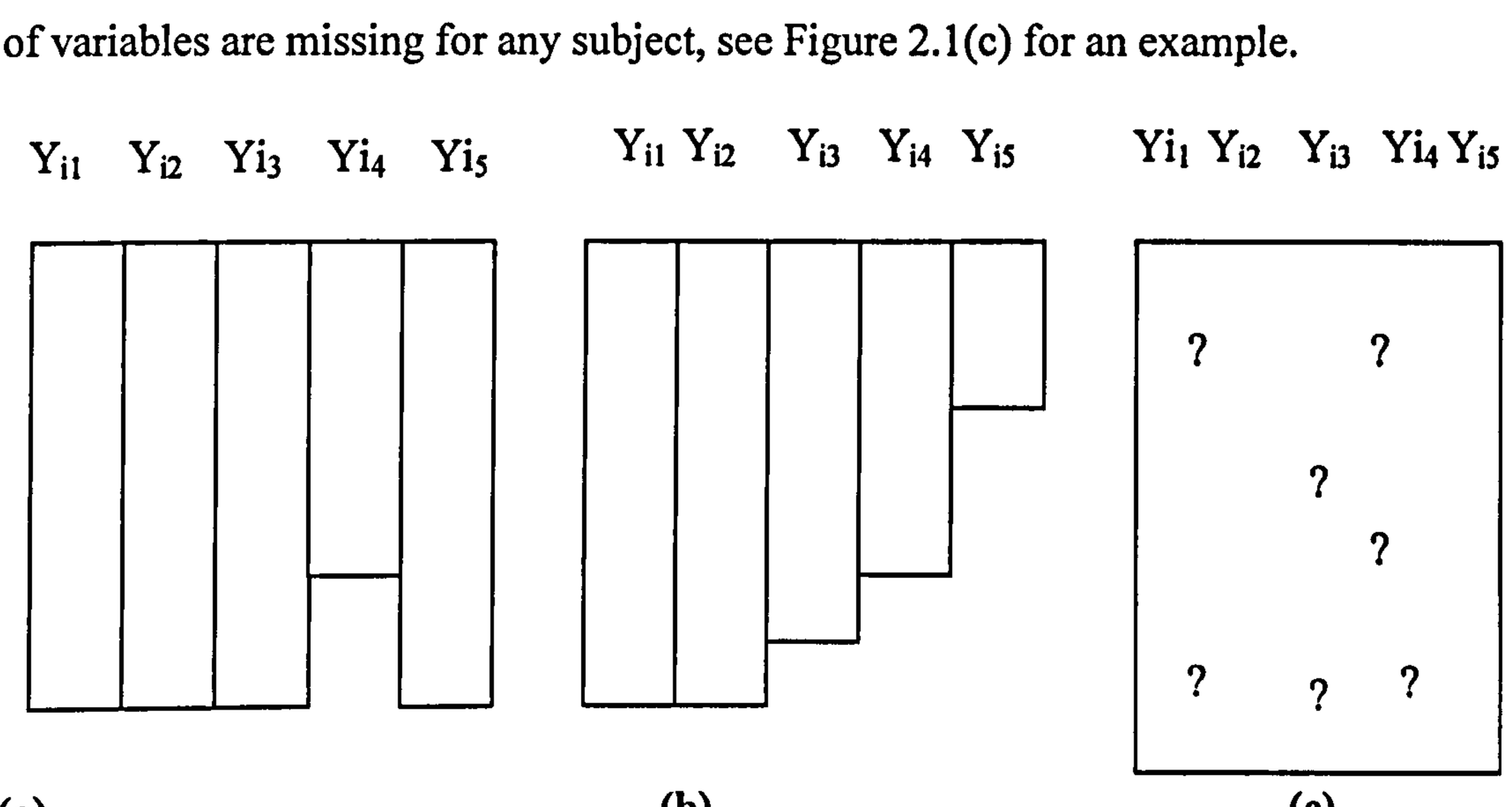


Figure 2.1: Patterns of non-response in rectangular datasets.

(a) univariate pattern, (b) monotone pattern, and (c) arbitrary pattern. In each case, rows correspond to observational units and columns correspond to variables. Adapted from;: Schafer, J. and Graham, John W. (2002). Missing data: our view of the state of the art. *Psychological Methods* 7, 147–177.

2.5 Handling missing data in the literature

When the researcher is faced with missing data and plans to analyse the incomplete dataset, the first priority should be to chose a method that is capable of maintaining or getting a close approximation of the original dataset. Missingness is usually a nuisance, not the main focus of inquiry in most fields of research. Roth, (1994) reviewed a random sample of 45 articles from *The Journal of Applied Psychology (*JAP), and 30 articles from *Personnel Psychology* (PP) between 1989 and 1991 on the methods used for handling missing data in the reported studies. He found interesting results; first it was difficult to understand how missing data were dealt with. Almost 42% of the articles in the *JAP* sample and 77% of the *PP* involving surveys did not mention if there were missing data on item level or the methods they used to deal with it. Second, in 37% of the *JAP* articles and 23% of the *PP* mentioned no need of any method to deal with missing data, suggesting that there was no missing data. The rest of the articles that mentioned using some sort of handling missing data did not discuss the topic of missing data. Only fifteen articles in the *JAP* discussed the issue of missing data, and six out of these discussed how they dealt with it. A similar review of published medical research examined randomised trials published between July and December 2001 in four major medical journals (BMJ, JAMA, The Lancet, and New England Journal of Medicine). Of the 63 trials with missing outcome data, 41 used complete case analysis, 14 studies used single imputation to handle missing data, these single imputation methods were as follows:-

- 7 last observation carried forward

- 5 worst-case

- 1 nearest value

- 1 regression imputation

One study used multiple imputation, 2 generalised estimating equations, and 3 used repeated measures analysis of variance (Wood *et al.*, 2004). No such review has been found in the field of nutrition.

2.6 Handling missing data in nutrition

The aim of this review is to provide a rigorous investigation of how missing data are dealt with in nutritional research.

The literature review was conducted systematically. The electronic search strategy was limited to a number of databases, which were Medline (1989 — present), EMBASE, a major biomedical database (1980 — present), PsycINFO (1985 — present). Within these databases all English language papers were included if abstracts included any of the key words or phrases in set 1 together with a key word from set 2.

Set 1: missing data, non-response, MCAR, missing completely at random, NMAR, not missing at random, MNAR, missing not at random, complete case analyses, complete case analysis, last value carried forward, listwise deletion, mean substitution, EM algorithm, multiple imputation

Set 2: nutrition, diet, nutrients, food frequency questionnaire

All studies were assessed for relevance, a total of 57 articles were retrieved. All retrieved references were checked for duplication and managed in the Reference Manager program (ISI ResearchSoft, Berkeley, CA). The precise strategy search is included in Appendix C.

In addition to this search, which identified all studies referring to missing data, studies published in the British Journal of Nutrition, the leading journal in this field in the UK, during July to December 2002 were also reviewed. This was to gauge the proportion of research articles that ignored the problem.

A large number of the articles mentioned missing data to report that it was excluded from the analysis. In a recent study Hsu *et al.* (2003) used the data from the National Health and Nutrition Examination Survey III, to study the prevalence of rheumatoid arthritis and hepatitis C in those aged 60 years and over. The study stated that 1,827 (27.7%) of the records were excluded because of missing data and all conclusions were based on complete cases. Sempos *et al.* (2000) and Gillum *et al.* (1996) applied a similar approach of excluding all records with parts missing.

Cowin, *et al.* (1999) studied the degree of underestimation of nutrient intake caused by missing data in two of the standard food tables used in the UK. Data were collected on 1,026 children aged 18 months, based on a 3-day dietary diary. Out of the 1,027 food items included in the analysis, 540 had missing values. Analysis on the complete cases was compared to the amount of nutrient intake after a guess was imputed for missing data. The study concluded that the

guesstimate altered the nutrient intake for 90% of the subjects. However, for some nutrients like B vitamins the mean percentage underestimate was small. This method appears to handle the problem of having a considerable amount of missing data. However, the method handles the missing data problem in a poor way. A guesstimate assumes that one knows exactly what the missing were. By imputing guesses, all information in incomplete records can be utilized. The main problem would be the reliability of such guesses, as it would not be possible to know if the same estimates can be replaced for missing data in other analysis applied to the same dataset.

Two studies assessed the potential bias due to non-response by comparing respondents to non-respondents, (Madigan *et al.*, 2000 and Turrell *et al.*, 2003). In the first study willing non-respondents completed a shorter interview. There was no evidence of difference between respondents and non-respondents with respect to smoking, family history of cancer and several dietary items. The study concluded that missing data would not affect the results and no attempt was made to handle the missing values before analysing the dataset. On the other hand, Turrell *et al.* (2003) studied survey participation and the error resulting from non-response in a population based-study which examined the relation between food purchasing behaviour and socio-economic status. The study reported a difference between respondents and non-respondents in their food purchasing behaviours as well as socio-demographic characteristics. The good effort of testing the difference between respondents and non-respondents was wasted. There was no attempt to re-analyse the data based on the additional information collected about non-respondents.

The National Health and Nutrition Examination (NHANES) are periodic surveys conducted in the USA to collect health and nutritional data on US population. Ezzati-Rice *et al.* (1995) used multiple imputation to account for missing data in 70 variables of NHANES III. This application was reviewed by Barnard *et al.* (1999), who concluded that although the application of multiple imputation is flexible in handling the missing data problem, caution is needed in creating the imputation models, and the missing data mechanism should be properly considered. Otherwise, and of with most other statistical methodology serious bias can result from analysing such imputed datasets. Hediger *et al.* (1999) also reported the use of multiple imputation to handle missing data. In this study the NHANES III dataset was used to compare young children who were born small for gestational age with those of appropriate size for their gestational age. Multiple imputation was used to handle 288 cases with missing values, or cases which reported unreasonable gestation. A regression model was used including race/ethnicity, infant's sex, mother's height and age, cigarette smoking, parity, family size, mother's race, and state of residence. For each missing value five imputed values were generated, using this model. A total of 267 values were imputed, leaving 21 cases missing because of simultaneous missing data in mother's height and cigarette smoking.

Multiple imputation has many attractive features. It allows the researcher to proceed with the analyses using exactly the same software that would be used if the data were complete. Information in the incomplete records can be used.

In studies that apply multivariable analysis, missing data are always a more serious problem, especially if values are missing for several covariates included in

the model. The analysis of complete cases can give biased estimates if missing is not completely at random. Zhao *et al.* (1996) applied a technique named as joint estimating equation (JEE), to handle missing data in a case-control study, which studied the link between diet and thyroid cancer. The joint estimating equation estimates regression coefficients, from linear and logistic regression models, provided that the missing data are either MCAR, or MAR. This technique is capable of using information in records with parts missing by imputing the estimates from regression coefficients for the incomplete variables, however the major drawbacks can be summarized as follows: -

- no attempt was made to investige the missing data mechanism
- Imputing values falling on the regressin line underestemates the varinace by ignornig the variability due to missing data.

This problem, and how it can be handled will be presented in Section 2.7.2.

Missing data were not mentioned in all the reviewed studies, which were published in the British Journal of Nutrition. There was no evidence of handling missing data even by simple methods like mean substitution and only one study applied single imputation using regression. This review suggests that methods used for handling missing data in the field of nutrition require more attention, and a lot of improvement.

2.7 Existing methods of handling missing data

Missing values are a problem in large-scale surveys with extensive questionnaires, because the motivation and concentration of the subjects is gradually eroded while completing the survey instrument, some items may be skipped, are difficult to interpret or they relate to the subject's specific circumstances, and the like.

The data collected by such questionnaires contain missing values. We refer to such data (databases) as *incomplete.*

When faced with missing data the researcher can use one of three techniques: -

1- Do nothing or analyse the available complete cases.

2- Generate a complete dataset by imputing or filling in the missing values, and then run the required statistical model on the complete dataset.

3- Use the more complex model based techniques, in which the missing data mechanism is included as part of the model.

Imputation methods are usually simpler but great precautions have to be taken to account for the uncertainty in the missing data; on the other hand, model based methods that are used to handle missing data, for example the Expectation-Maximization algorithm, see Section 2.7.3. The model-based techniques can be very efficient in solving the missing data problem for a specific model, taking into account the missing data mechanism. However, these types of models cannot be generalized to handle missing data in large databases, within which different analyses are to be carried out.

In this Section the most widely used methods of handling missing data will be explored, with their advantages and disadvantages, with more focus on imputation methods.

2.7.1 complete case analysis

Listwise deletion

The easiest solution of handling missing data is to exclude all records (cases) that are incomplete. This is referred to as the complete case analysis or listwise deletion. When the incomplete cases comprise only a small fraction of all cases (say, five percent or less) then complete-case analysis, i.e. the analysis of records with no missing values or complete information in all its variables, may be a perfectly reasonable solution to the missing-data problem.

One of the major disadvantages of this method is that it assumes that the missing data were MCAR, a strong assumption that is not realistic in most datasets. Little and Rubin (1987) demonstrated that if the incomplete cases differ systematically from complete cases, complete case analysis may lead to biased estimation of parameters.

The main advantage of this method is simplicity. In a study aimed to describe the systematic development and reproducibility of a FFQ, 539 students completed the questionnaires. Only 415 were included in the analysis as a result of missing data, see Buzzard et al (2001). Conclusions were based on this complete case analysis

with loss of 23% in sample size. No effort was made to check if the observed cases were different in any way from the incomplete cases, which were discarded.

Pairwise deletion

If a record has missing data for any one variable used in a particular analysis, the entire record is omitted from that specific analysis. For example if a dataset is made up of five variables, height, weight, smoking, body mass index, and alcohol, and in the first analysis only height and weight will be analyzed, then records with missing weight and height will be omitted from the analysis. This approach is implemented as the default method of handling incomplete data by many statistical procedures in commonly used software packages. In the UKWCS there are many confounding variables adjusted for in each analysis, leading to many records being excluded even using just pairwise deletion. Therefore, pairwise deletion is nearly as bad as listwise deletion, and neither is suitable for handling the UKWCS missing data problem.

2.7.2 Imputation

Researchers reached the conclusion that the obvious complete case analysis is not a viable option with the substantial presence of missing data. Imputation, the practice of 'filling in' missing data with some values, is an attractive approach to analysing incomplete data. It apparently solves the missing-data problem at the beginning of the analysis. Each missing observation is assigned a value; hence creating a completed dataset with no missing values. The completed dataset is

then analysed using standard statistical methods as if the new completed dataset made up of the observed and imputed data were the true one. However, a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests, as documented by Little and Rubin (1987); Schafer, (1997), and others. For instance, if each missing value is replaced by a default, such as `Never' in a question such as "how often, if ever do you drink alcohol?", the problem has seemingly disappeared. However, having imputed an extreme category will bring about bias in all the inferences, and overstate the precision of the estimates, because we pretend to possess more information that was in fact collected. Imputing a different value, such as the average frequency, leads to other distortions because the imputations reduce the dispersion of the values.

The most common problems with single imputation are that

- the sample size is overstated
- the resulting estimated variance of the parameters will be biased twoards zero as a result of ignoring uncertainity about the predicitons of the missing values
- the confidence intervals are too narrow

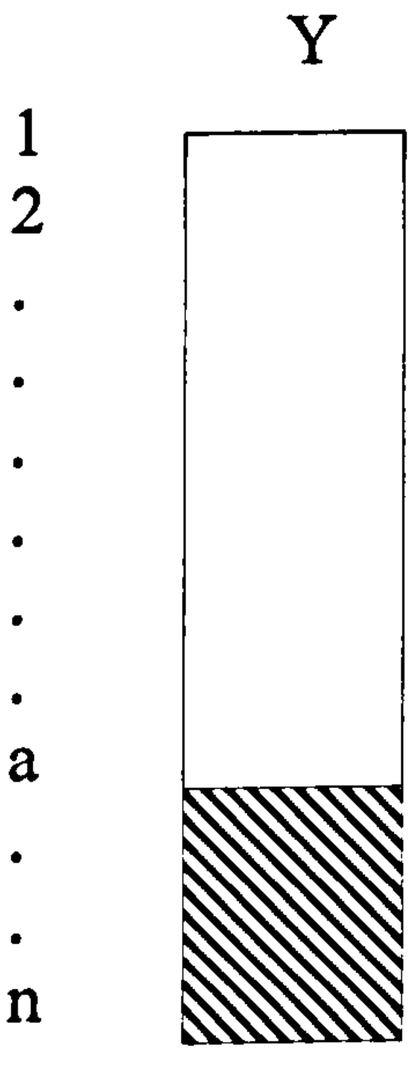
The above problems become worse as the amount of missing data and the number of parameters increase.

- Mean substitution

A particularly simple form of imputation is mean substitution, which is also referred to as unconditional mean (Little and Rubin, 1987). In a multivariable

dataset, each missing value may be replaced by the observed mean for that variable, see Figure 2.2 for demonstration. After the data have been altered, the research usually proceeds as if the omitted cases had never occurred, or as if the imputed values were the real data.

Mean substitution has some undesirable properties. First, the estimate of the sampling variance derived by the standard formula applied to the completed data is not valid. Since the sample size is effectively reduced by non-response, standard variance formulas underestimate the true variance. Second, estimates of quantities that are not linear in the data, such as the variance of Y or the correlation between a pair of variables, cannot be estimated consistently using standard complete-data methods, see Figure 2.3



$Y_1, Y_2, Y_3, \ldots\ldots\ldots\ldots Y_a$ Observed = Y_{obs}

$Y_{a+1}, Y_{a+2}, \ldots\ldots\ldots\ldots Y_n$ Missing = Y_{mis}

Replace each missing value with

$$\overline{Y}_{obs} = \sum_{i=1}^{a} \frac{1}{a} Y_{obs}$$

Figure 2.2: In a sample of n subjects the mean $\overline{Y}_{obs}$ of the observed is imputed for $Y_{a+1},\ldots,Y_n$ missing.

Third, imputing means distorts the empirical distribution of the sampled Y values see Figure 2.4. This is important when studying the shape of the distribution of Y using histograms or other plots of the data. Little (1992) showed that by assuming MCAR, the sample variance is biased by a factor $(n_{obs} - 1)/(n_{obs} + n_{mis} - 1)$, where n_{obs} is number of observed values, and n_{mis} is the number of missing values in a given variable.

The major deficiencies of the mean substitution can be summarized as follows: -

(a) It assumes that the data are MCAR.

(b) By substituting the mean for missing values the sample size will be overstated.

(c) Since the mean substitution will have zero variance, the variance and covariance will be underestimated, creating the false impression that we have more confidence in the completed data than we had for the observed data.

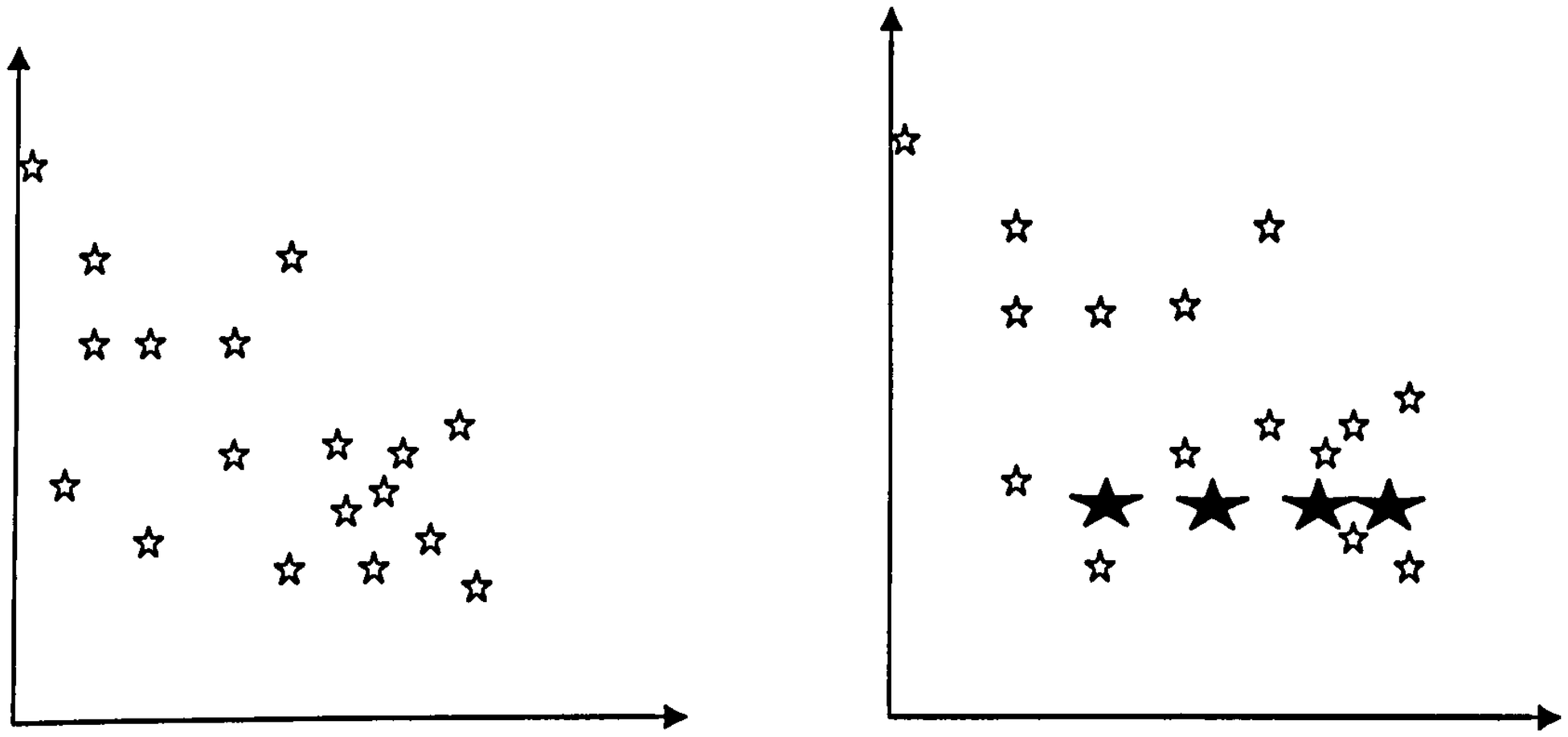


Figure 2.3: Illustration of how mean substitution corrupts covariances and correlations with other variables

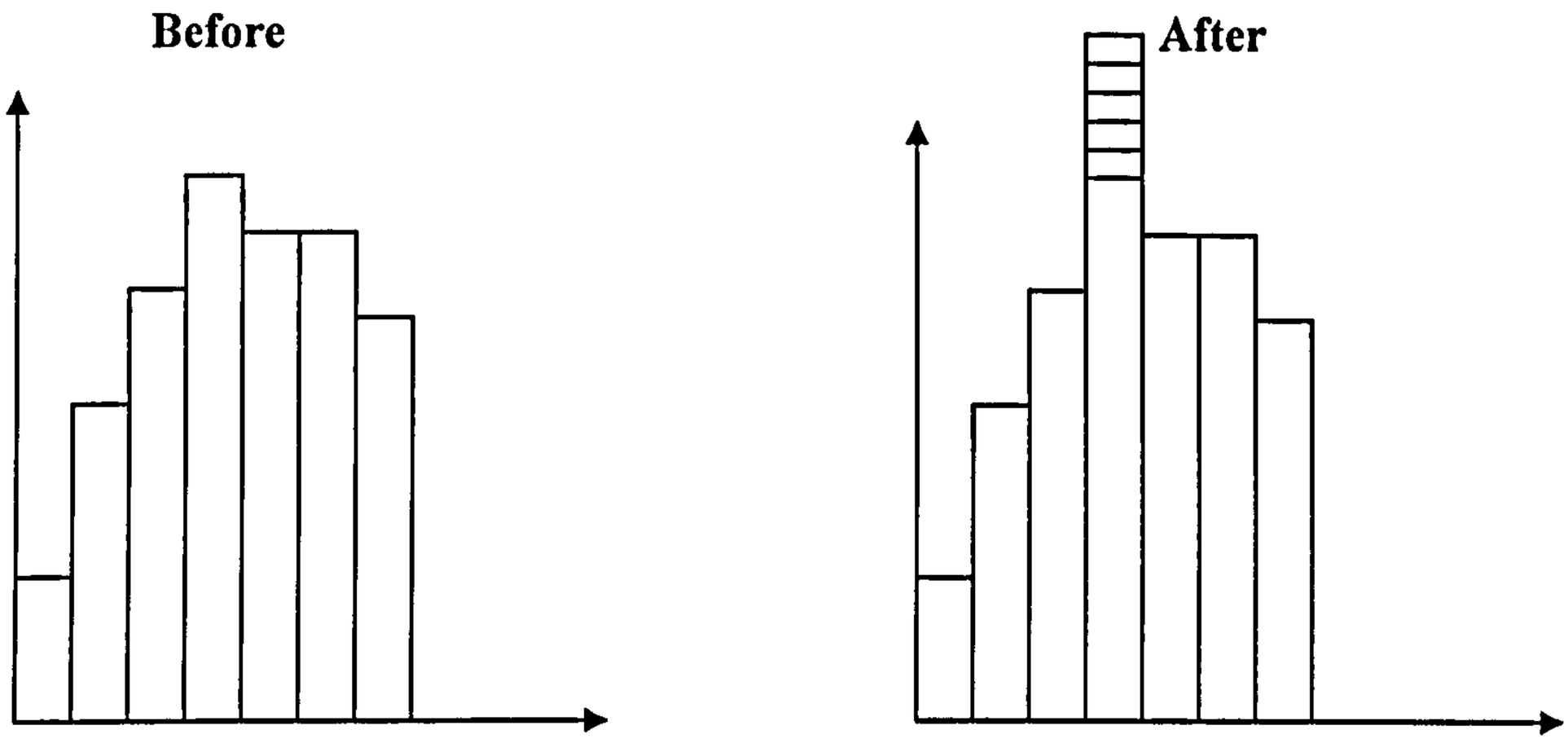


Figure 2.4: Illustration of how mean substitution corrupts marginal distribution of Y

- Indicator method

The indicator method is a simple idea developed before the new generation of fast computers (Jones, 1994; Chow, 1979). This method was first used in regression analysis, in which a missing value category is created for each incomplete independent categorical variable. The indicator takes the value 1, if the value is missing for that variable and zero otherwise. For example suppose the variable X_i, smoking is a categorical variable (3 categories), and some of its values are missing. The missing values of the variable in a regression analysis are replaced by $K_i + \beta_i X_i (1 - K_i)$ where the indicator K_i =1 if X_i is missing and zero otherwise. This is similar to creating an additional category for missing values. This method was widely used by epidemiologists for handling missing data because of its simplicity, but it was found to produce biased estimates under most

conditions (Greenland, 1995; Jones, 1994). The impact of this method depends on how the missing values are divided among the real categories, and how the probability of a value being missing depends on other variables. This method can lead to misleading results as very dissimilar classes may be lumped into one category.

- Regression methods

A seemingly better option is to fill in missing values conditional on other information in the data. One can construct a regression estimate for missing data based on observed data. Let us assume that Y_{1j} is fully observed for j=1,...,n and Y_{2j} is observed for j=1,…,a and missing for j= a+1, a+2,….n . Regress Y_{2j} on Y_{1j} for cases 1, ……,a. In the second step impute

$$\hat{Y}_{2j} = \hat{\beta}_0 + \hat{\beta}_1 Y_{1j} \qquad \text{for } j= a+1,...,n \qquad (2.1)$$

Little (1992) reviewed various methods of regression imputation, and stated, as have others, that the estimated standard errors of the regression coefficients on the completed data will tend to be too small as the imputation error is not taken into account. An improvement to this method, suggested by Little and Rubin (1987), is a stochastic regression method. In this case, the missing values Y_{2j}, j= a+1,…,n are

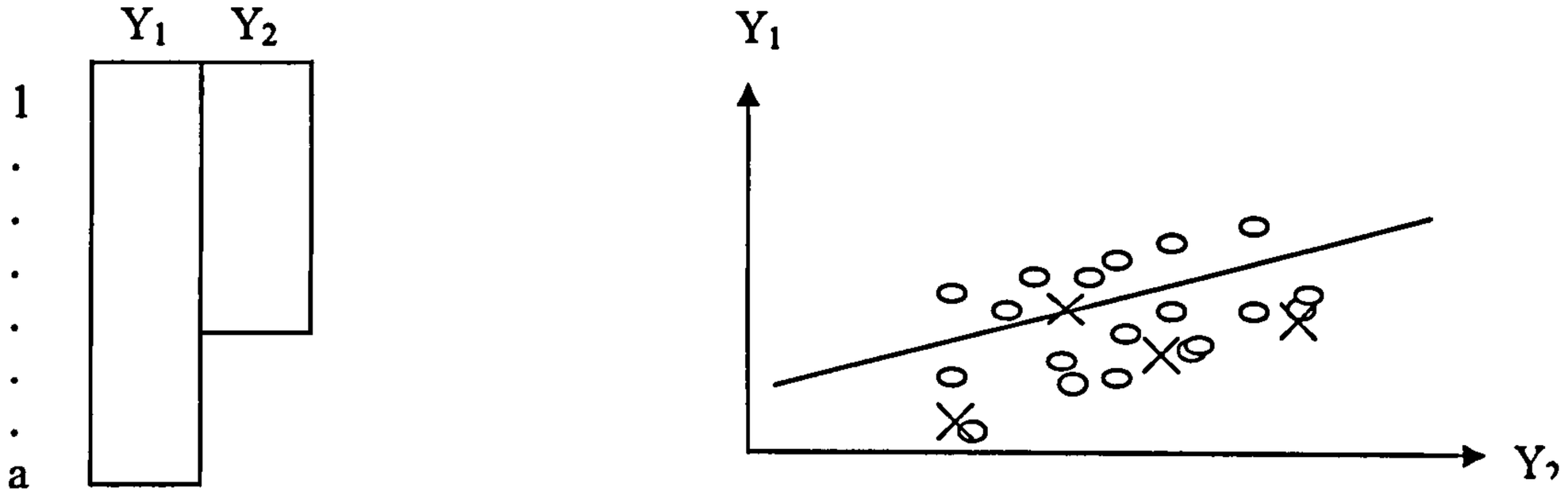


Figure 2.5: Imputing using regression method

replaced by $\tilde{Y}_{2j} = \hat{Y}_{2j} + \varepsilon_j$, where $\hat{Y}_{2j}$ is given by (2.1) and ε_j is the normal deviate with mean 0 and variance σ_j, where σ_j is estimated as the variance of the residual ε_j. The complete equation is

$$\tilde{Y}_{2j} = \hat{\beta}_0 + \hat{\beta}_1 Y_{1j} + \varepsilon_j \qquad (2.2)$$

This method is an improvement because the distributional characteristics of Y_{1j} and Y_{2j} are maintained (assuming MAR), but the variability due to imputation is still ignored, and so the completed-data estimates, although without bias, have too small estimated standard errors.

Hotdeck imputation

Hotdeck imputation identifies the most similar case (donor) within the dataset, to the case with a missing value (recipient), and substitutes the donor's Y_{obs} for the recipient Y_{mis}. The hotdeck refers to the deck of matching computer "cards" for the donors available for a respondent.

Hotdeck imputation has a long history of use, including years of use by the United States Census Bureau. Applications of hotdeck include sequential hotdeck used by the United States Bureau in the Current Population Survey (Bailar *et al.*, 1978),

the nearest neighbour hotdeck, in the Census of Construction in Canada (Colledge *et al.*, 1978), and another hotdeck procedure for the income supplement to the March Current Population Survey (Coder, 1978). Hotdeck imputation is superior to listwise deletion and mean substitution approaches to handling missing data. Among hotdeck's advantages are its conceptual simplicity, its maintenance of the proper measurement level of variables (categorical variables remain categorical and continuous variables remain continuous), and the availability of a complete data matrix at the end of the imputation process that can be analysed by complete analysis methods. One of hotdeck's disadvantages is the difficulty in defining "similarity"; there may be any number of ways to define what similarity is in any particular context. Thus, the hotdeck procedure can be considered as a simple approach to handling incomplete data. However, it requires customised software to perform the selection of donor cases and the subsequent imputation of missing values in the database.

Person	Sex	Age Group	Marital Status	Salary	Owns a house
1	M	2	S	50	N
2	F	1	M	80	Y
3	F	2	M	90	Y
4	M	2	S	60	-
5	M	2	M	40	N
6	F	1	M	30	-
7	M	1	M	30	N
8	M	2	S	-	-
9	F	2	S	100	Y
10	F	1	M	40	-

Table 2.1: Illustration of hotdeck imputation in an incomplete data matrix

Person 4: impute from person 1	owns a house = No
Person 6 impute from person 2	owns a house = Yes
Person 8 impute from person 1	salary = 50 , owns a house = No
Person 10 impute from person 2	owns car = Yes

More sophisticated hotdeck algorithms identify more than one similar record and then randomly select one of those available donor records to impute the missing value or use an average value if that is appropriate. Table 2.1 illustrates a simple example of hotdeck: -

(a) Person 8 can receive values *owns a house* = *No* and *salary* = *50*, from person 1 as both share being Males, in age group 2 and single for their marital status,

(b) Person 6 can receive the value *owns a house* = *Yes*, from person 2, as they share being Females, age group 1 and married.

(c) Person 10 impute from person 2, *owns a house* = *Yes*, as they share being female in age group 1 and both married.

Shen, et al (2000) compared the random hotdeck to complete case analysis and mean substitution in an artificial dataset generated by simulation. The dataset was developed following conditions encountered in Quality of Life data (QOL). The dataset consisted of k variables and 10,000 records which was drawn from a multivariate normal distribution with mean 3, and variance 0.8 and correlations varying between 0.2 and 0.8. A random sample of size 500 (records) was next drawn without replacement. Missing values were generated using independent Bernoulli trials with parameters q ($0 \leq q \geq 1$) as the probability of occurence of a missing value. Analysis was carried using three methods:-

- complete case analysis.
- mean substitution, in which missing values for each variable were imputed by the sample mean.

- the random hotdeck in which the missing entries of each complete unit were filled in by the counterparts of a randomly selected completely observed unit from the same dataset.

The experiment was repeated 1,000 times, for each level of missing probability q. The simulation study concluded that complete case analysis tends to give estimates with variance larger than the imputation methods. The mean substitution did not preserve the population distribution, as it was imputing values at the centre of the distribution. The mean imputation was found to underestimate the population variance. The study applied hotdeck imputation, in which a record with missing observation was filled in by the counterparts of a completely observed record selected at random from the dataset. The random hotdeck was found to give reasonable estimates for the population mean, while preserving the distribution of the population.The study might have benefited and thereby given better results if the selection, at random, were based on a pool of donors with complete records chosen for each recipient with incomplete records based on a definition of similarity.

The statistical package STATA version 7 (Stata Corp, 2000), has implemented a hotdeck procedure by tabulating the missing data pattern in a list of variables. A missing line is defined as a record from the list of variables with a missing value in any of its variables; a complete line is a row of the list of variables where all the data are observed. The hotdeck will then replace the variables in the missing line with the corresponding values in the complete line. Missing values are imputed stochastically rather than deterministically; therefore hotdeck should be used several times within a multiple imputation procedure. The variables with missing

values in each stratum of the data described are replaced by values sampled from variables with complete records in the same stratum. A bootstrap sample of complete records is sampled with replacement from the observed values, and the records with missing values are sampled at random (also with replacement) from this bootstrap sample.

The UKWCS data have missing values in almost every variable. Handling missing data by applying hotdeck in STATA may not be appropriate, as a row with missing values will be substituted for a complete row, i.e. even genuine recorded values are overwritten.

A simple demonstration of this is shown below. Let us assume that the data are made up of 3 variables

A	B	C
1	10	100
2	20	200
3	30	300
3	.	301
3	.	.

hotdeck missing for B using A.

A	B
1	10
2	20
3	30
3	30
3	30

hotdeck missing for B and C using A

A	B	C
1	10	100
2	20	200
3	30	300
3	30	300
3	30	300

From the above one can figure out that a valid value of C "301" is replaced by "300" because B is missing. Although this hotdeck procedure can result in a complete dataset, which is easy to analyse by complete data methods, however, such risk is not acceptable in a medical dataset as in our case. First, recorded values can be changed as a result of changing incomplete records with the complete ones; second the applied method underestimates variability due to missing data.

2.7.3 Model based methods

In the previous Section, imputation techniques that attempt to fill-in the missing data before fitting analysis model were introduced. A different technique is model-based methods. This technique fits the missing data mechanism as part of the model and hence account for the missing data by fitting models that are more complex. Imputation models are more straightforward, but they perform very badly if they fail to account for the missing data uncertainty. However, model based methods can be unattainable for very complex models.

- The EM algorithm

The Expectation-Maximization algorithm, Dempster, Laird and Rubin (1977) known as the EM algorithm, is an iterative technique for finding maximum likelihood estimates (MLE) when the data are incomplete or has missing values.

To clarify the concept of the EM algorithm it is first necessary to describe the *likelihood function.* The likelihood function is a different way of viewing the

probability function. The probability function assumes that the parameter (θ) is given while the likelihood function $L(\theta|y)$assumes the data (Y) is given. The likelihood and the log-likelihood $l(\theta|y)$ are the basis of estimating parameters given the data. The log-likelihood function links the data with the unknown parameters through a mathematical model and makes understood assumptions.

It is more appropriate to use the log-likelihood function $l(\theta|y)$, to provide an optimal basis for the estimation of parameters as well as their precision e.g. coefficients and standard errors etc. Since log Y is a strictly increasing function of Y, in order to maximize $L(\theta|y)$ it suffices to maximize log likelihood $l(\theta|y)$, which is generally easier, and while the shape of these two functions are different they have their maximum at exactly the same point.

In the presence of missing data Y is partly missing and partly observed. The EM algorithm maximizes $l(\theta|y)$ iteratively. Each iteration of the EM algorithm consists of an E-step (expectation-step) followed by an M-step (maximization-step). The E-step computes the expected values of the complete data sufficient statistics given the observed data. The M-step maximizes the loglikelihood computed at the E-step. The E-step followed by the M-step are repeated until convergence. Convergence is achieved when parmeters produced at the M-step are close to those computed at the previous M-step.

The EM algorithm can be efficiently applied when both the E and M steps are easy to implement. Despite its efficient ability to solve the problem of missing data it has the following drawbacks: -

- The algorithm can be extremly slow if a lot of data are missing (Laird, 1988; Little and Rubin, 1987), and in some cases it is difficult to judge whether convergence has occurred.
- The concept, and most of the material explaining it, is very complicated for a medical researcher, and its application requires a statistician with a strong mathematical background to be applied.
- In its analytic form it can involve difficult expectations.
- Standard errors are not directly available.
- No general code is available.

- Applications of the EM algorithm in the literature

This review explores the applications of the EM algorithm to handle missing data.

The literature was searched using the databases Medline (1989-present), Embase (1988-present) and PsycINFO (1985-present). Within these databases all English language papers were sought if the abstract included any of the keywords in set 1 together with the keyword "EM algorithm"

Set 1: missing data, non-response, MCAR, missing completely at random, missing at random, NMAR, missing not at random, MNAR, MAR

The precise search strategy is included in Appendix C. The database Web of Science was also searched for any abstract that included the keywords, "EM algorithm" and "missing data".

The first search retrieved 53 articles in the medical literature; the second search on the web of science retrieved 228 articles. Methodological articles as well as mathematical articles were covered. The web of science retrieved applications of

the EM-algorithm in marketing, political research as well as research in the field of economics, which were then excluded. All studies were assessed for relevance, and the references cited in each article were browsed for further relevant research. The key article referenced by most applications as well as methodological or mathematical research on the EM algorithm, is Dempster *et al.* (1977). The article gives a detailed description of the iterative computation of the maximum likelihood estimates when the data are not complete. Many methodological papers discussing the EM algorithm, as well as applications using simulated or original data, were then published after 1977. The EM algorithm however, was earlier presented by (McKendrick (1926); Hartley (1958); Woodbury (1971) and Sundberg (1976)); these articles have detailed discussion of the equations and applications of the algorithm. Meng (1997) examined the link between McKendrick's (1926) method and the EM algorithm. McKendrick (1926) was the earliest reference cited by Dempster *et al.* (1977), which defined and popularised the algorithm. The article used data from McKendrick (1926) on a cholera epidemic to demostrate the EM algorithm. The same author published another article celebrating 20th anniversary of Dempster's key article on the EM algorithm, Meng (1997), in which a modification of the algorithm was presented which gave faster convergence and less computational effort.

Applications of the algorithm in the medical literature are quite limited, and most of them are published in statistical journals like *Biometrics* and *Statistics in Medicine*. Meng (1997) illustrated the properties and the central ideas underlying the EM algorithm and its applications. The EM algorithm was used to obtain maximum likelihood estimates when one covariate is partially missing and the missing data mechanism is non-ignorable in an article presented by Lipsitz *et al.*

(1999). The method presented was applied to a dataset concerning quality of life of breast cancer patients in a clinical trial. The EM algorithm proved to be quite efficient in solving this missing data problem, but in medical research, missing data are usually missing in more than one covariate. Lipsitz *et al.* (1999) stated that extending the technique to missing data in many covariates would be rather complicated. This shows that the application, although quite effective, cannot deal with all missing data problems. A similar extended application was published by Schill and Drescher (1997), on a logistic regression model with missing data on several covariates. This study compared the application of the EM algorithm with three other approaches, the weighted pseudo-likelihood method of Flanders and Greenland (1991), the pseudo-conditional likelihood methods of Breslow and Cain (1988) and Schill (1993). The study concluded that although the EM algorithm is quite efficient it is quite difficult to implement when several covariates have missing values.

A few articles presented a modification of the EM algorithm for specific statistical models. For example Chen and Ibrahim (2001) discussed the application of the EM algorithm in a semi-parametric survival model with missing covariates, and applied the method to a dataset from a melanoma cancer clinical trial. Ibrahim (1990) provided a general method for estimating generalized linear regression models with incompletely observed covariates. The results were illustrated with two examples one using a logistic regression model and the other using a gamma regression. Ten years later Horton (2001) presented maximum likelihood estimation of logistic regression model with missing data in the covariates with available auxiliary information.

All these applications of the EM algorithm solved the problem of missing data in the covariate for a specific model such as logistic regression or survival analysis. This makes the EM algorithm an inappropriate solution for handling missing data in datasets intended for a multitude of purposes, as the algorithm has to be modified for each additional analysis.

Although all the above references simplified the idea by examples, they are too mathematical for a medical researcher to understand and they need a strong statistical background as well as strength in mathematics.

Although the EM algorithm is an efficient tool for dealing with missing data, it is difficult to implement when the sufficient statistics are not easily calculated, especially when their conditional expectations cannot be expressed in a closed form. For most of the applications we consider in this thesis, the EM algorithm is poorly suited, because our problems rarely have a short list of sufficient statistics, and often the complete-data solution is not by maximum likelihood. Another disadvantage of the EM algorithm is that it sometimes needs a large number of iterations to achieve convergence, and for some statistical models the M-step is complex and in cases impossible to formulate.

2.8 Multiple imputation

Many of the imputation methods described in Sections 2.7.1 and 2.7.2 are simple, but may lead to biased estimates as the uncertainty of missing values is not considered in most of them. On the other hand, model based-method such as the EM algorithm can be impractical for complex problems. The method of multiple imputation gets around the pretence of certainty about the missing values by

generating multiple completed datasets, and its application to complex problems is much simpler than the model-based methods. Multiple imputation is a three-step method:-

1. A set of m>1 completed datasets are created by imputing for the unobserved data m times using m independent draws from an imputation model. The imputation model is constructed to reasonably approximate the true distributional relationship between the unobserved data and the available information, and thus reduce potentially very serious non-response bias due to systematic difference between the observed data and the unobserved ones.
2. The m complete-data analyses are performed by treating each completed-dataset as a real complete dataset, using the procedures and software that would be appropriate on their own if the data were complete.
3. The results from the m completed-data analyses are combined in a simple, appropriate way to obtain imputation inference, which properly takes into account the uncertainty in the imputed values.

See Figure 2.6 for an explicit sketch of the method.

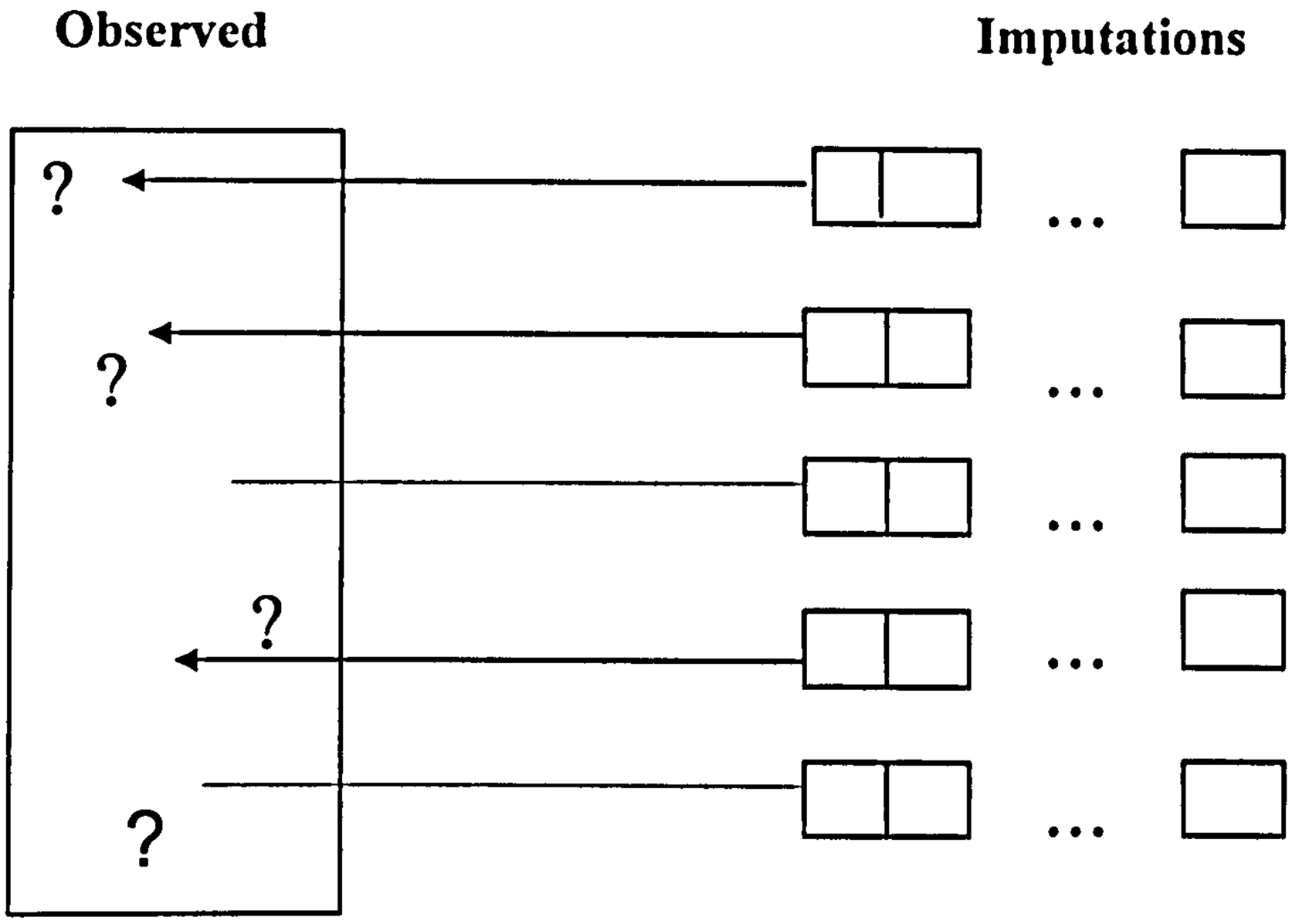


Figure 2.6: Multiple imputation replaces each missing value in the dataset by m imputed values.

Rubin (1987) describes a procedure for combining the m estimates from the m analysed datasets. Examples of these estimates are regression coefficient estimates or means.

To demonstrate this procedure let us assume that following the analysis of the m completed datasets, there are now m estimates $\hat{P}_j$, j=1,...m together with their sampling variance $\hat{s}^2_j$, j=1,...m. The mean of $\hat{P}_j$ is then given by

$$\overline{P} = \frac{1}{m}\sum_{j=1}^{m}\hat{P}_j \qquad (2.3)$$

The variability of $\overline{P}$ is divided into two components (Rubin, 1987). The within imputation variance

$$\overline{U} = \frac{1}{m}\sum_{j=1}^{m}\hat{s}^2_j \qquad (2.4)$$

and the between imputation variance,

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{P}_i - \overline{P}\right)^2 \qquad (2.5)$$

This between imputation variance is the additional variance due to uncertainity about the missing values.The total variance T is the sum of B and $\overline{U}$ corrected for m being finite by $\frac{m+1}{m}$

$$T = \overline{U} + (1+m^{-1})B \qquad (2.6)$$

The overall standard error is the square root of T. The confidence intervals are calculated by taking the average estimate plus or minus a number of standard errors,

$$\overline{P} \pm t_{df}\sqrt{T} \qquad (2.7)$$

where that number is a quantile of a t-distribution with degrees of freedom

$$df = (m-1)\left[1+\frac{\overline{U}}{(1+m^{-1})B}\right]^2 \qquad (2.8)$$

The relative increase in variance due to non-response is a comparison between the between imputation variance B and the estimated samling variance $\overline{U}$, this can be defined as :-

$$r = \frac{T-\overline{U}}{\overline{U}} = \frac{(1+\frac{1}{m})B}{\overline{U}}$$

The fraction of missing information (λ) about the estimate P due to non-response, compares the between imputation variance B with the total variance T. This is estimated as:-

$$\lambda = \frac{r + 2/(df + 3)}{r + 2}$$

Rubin (1987) show that the efficiency of an estimate based on m imputations is approximately $\left(1 + \frac{\lambda}{m}\right)^{-1}$ where λ is the proportion of missing information for the quantity being estimated. This proportion of missing information quantifies how much more accurate the estimate would have been if the data were complete. The efficiencies achieved for various values of m and rates of missing information are shown in Table 2.2. This effeciency depends on λ, and there is little advantage in increasing the number of imputations m beyond a small number. For example, if 50% of the information is missing, the effeciency of the estimate from five imputations is around 91%, doubling the number of imputations to 10 only increases the effeciency to 95%. This justifies that only few imputations are needed, usually in the range 3-10 imputations.

m	λ				
	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

Table 2.2: The proportion of missing information λ for the quantity estimated, number of imputations m and the efficiency of the multiple imputation procedure

The difficult and the most challenging part of the procedure is how to create the m-completed datasets. The imputed values are drawn from the estimated sampling distribution of the missing data given the observed data and the assumed model for missingness (Longford *et al.*, 2000). If the data are used for a specific analysis, the analyst can choose the model that would be easiest to implement and as close as possible to the particular analysis model. However, in public use datasets i.e. datasets that are to be used for more than one analysis and by more than one analyst, the imputation and the analyst models are incompatible most of the time.

The first assumption is that the data are MAR, i.e. the probability of missing values depends on the observed values (Y_{obs}), not on the missing values (Y_{mis}), as is discussed in Section 2.2.2.

There are often strong reasons why data may be missing not at random and testing the hypothesis that the mechanism is MAR is not easy. Although it is possible to formulate models for the data that are not missing at random (NMAR), these models are usually very complex. Rubin (1996) and Longford *et al.* (2000) suggested that introducing the maximum possible set of explanatory variables in the model for missing values could eliminate or reduce this difficulty. Even if the assumption of MAR does not hold, the procedure based on this assumption is less biased than the naïve method such as the complete case analysis, see Section 2.7.1.

The second assumption, is normality. An asssumption which can easily be violated in real data. Transformation of the variable can help in making the normality assumption more plausible (Box and Cox, 1964), for example a log transofrmantion of a positive variable will guarantee that the imputed variables

are always positive. Inference by multiple imputation may be robust for departures from the imputation model if the amount of missing information is not large, because the imputation model is effectively applied not to the entire dataset but only to its missing part. For example it may be quite reasonable to use a normal model to impute a variable that is ordinal (consisting of a small number of ordered categories), provided that the amount of missing data are not excessive. When using the normal model to impute for categorical data, however, the continuous imputed values should be rounded off to the nearest category to preserve the distributional properties as fully as possible and to make them intelligible to the analyst. According to Schafer (1997) the normal model, when used in this fashion, can be effective for imputing ordinal and even binary data in instances where constructing a more elaborate categorical-data model would be impractical (Ezzati-Rice, Khare and Schafer, 1993). Graham and Schafer (1999) used simulation in which nonnormal variables were imputed under normality assumptions, without any transformations. These simulations reported excellent performance with linear regression even when the sample size was small.

In this thesis the imputation of categorical and binary variables was further improved by the application of polytomous and logistic regression models for categorical and binary variables, discussion and application of these types of models are presented in Chapter 4 & 5.

2.8.1 Multiple imputation with categorical data

- The model for missing values

When a value is missing in a specific variable, the values of the other variables inform about the missing item. The model for missing values seeks to exploit such information, while reflecting the uncertainty about the missing item. I illustrate generating plausible values on an example of two questions, A and B. For simplicity, I assume that each frequency question has only three response options, 1, 2 and 3. Suppose a subject failed to respond to item A, but responded with B=2. I refer to this pattern of responses as (A=?, B=2). Similarly, I use the symbol + for an (unspecified) response. Throughout, I assume that there is a well-defined value, equal to one of the options given, in each instance of non-response. Suppose the cross-tabulation of the responses to the items A and B yields the Table 2.3

A	B			Total
	1	2	3	
1	100	25	20	145
2	10	155	30	195
3	20	20	200	240
Total	130	200	250	580

Table 2.3: Two frequency questions each with 3 categories

Since the vast majority of (responding) subjects with B=2 declared A=2, it would seem reasonable to choose A=2 also for the subject in question. This would be done by a typical simple imputation procedure. However, we cannot be certain that the genuine value of A is 2; after all, only 155/(25+155+20) = 77.5% of the subjects with complete responses to A and B and B=2 responded with A=2. So, a more appropriate way of filling in for A is to draw a value at random, with 25/200 = 12.5% chance of A=1, 77.5% chance of A=2 and 10% chance of A=3. With multiple imputation, a small preset number (say, five) of such draws is made for each subject with pattern of responses (A=? B=2). In this way, the uncertainty about the response to A would be appropriately reflected, if we were certain about the percentages 12.5, 77.5 and 10.0 for the three possible values of A.

- Reflecting uncertainty and MAR

In this method of generating replacements, we have made two assumptions. First, we have pretended that these three percentages are known, whereas they were merely estimated. Second, we have assumed that the pattern of values of A among those with A=? and B=2 is the same as among those with A= + (response) and B=2. The first assumption is not valid, but we can easily remedy it. The triplet of estimated percentages (12.5, 77.5, 10) is associated with uncertainty, which can be represented by sampling variation. For instance, the standard error associated with (A=1 | B=2) is

$$\sqrt{(0.125 \times 0.875)/200} = 0.0234$$

Since this is based on a large enough sample for the normal approximation to apply, the percentage of (A=1 | B=2) is estimated to have a normal distribution with mean 12.5% and standard deviation 2.34%. This sampling distribution is estimated similarly for the other categories of A given that B=2. Note that the three distributions are correlated because the three percentages have to add up to 100.

- Multinomial uncertainty

The uncertainty about the percentage is reflected as follows. For the first set of plausible values, we draw at random a triplet of plausible percentages for (A | B=2), and use these percentages to draw the plausible responses to A for all subjects with pattern (A=?, B=2). For the second and subsequent sets of plausible values, we draw, independently, other triplets of plausible percentages for (A | B=2), and use these percentages as above.

This procedure is repeated for the other patterns of missing values (A=?, B=1; A=?, B=2; A=?, B=3) and, by reversing the roles of A and B, also for (A=1, B=?; A=2, B=?; A=3, B=?). This will take care of all missing values on A and B except when both responses are missing. Since no information about each subject with this pattern is available, we consider the multinomial distribution with the nine categories implied by Table 2.3, draw a set of plausible probabilities for each set of plausible values, and use these probabilities to draw a response pattern for

A and B. This is in complete analogy with how we dealt with patterns (A=?, B=+) and (A=+, B=?).

The theory, given in detail in Rubin (1987) is rather complex, but implementing the method of multiple imputation is feasible in modern computing environments in which large-scale databases are stored and maintained. Depending on the number of missing values, the sets of plausible values can be generated immediately after the data have been collected and then used in every subsequent secondary analysis, or running the programme, which generates sets of plausible values for the required variables, may precede each analysis. In a compromise solution, plausible values are generated for the most frequently used (principal) variables, and a programme is provided for generating values for the other variables. The method has a version for continuous (normally distributed) variables and for datasets, which contain both categorical and continuous variables (Longford, 2000). This will be dealt with in Chapters 3, 4 and 5.

The second assumption made is that the distribution of A among subjects who responded to B is the same for those who responded to A (A=+) and those who did not (A=?). This is a key assumption for the validity of the method. Note that we can condition on several variables, not only on B. In general, the more variables we condition on, the better the chance that the two distributions are close to one another. This suggests that we should condition on as many variables, or use as fine a stratification (B), as is feasible.

A related issue is how to choose variables to condition on. First, we should prefer variables that are closely associated with A; second, we should prefer variables

that have fewer missing values; and third, we should prefer variables, which bring about MAR. The latter criterion is often difficult to check.

2.8.2 Multiple imputation with continuous data

For a dataset with incomplete continuous variable Y and a set of complete variables X_j, j=1,...m , the general recommendation for imputing for Y, is to use values generated by a model that conditions on observed variables X_j. Barnard and Meng (1999) mentioned that a balance should be kept in choosing an appropriate imputation model. A very simple model, might not reflect the data well, while a too complex model, i.e. a model with too many high order interactions, can be extremely difficult to implement and program. A more serious issue is that such complex models would have poor prediction, as they tend to 'over-fit' the existing data. One therefore should maintain a reasonable balance in fitting the imputation model. For continuous variables, plausible values can be generated using a regression model. In a regression imputation a regression model is fitted for Y_j given $X_1, X_2, \ldots, X_{j-1}$ complete cases j, j=1,...,m. When both X and Y are continuous using a linear model, see Figure 2.7.

$$Y_{ij} = \beta_{i0} + \beta_{i1}X_{i1} + \beta_{i2}X_{i2} + \ldots\ldots\beta_{ij-1}X_{ij-1} + \varepsilon_{ij}$$

the residuals are then computed for the complete cases

$$\hat{\varepsilon}_{ij} = Y_{ij} - (\hat{\beta}_{i0} + \hat{\beta}_{i1}X_{i1} + \ldots..\hat{\beta}_{ij-1}X_{ij-1})$$

and in the second step for each missing value of Y_{ij} a fitted value is computed

$$\hat{Y}_{ij} = \hat{\beta}_{i0} + \hat{\beta}_{i1}X_{i1} + \ldots..\hat{\beta}_{ij-1}X_{ij-1}$$

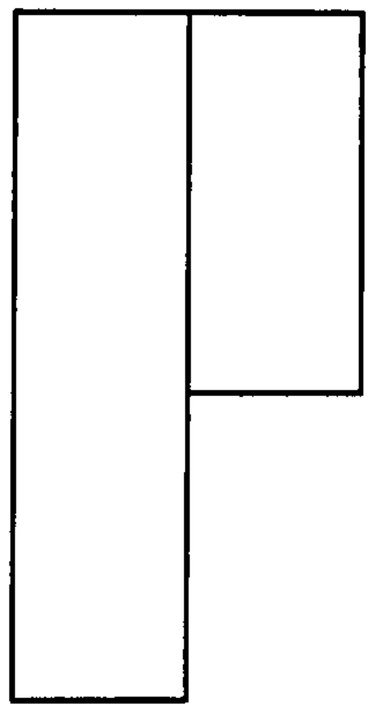


Figure 2.7: An example where Y values can be imputed from X values by regression imputation

To improve the regression model, uncertainty is added to the imputation of Y_{ij} so that the mean response is not always imputed. To achieve this $\varepsilon_{ij}^{\bullet}$ can be generated as a normally distributed random variable with mean 0 and variance $\hat{s}_t^{\,2}$, where $S_t^{\,2}$ is an estimate of var $(\hat{\varepsilon}_{ij})$. The missing values Y_{mis} are then replaced by the

$$\tilde{Y}_{ij} = \hat{Y}_{ij} + \varepsilon_{ij}^{\bullet}$$

In the ideal case for each missing data point, a model has to be fitted for the missing variable given the set of complete variables. This technique can easily be implemented using a regression model if the values are missing in one continuous variable. The difficulty arises in large datasets, which consists of a mixture of binary, continuous and categorical variables that are all incomplete. Regression imputation models will not be capable of handling the problem efficiently.

Modified methods that can handle this difficulty will be discussed later in this Chapter and implemented in Chapters 4 & 5.

2.9 The motivation for multiple imputation

A large number of researchers use databases constructed from surveys. These databases are frequently collected and designed to answer many questions. With missing data, each researcher should then decide on how to treat such incompleteness. Some of the analysts may not have access to key information to which the primary data collector has access. This key information (e.g. age, sex and ethnic group) can help to solve part of the missing data problem. Rubin's multiple imputation was set so that the primary investigator can solve the missing data problem for all future analysis. This is achieved by preparing m imputed datasets, with the help of key information available to primary investigator but not to all future researchers analysing all or part of the database.

These imputed datasets would be complete, and the analysts can then carryout straightforward analysis, with available software tools without the need to develop a missing data solution tailored to their particular analysis.

2.10 Bayesian approach to multiple imputation

In the past Bayesian statistics has been quite theoretical, until around two decades back it had been impossible to solve practical problems through the Baysian theory due to the complexity of the integrations involved. The new generation of computers with large storage and speed helped in solving practical Bayesian statistical problems. In the Bayesian theory for an unknown parameter θ a prior belief is condensed into a prior distribution $p(\theta)$. The likelihood function $L(y|\theta)$ produced from the collected data Y is then combined with the prior distribution to produce a posterior destribution for θ.

$$p(\theta|y) \propto p(\theta)L(y|\theta)$$

The prior distribution contains the prior knowledge about the unobserved parameters (which may be model parameters, missing data, or events which have not been directly or exactly observed), provided that enough information exists. However, it is perfectly possible for the prior to be vague, to the point of contributing little to the posterior distribution. The posterior distribution $p(\theta|y)$, is the subject of exploration in Bayesian inference. Any function of the posterior such as the mean, median, etc. is permitted. The posterior expectation of such a function $f(\theta)$ is defined in Gilks *et al.* (1996) as:

$$\mathrm{E}(f(\theta \mid y)) = \frac{\int f(\theta)p(y \mid \theta)p(\theta)d\theta}{\int p(y \mid \theta)p(\theta)d\theta}$$

Where Y are observed, θ are unknown parameters, and the denominator represents the marginal distribution of Y, sometimes called the prior predictive

distribution. The denominator generally need not be computed because it is independent of θ, (Gilkes et al.,1996 and Congdon, 2001). These integration have always proved difficulty for practical Bayesian inference. The Markov chain Monte Carlo methods (MCMC) avoid the complexity of the integrations of the Bayesian approach by producing simulated draws from the posterior distribution instead of calculating its exact form. The MCMC is a simulation-based procedure, in which instead of producing point estimates (frequentist methods) this proceeds in a sequence of iterations within which an estimate for each unknown parameter is produced. A large number of samples are drawn from the posterior distribution, sample means of these are then taken as an estimate of the population mean (Gelfand *et al.*, 1990 and Robert and Casella, 1999). More than one sequence of draws is required. Each one of this sequence of independent draws start at the same initial value for the parameter θ^0, and then values for $\theta^1, \theta^2, \theta^3$, etc. are drawn from a transitional distribution. The transition distribution for the draw θ^t, depends on the previous draw θ^{t-1}. An important aspect of the MCMC is that the transition distribution converge to one unique stationary distribution which is equal to the posterior distribution $p(\theta|y)$.

Schafer (1997) discussed the Bayesian approach to multiple imputation, which was first developed by Rubin (1987). When the data are not complete $Y=(Y_{mis}, Y_{obs})$, a multivariate normal model for the entire dataset is specified, conditional on completely observed variables. Imputations for the missing values are generated from a predictive distribution $P(Y_{mis}|Y_{obs})$ under an appropriate model, and the posterior distribution of the parameters. This is to say that multiple imputations works by averaging the complete data posterior over the predictive distribution of the missing values using MCMC

$$P(\mathrm{Y}_{mis} \mid \mathrm{Y}_{obs}) = \int P(\mathrm{Y}_{mis} \mid \mathrm{Y}_{obs}, \theta) P(\theta \mid \mathrm{Y}_{obs}) d\theta \qquad (2.9)$$

where

$$P(\theta \mid \mathrm{Y}_{obs}) \propto L(\theta \mid \mathrm{Y}_{obs}) p(\theta) \qquad (2.10)$$

for a prior distribution $p(\theta)$.

Gibbs sampling is one of the two most popular methods of MCMC. Other types of MCMC methods, which were proven to be useful in the analysis of incomplete multivariable data, are given by Gelfand and Smith (1990); Gelman and Rubin (1992a), Geyer (1992) and Smith and Roberts (1993). Applications of MCMC were also discussed by Gelfand *et al.* (1990), Casella and George (1992) and Smith and Roberts (1993).

2.10.1 Gibbs sampling

The Gibbs sampling was first introduced by Geman and Geman (1984). In the Gibbs sampling each parameter in a specified model is simulated in turn by conditioning on other parameters, while assuming that the current values of the other parameters are the true ones. Gibbs sampling is applicable when the joint distribution is not known explicitly but the conditional distribution of each variable is known, therefore it is well suited to cope with incomplete datasets. However, this suitability comes with the cost of more extensive computation. Let $Y=(Y_1, Y_2, \ldots\ldots Y_n)$ be a set of n random variables, each Y variable is partially observed, i.e. $Y_i=(Y_{obs}, Y_{mis})$, with i= 1,2,.....n . Let t denote an iteration counter. Assuming that the data are missing at random, the following iterations of Gibbs sampling sequence are repeated: -

For Y_1: draw imputations Y_1^{t+1} from $P(Y_1| Y_2^t, Y_3^t,.....Y_n^t)$

For Y_2: draw imputations Y_2^{t+1} from $P(Y_2| Y_1^{t+1}, Y_2^t,.....Y_n^t)$

.

.

For Y_n: draw imputations Y_n^{t+1} from $P(Y_n| Y_1^{t+1}, Y_2^{t+1},.....Y_{n-1}^{t+1})$

In this sequence one should condition each time on the most recently drawn values of all other variables. The initial iterations of the Markov Chain are influenced by the starting distribution, therefore they are discarded and this is known as the burn-in stage.The sequence is repeated until it reaches equilibrium state or convergence that is independent of the starting values. The common problem of MCMC methods relate to slow convergence, and the difficutly in assessing this in practice. A number of statistical packages were develped in recent years, applying MCMC methods. One of these is *MLwiN* a multilevel modelling software, which was originally developed for fitting hireachical models and introduced MCMC in its latest versions. The application of MCMC to handle missing data using *MLwiN* will be presented in Chapter 3. Multiple imputation by Chained equations, a technique capable of recovering missing data in covariates using Gibbs sampling will be discussed and applied in Chapter 4.

2.11 Applications of multiple imputation in the literature

The literature review was conducted systematically, to explore applications of multiple imputation. The electronic search strategy was limited to a number of databases, which were Medline (1966, February 2004), EMBASE a major biomedical database (1988, February 2004), PsycINFO (1985, March 2004), Web of Science (1988, February 2004). Within these databases all English language papers were included if abstracts included the keyword 'multiple imputation'. This search was further reduced to exclude articles not in the field of medical research. Nevertheless all methodological papers were taken into consideration. The final search retrieved 305 articles.

Table 2.4 show that handling missing data by multiple imputation became more popular in recent years. The retrieved literature covered applications in a range of research areas within the medical field. These included applications on clinical trials research, HIV research, cancer research, as well as many other epidemiological researches.

Publication Year	Number of publications
2003- Jan/2004	65
2000-2002	118
1997-1999	66
1994- 1996	31
Before 1994	25

Table 2.4: Number of publications of multiple imputation by year

In the same way that most applications in the EM algorithm referred to Dempster *et al.* (1977), the majority of applications on multiple imputation referred to Rubin (1987) and Little and Rubin (1987). These two books introduced the idea of multiple imputation, in which each missing value is replaced by m>1 simulated values generated before the analysis. Although the technique described in the books could solve the problem of missing data for the majority of applications in the medical field, understanding the mathematics could be very difficult for a non-mathematical researcher. The search showed that most of the publications on multiple imputation which were written in the early to mid 90's were methodological papers, published in statistical journals, some of these papers used data for demonstrations. Heitjan and Rubin (1990) demonstrated the efficiency of multiple imputation using a dataset from rural Tanzania (Kimati, 1985). In this dataset variables were partially observed, mothers reported age of their children rounded to the nearest year or half year. The authors described the mathematical theory of multiple imputation in detail. They then presented how the analysed multiple imputed datasets could obtain inferences that adjust for the coarseness of the data. Heitjan and Little (1991) also demonstrated multiple imputation using data collected by the Fatal Accident Reporting System in the USA. The database included information on location and time of accident, age, sex and the driving record of the driver as well as seat belt use and blood alcohol content. The last two variables had substantial proportions of missing data. The paper compared imputation of missing data by predictive mean matching, a method similar to hotdeck to complete case analysis and multiple imputation. Multiple imputation results provided substantial improvement over single imputation. Similar

methodological papers with applications were (Rubin and Schenker, 1991; Little, 1992 and Efron, 1994).

The complexity of the first two books might have been the main reason that multiple imputation was not widely applied until around ten years later when Schafer (1997) was published. This book described the methods in an easier format, and used examples to simplify the difficulty of the techniques.

Barnard and Meng (1999) reviewed three applications of multiple imputation in medical research. The first application was on data collected and maintained by health surveillance systems, inferences based on the data were suspected to be biased because of the incompleteness. These data are usually the only source for estimating prevalence, the rate of mortality as well as important features of common diseases. An example was the data collected by the AIDS surveillance systems. A fraction of the deaths among the recorded AIDS patients was never reported to CDC (Centres of Disease Control) in the USA. Therefore the survival time of all the reported cases, without death certificates might not reflect actual survival time. Only cases with reported deaths were used and those with deaths not reported were excluded, but there was a delay in reporting even among the reported deaths. Multiple imputation was used to handle the delay in reported deaths. The second application was to a longitudinal study with a randomised block design, where the missing data pattern for the variables was not monotone. In these applications a school choice was used for illustration. Multiple imputation was used to handle non-response in the United States National Health and Nutrition Surveys (NHANES). The study concluded that multiple imputation was an essential tool in handling missing data, in terms of statistical efficiency and

computational efficiency. However, the imputation model should be chosen with great care, and the missing data mechanism had to be considered.

In the last decade with the new generation of fast computers that have high storage capacity, and coinciding with researchers becoming aware of missing data and all its problems, there was tremendous growth in multiple imputation applications. In the retrieved publications few presented comparisons of simple methods of handling missing data to multiple imputation. Penny (1999) considered the application of multivariate outlier detection methods to multiply imputed laboratory datasets; the study tested the efficiency of multiple imputation in different proportions of missing data. Greenland and Finkle (1995) reviewed a number of methods of handling missing covariates in regression analysis data. The reviewed methods ranged from the simple mean imputation to multiple imputation, they concluded that results from multiple imputation was preferable but implementation was rather difficult because of the lack of software. Hunsberger (2001) applied three different methods of handling missing data to a multi-centre school-based trial, testing the efficacy of an obesity prevention intervention in American Indian children. The study first applied a multiple imputation procedure in which missing data were re-sampled from the observed data, the second procedure used a Wilcoxon rank sum test in which missing data in the intervention group received the worst ranks, and the last procedure was once more multiple imputation in which missing values were replaced with plausible values from a regression equation estimated from baseline values and follow-up data. The study then concluded that multiple imputation using the regression equation gave the best results. The search retrieved other studies,

which compared multiple imputation to simple methods of handling missing data e.g. (Xie *et al.*, 1997; Collins *et al.*, 2001; Arnold *et al.*, 2003).

Most HIV and AIDS clinical trials and cohort studies are designed to study change of the disease over time. The major problem of such studies is the drop out of patients over time. The search retrieved a number of applications of multiple imputation in this field. For example Wu *et al.* (2001) compared the application of multiple imputation implemented by the Gibbs sampler for estimating parameters in non-linear fixed effects model with missing covariates, with estimates obtained by the mean imputation method and the complete case analysis. The three methods of handling missing data were applied to modelling HIV viral dynamics from an AIDS clinical trial, the study concluded that the results from the multiple imputation were more reliable. Similar applications of multiple imputation in HIV research were published by Geskus (2001); Lyles (2001) and Touloumi *et al.* (2003).

It can be concluded that handling missing data by multiple imputation was very efficient in the majority of applications reviewed. The method was by far superior to any simple ad hoc method. However its implementation is rather complex, and the major complexity arises in specifying and fitting the imputation model. Data preparation and variable selection are time consuming. Another difficulty that faces researchers is the unavailability of software. Implementation of multiple imputation requires at least a basic knowledge of the Bayesian theory. However these difficulties can be sorted once and secondary analysts may not even notice the existence of these problems.

2.12 Software for handling missing data

When dealing with a large dataset, that contains some missing values, the major concern will be the software that can assist with the problem. I have already reviewed a number of methods for handling missing data that appeared in the literature. Most of the widely used statistical packages started introducing routines to handle the problem of missing data. For example most of the procedures in SAS exclude observations with any missing values from any analysis. Multiple imputation was then introduced in one of the SAS procedures in three steps. In the first step, m imputed datasets are created. The imputed datasets are then analyzed using standard procedures. The last step produce statistical inferences about the parameters of interest by combining the m results produced in the second step. SPSS also provide some imputation routines, which require the user to fit the individual complete data models.

Following the high demand for software to handle missing data by multiple imputation in the last years, more specialized packages were developed: -

- MICE (Multiple Imputation by Chained Equation) was produced by a group in the Netherlands working on applications of multiple imputation in public health. MICE was written using S-Plus V4.5 and S-Plus 2000 for Windows. A number of papers document the method (Van Buuren *et al.*, 1999; Brand, 1999). The chained equation implemented in MICE requires assumptions about the multivariate posterior distribution. Nevertheless, it is not always certain that such a distribution exists (Van Buuren, 1999).

- SOLAS is another widely used package, which was designed for the analysis of datasets with missing values. The package performs imputation for missing data by a number of methods, including the last value carried forward (LVCF), hotdeck imputation and multiple imputation, using propensity score models, a method which is beyond the scope of this thesis. Once the multiple datasets are created, SOLAS provides summary statistics and combines the results from the multiple analyses. A limitation of SOLAS is that although it handles linear regression it lacks the ability to handle non-linear regression, such as logistic or survival models (Horton and Lipistz, 2001).
- NORM applies multiple imputation routines of multivariate continuous data under a normal model. Computational routines used in NORM are described by Schafer (1997).
- *MLwiN*, the multilevel modelling software, introduced multiple imputation by Markov Chain Monte Carlo methods in its latest version. The software uses Gibbs sampling for the generation of imputed values. This software will be applied and compared to other methods of dealing with missing data in Chapter 3.

Other packages that provide some support to imputation include AMELIA, IVEWARE, and EMCOV. The statistical package STATA implements routines for single imputation using linear regression and hotdeck imputation, see Section 2.7.2 for hotdeck in STATA.

NORM has some limitations in transforming non-normal variables. Although the NORM package has a help menu, for all its routines, it was found not flexible enough to handle the UKWCS missing data. The data as will be discussed in the

following chapter is a mixture of continuous and categorical variables, and the questionnaire itself is divided into two Sections FFQ questions and long questions. Different format of variables required different consideration. To provide necessary flexibility for the UKWCS, all the routines used in this thesis for multiple imputation, were programmed using STAT A 8, see Appendix C.

2.13 Discussion

The problem of missing data should be avoided whenever possible, by close follow-up, and repeated calls. Nevertheless, if all efforts fail, doing nothing about missing data, or in other words the complete case analysis described in Section 2.7.2 is not an acceptable option. Analysing complete cases will reduce the sample substantially, and lead to serious bias as the missing sample might differ from the complete cases. Single imputation can be a solution when the fraction of the missing data is very small, and there is substantial evidence that the missing data does not differ from the observed or MCAR.

EM and multiple imputation are the only efficient options for missing data. EM can be the solution if the problem is simple. Finding the maximum likelihood can sometimes be rather difficult, and the EM algorithm can be very complex when the sufficient statistics is not easily calculated.

By multiple imputation all the information in the incomplete records is used. The application of multiple imputation can be complex nevertheless it leads to valid inferences. With the great improvement in computing environment, faster capabilities and greater storage, multiple imputed datasets can be stored for any

secondary analysis. Software for the developments of multiply imputed datasets and the analyses of imputed datasets are needed, as today's available software might not satisfy individual needs.

As a result of a great demand for handling missing data especially in the medical field, many statistical packages have implemented routines to handle missing data. Furthermore, specialized statistical packages for multiple imputation, which range from expensive to free packages that can be downloaded from the web, became available in recent years. Most of these routines had some limitations. Handling missing data in the UKWCS was in need of careful consideration for its large number of variables and their diversities. Therefore, all the routines used in this thesis were written using STATA 8, to handle the complexity of incomplete outcomes and covariates, as well as the different imputation models that are required to generate imputations for continuous, categorical and binary variables.

Chapter 3

Alcohol consumption

3.1 Introduction

Food supplies energy and provides essential nutrients needed for body functions. The three basic nutritional components of food are protein, carbohydrates and fats, known as the macronutrients. After being converted into simpler products, the body uses them as a source of energy.

Unlike protein, fats and carbohydrates, alcohol is not essential to the body. There has been a long debate on alcohol and its effect on the body. Moderate consumption of wine was found to have a beneficial effect and to protect from coronary heart disease and cancer (Gronbaek *et al.*, 2000). The joint WHO/FAO, (2003) report also stated that low to moderate consumption of alcohol lowers the risk of coronary heart disease. However, other cardiovascular and health risks associated with alcohol do not favour a general recommendation for its use. In developed countries alcohol is considered as one of the main risk factors for cancers of the oral cavity, pharynx and oesophagus and 75% of such cancers are attributed to alcohol and tobacco (International Agency for Research in Cancer, 1990). The same report mentioned that excessive alcohol consumption is the main diet risk factor related to cancer of the liver. Studies also found that alcohol is the main dietary factor which increases the risk of breast cancer, with around 10% increase in the risk for an average one alcoholic drink per day (Smith-Warner and

Spiegelman, 1998). One in four men and one in ten women are believed to be drinking above the sensible limit (Paton, 1994). However, drinking is often under-reported in surveys. For most communities, alcoholism could be regarded as a stigma, and one would expect it to be even harder for women to admit alcohol problems if they had any, so care had to be taken in questions on alcohol consumption. These questions should be asked in the same manner as other questions. They may be combined with lifestyle questions, such as smoking; diet and exercise; see for example Paton (1994).

Information on alcohol consumption was collected in two parts of the UKWCS questionnaire, in two different ways. The first part of the alcohol consumption questions consisted of a block of five items in the FFQ. For each item there were ten response options ranging from "never" (coded as 0) to "six or more times per day" (coded 9), in response to the question:

"How often have you eaten these foods in the last 12 months" see Figure 3.1.

ALCOHOLIC BEVERAGES										
Wines (wineglassful)	0	1	2	3	4	5	6	7	8	9
Beer, Lager (half pint)	0	1	2	3	4	5	6	7	8	9
Cider (half pint)	0	1	2	3	4	5	6	7	8	9
Port, Sherry, Liqueurs (glass)	0	1	2	3	4	5	6	7	8	9
Spirits e.g. Whisky, Gin, Vodka, Brandy (single/1 measure)	0	1	2	3	4	5	6	7	8	9

Figure 3.1: Alcoholic beverages block of the FFQ

In the second part of the questionnaire, the question on alcohol consumption consisted of three parts, in the format of long questions on the amount of alcohol consumed per week; the questions read as follows:

(A) In a typical week, how much do you drink?

More than once a week ☐ Once a week ☐

Less than once a week ☐ Never drink alcohol ☐

(B) In a typical week, how much do you drink?

Beer or cider (pints per week) ☐

Wine (glasses each week) ☐

Sherry/Fortified Wines (glasses each week) ☐

Spirits (glasses (singles) per week) ☐

(C) Five years ago, how many alcoholic drinks did you have each week?

Beer or cider (pints per week) ☐

Wine (glasses each week) ☐

Sherry/Fortified Wines (glasses each week) ☐

Spirits (glasses (singles) per week) ☐

These questions were used for two purposes:

- As a crosscheck to the reported alcohol consumption of the FFQ, see Section 1.2.

- To provide a more detailed record of this important energy source.

Alcohol questionnaire items were presented in a different way in the two methods; for example, the second item of the alcohol consumption of the FFQ Figure 3.1 dealt with beer, lager (half pint) and the third item was cider (half pint), while the first item of the average weekly recall (B and C) was beer or cider combined as a single item.

One aim of collecting information on alcohol consumption was to measure the impact of alcohol on health. It was also important to find the magnitude of its contribution to the total energy intake for nutritional analysis.

Estimates of overall nutrient intake were calculated by adding up the product of the reported frequency of each food by the amount of nutrient in a specified portion of that food. The total alcohol nutrient intake was found by adding up the intake of the different types of alcohol consumed per week. For example, the total alcohol nutrient intake of a subject who reports consuming 2 pints of beer, 3 glasses of wine, 2 glasses of sherry and a glass of spirit per week, can be found by summing the amount of alcohol nutrient in 2 pints of beer, 3 glasses of wine plus the amount of alcohol nutrient in a glass of spirit and 2 glasses of sherry, i.e.,

$$\text{Total alcohol intake} = (\{2 \times 287\}/100 \times 3.08) + (\{3 \times 125\}/100 \times 9.25) + (\{2 \times 40\}/100 \times 16.65) + (\{23/100\} \times 31.70)$$

$$= 72.98 \text{ g/week}$$

Here 287, 125, 40 and 23 are the quantities in grams, of alcohol nutrient in a pint of beer, and a glass of wine, sherry and spirit respectively. 3.08, 9.25, 16.65 and

31.70 are the quantities of the nutrient in 100 grams of beer, wine, sherry and spirits, respectively (Table 3.1).

Alcohol	Grams per pint/glass	Alcohol nutrient/100g
Wine	125	9.25
Beer	287	3.08
Cider	287	5.98
Spirit	23	16.65
Sherry	40	31.70

Table 3.1: Alcohol nutrient in a pint of beer and a glass of cider, spirit and sherry

In this chapter, the two blocks of alcohol consumption are investigated. The impact of missing data, their patterns, and how different methods of handling missing data can affect the results are discussed. The response rates for the different questions, and their dependence on the format of the question were compared. In Section 3.5 the impact of handling missing data by imputing zeros, the default value imputed in missing items by nutritionists, were compared to the complete case analysis and multiple imputation. For simplicity, the alcohol consumption questions of the FFQ were referred to as the FFQ questions (Figure 3.1) and the three parts of the alcohol consumption long question as (A), (B) and (C), defined in this Section.

3.2 Missing data in alcohol consumption (FFQ)

The total sample size of the UKWCS is 35,367. For the alcohol consumption block of the FFQ, Table 3.2 showed that the largest frequency of missing values occurred for cider where 1.0% of the values were missing. For the total consumption of alcohol nutrient, all five variables are used. Only 34,840 subjects had complete records on these variables, which accounted for 98.5% of the sample. This response rate was very satisfactory and the loss of information due to missing data was negligible.

Missing values might have occurred accidentally, i.e. respondent missed the question, or forgot to tick the appropriate box, because subjects did not want to reveal their alcohol consumption, or because they confused "Never" with no response. The 1.5% loss of information was quite low and does not need complex imputation to account for the loss of information from the incomplete records. One could assume that missing responses were meant to be coded zero or "never", i.e. respondents left the questions blank to say that they did not drink that type of alcohol. That suggests the simple solution to impute zeros for every missing alcohol item. It is reasonable to assume that the 1.5% of the sample loss has a negligible impact on any data summary, even if the subjects with incomplete records on alcohol differ systematically from those with complete records.

	Wine	Beer	Cider	Sherry	Spirits
Never	4,765	19,483	25,080	13,865	14,405
Less than once a month	4,805	5,972	6,674	11,229	7,850
1-3 per month	4,843	3,666	1,676	4,527	4,420
Once a week	4,722	2,762	793	2,684	3,007
2-4 per week	7,156	2,044	496	1,732	3,080
5-6 per week	3,463	602	133	356	970
Once per day	3,020	324	86	564	911
2-3 per day	2,213	170	45	102	426
4-5 per day	175	44	10	5	56
6+ per day	34	11	4	9	22
Missing (Missing %)	171 (0.5%)	289 (0.8%)	370 (1.0%)	294 (0.8%)	220 (0.6%)

Table 3.2: Observed distributions of responses to Alcohol consumption block of FFQ

Schafer (1997) stated that when the incomplete cases amount to a small percentage of all cases, say five percent or less, case deletion may be a perfectly reasonable solution to the problem of missing data.

Table 3.3 shows that the mean and standard deviation of alcohol nutrient intake per week had changed only slightly by filling in zeros for the missing values. This could therefore be a satisfactory solution in this case. The same process of handling non-response, can lead to biased results in other situations see Section 3.5.

	Complete cases	Impute zero
N	34,840	35,367
Mean	13.20	13.16
Std. Dev.	11.59	11.59

Table 3.3: Imputing zero for missing values in alcohol consumption of the FFQ

3.3 Missing data in alcohol consumption (long questions)

The long questions on alcohol consumption (A, B and C), described in Section 3.1, were constructed as crosscheck questions to the FFQ. The response rate to question B was poorer than to the same question of the FFQ, see Table 3.4. The rate of non-response ranged from 18% for wines to more than 52% for beer and cider. For question (C), alcohol consumption before 5 years, response rate was just as poor, this ranged from 18% for wine consumption to 52% for the consumption of beer and cider, see Table 3.4.

One could think of many reasons for the poor response rates to these questions: -

- The format of the questions may have been one of the major causes of the different rates; it must have been much easier for the respondents to tick the frequency of their alcohol consumption than filling in the number of glasses or pints consumed.
- The women may not be able to remember the exact amount, and so left their answer blank
- Beer or cider was combined as one category in questions B and C, compared to two separate categories on the FFQ.
- Maybe subjects did not want to repeat themselves, and thought that they already submitted enough information about their alcohol consumption in the FFQ.
- Collecting information on alcohol had always been considered a difficult task, and it is generally more embarrassing for women responders to reveal their alcohol consumption habits precisely.

To understand the impact of imputing values on the basis of several simple criteria combined and the effect on reducing missingness, a combination of three different types of single imputation was applied to these crosscheck questions in Sections 3.3.1, 3.3.2 and 3.3.3.

For simplicity, variables of question B were referred to as *B-beer, B-wine, B-sherry*, and *B-spirits*, variables of question C were referred to as *C-beer, C-wine, C-sherry and C-spirits.*

	B		C	
Alcohol	**Recorded**	**Missing**	**Recorded**	**Missing**
Beer or Cider	16,877	18,490 (52.3%)	16,973	18,394 (52.0%)
Wine	28,937	6,430 (18.2%)	28,879	6,488 (18.3%)
Sherry	17,122	18,245 (51.6%)	17,459	17,908 (50.6%)
Spirits	27,620	7,747 (21.9%)	20,629	14,738 (41.7%)

Table 3.4: Response frequencies and percentages to question B

3.3.1 Row borrowing

Questions B and C were on drinking habits now and five years ago respectively. It is reasonable to assume that drinking habits do not change greatly in 5 years and especially at this age of respondents 35-69 years, this assumption was also supported by the strong association between each pair of alcohol type (now and

five years ago), see Table 3.8, and that subjects may have been inclined not to fill in the same number twice.

For all those who completed question B, and left all of question C missing, values for question C were borrowed from the relevant answers to question B, 96 values were imputed in this way. The same type of "borrowing" was carried out for people who completed the 4 types of alcohol consumption in question C and left the whole of question B missing. In this case 38 values were borrowed. In total 144 missing values were imputed, which accounts for a very small percentage of the total missing values in these two questions, see Table 3.4.

This type of imputation could be considered as a simple form of hotdeck imputation, and was referred to as row borrowing in this Chapter.

3.3.2 Column borrowing

It was assumed that respondents who left the entire second question of alcohol consumption i.e. both B and C blank and filled their counterparts of the FFQ by zeros meant to say that they didn't drink that type of alcohol. Hence, zeros were imputed for *B-beer* and *C-beer* if *beer* or c*ider* variables of the FFQ were filled in as "never", coded 0 and *B-beer*, *C-beer* were both missing. The same procedure was applied to *B-wine, C-wine, B-sherry*, *C-sherry* and *B-spirits*, *C-spirits.* By this imputation the number of missing values was reduced considerably, for example more than 2,000 values were imputed for *B-beer*, 8,000 values for *B-sherry* and 6,000 values for *C-spirit,* see Table 3.5. This step was referred to as column borrowing.

	B		C	
Alcohol	**Recorded**	**Missing**	**Recorded**	**Missing**
Beer or Cider	19,257	16,110 (45.6%)	19,353	9,271 (26.2%)
Wine	29,064	6,303 (17.8%)	29,006	5,597 (15.8%)
Sherry	18,570	10,490 (47.5%)	18,899	10,560 (29.9%)
Spirits	27,620	7,747 (21.9%)	26,160	14,738 (41.7%)

Table 3.5: Column borrowing for alcohol items if response was "never" in the same alcohol item of the FFQ.

The two steps of single imputation (row borrowing and column borrowing) reduced the amount of missing data in the alcohol consumption block. Percentage of missing was reduced from 52.3% to 45.6% for beer and cider, from 18.2% to 17.8% for wine, from 51.6% to 47.5% for sherry and the missing percentage for spirits stayed the same at 21.9%. This showed that only minor amount of information was gained, based on the above logical decision.

3.3.3 Imputing means for the cells

Mean substitution, is a common method used in a number of statistical packages, see Section 2.7.3. This method replaces missing data with the average for the specific variable. In this Section this method is modified by instead of substituting the same number (average) for every missing value in the variable, for each code

of response to each alcohol type from the FFQ, the means of the responses to the counterpart in question B or C were calculated and imputed for missing values in questions B and C. To simplify this I will describe imputation of beer as an example, the imputation proceeded in the following steps: -

- In the FFQ 3,666 responded with the code 2 (1-3 per month) for consumption of beer, see Table 3.6.
- The mean quantity of these 3,666 responses for *B-beer* was then calculated.
- For every subject who filled in the code of 2 for beer in the FFQ and left *B-beer* (its counterpart of the long questions) missing this mean was imputed.

This technique of imputing mean of each code of FFQ was applied on all the relevant missing variables of the long questions B and C.

Code	B-beer	B-wine	B-sherry	B-spirits
Never	19,483	4,765	13,865	14,405
Less than once a month	5,972	4,805	11,229	7,850
1-3 per month	3,666	4,843	4,527	4,420
Once a week	2,762	4,722	2,684	3,007
2-4 per week	2,044	7,156	1,732	3,080
5-6 per week	602	3,463	356	970
Once per day	324	3,020	564	911
2-3 per day	170	2,213	102	426
4-5 per day	44	175	5	56
6+ per day	11	34	9	22

Table 3.6: Frequencies of responses by code to alcohol FFQ.

In Table 3.7 the numbers of missing values were reduced in all alcohol types, for example for *B-beer* the percentage of *missing* values was reduced from 52% to 0.9%.

These single imputation methods (Section 3.3.1, 3.3.2 and 3.3.3), in which missing values of alcohol consumption in questions B and C were imputed using row borrowing, column borrowing and lastly imputing recorded mean of the cells, were easy to implement. The sequence of imputations was arbitrary, and was based on logical decision. However, the sequence of single imputations, although based on logical assumptions, had the following deficiencies. The completed dataset appeared to have more information than the observed dataset, so the assessment of the precision of the estimator could be incorrect. These methods also underestimate the sampling variance associated with the incompletely recorded variables.

In the next Section, these methods were improved by applying regression imputation instead of mean imputation.

	B		C	
Alcohol	**Recorded**	**Missing**	**Recorded**	**Missing**
Beer or Cider	35,042	325 (0.9 %)	35,198	169 (0.5%)
Wine	35,102	265 (0.7%)	35,355	12 (0.0%)
Sherry	34,972	395 (1.1%)	35,186	181 (0.5%)
Spirits	35,367	0 (0.0%)	28,048	7,319 (0.4%)

Table 3.7: Frequencies and percentages of missing items to question B and C after row borrowing, column borrowing and imputing means to the cells.

3.3.4 Regression imputation

Consumption of alcohol now and five years ago had to be related for the same person. Each identical pair of alcohol (now/five years ago) consumption was studied separately in detail. The strength of association between all available pairs of alcohol was presented in Table 3.8. Pearson correlation coefficients showed that the relations were highly correlated for each similar pair. The association was further tested using the non-parametric Spearman correlation, which ranks the two variables and so does not make any assumptions about the distribution of the variables. The association was even stronger using the later test. Table 3.8 gives evidence that a middle-aged person's consumption of alcohol now and five years ago is bound to be related, as a result of the habits formed over the many years of earlier life.

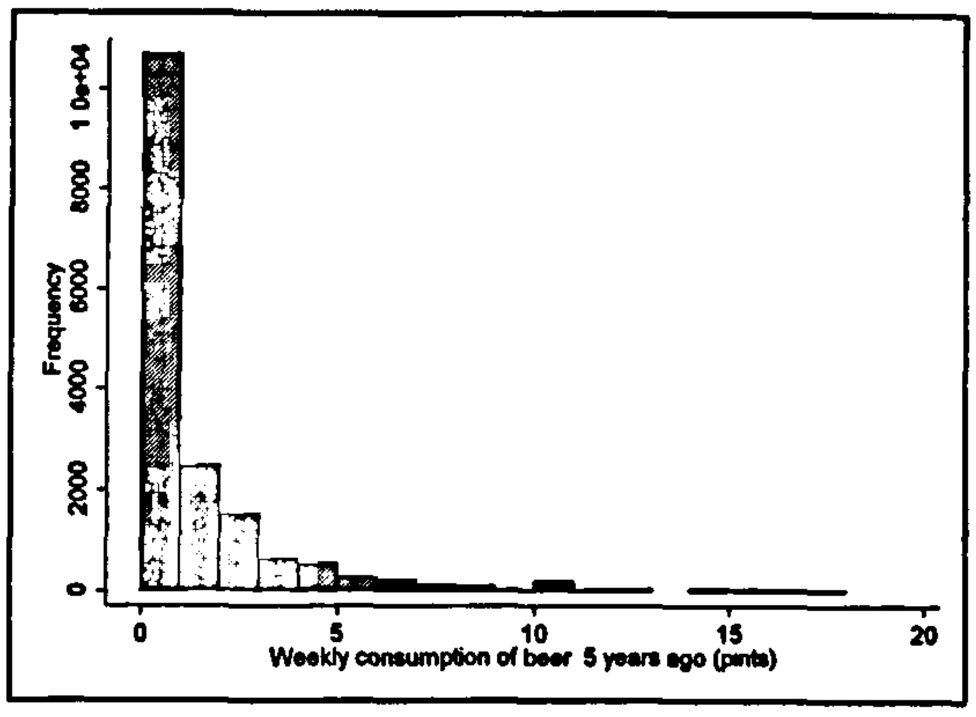
Variables			Pearson Correlation coefficients	Spearman Correlation coefficients
B-beer	__	*C-beer*	0.75	0.88
B-wine	__	*C-wine*	0.81	0.84
B-sherry	__	*C-sherry*	0.77	0.87
B-spirit	__	*C-spirit*	0.71	0.81

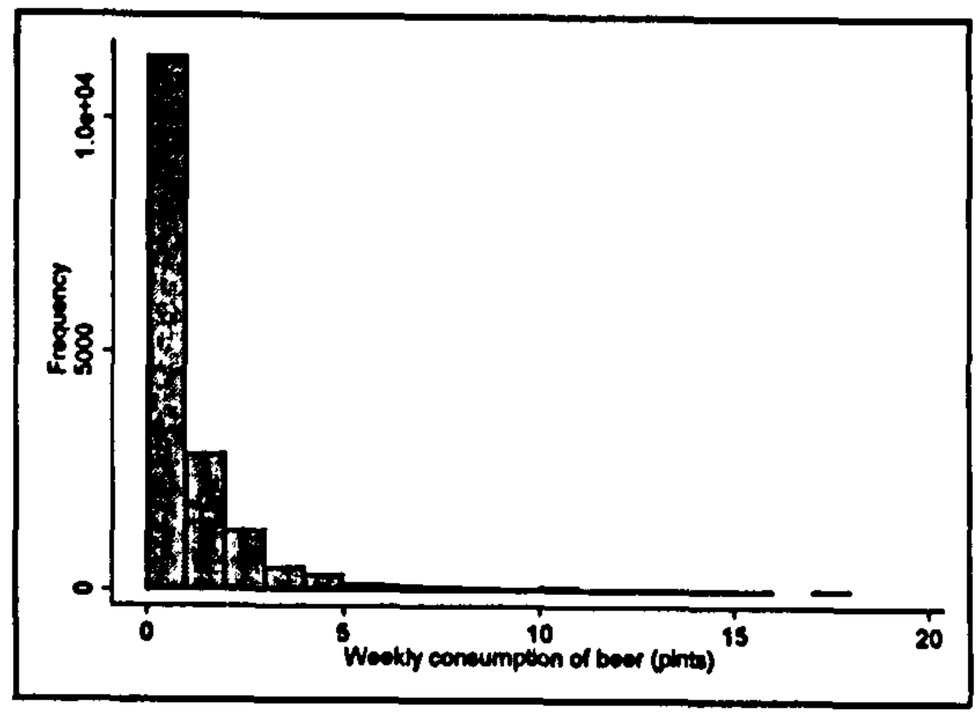
Table 3.8: Correlations between identical pairs of alcohol consumption.

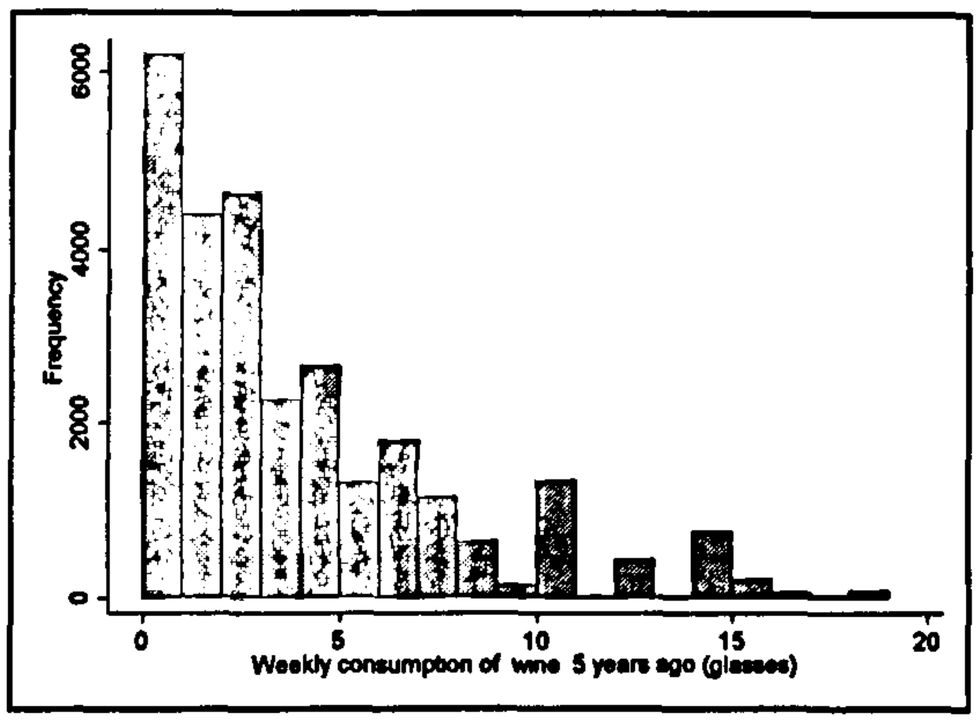
The distribution of each type of alcohol was tested by histograms presented in Figure 3.2. It was very clear that all the variables were highly skewed. No transformation could help to yield normally distributed data, because a substantial proportion of the quantities were equal to zero.

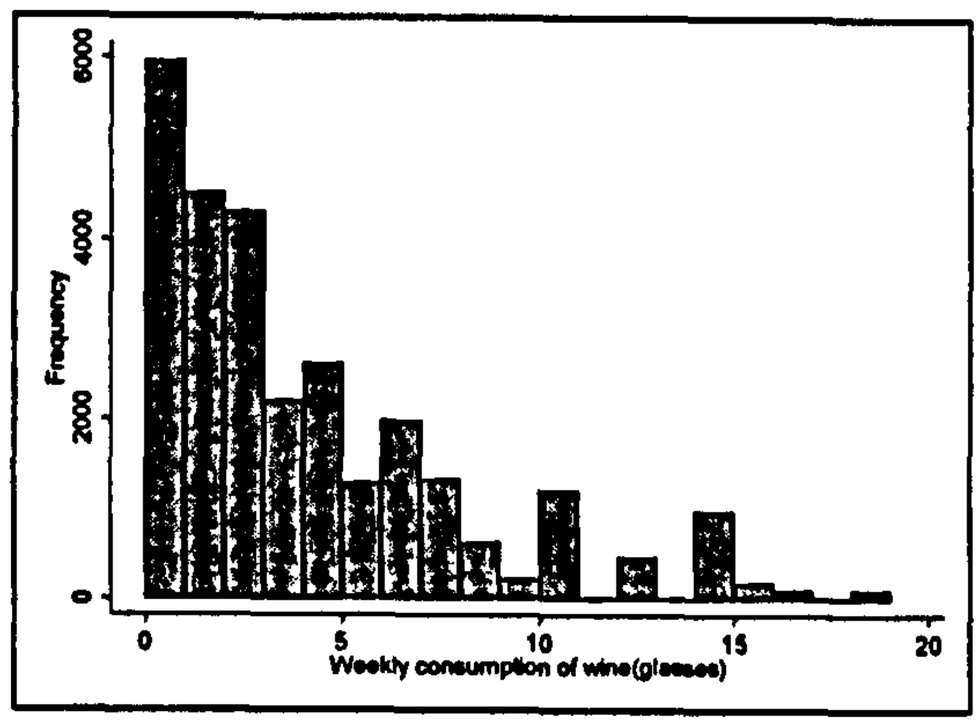
The imputation of missing values proceeded as follows:-

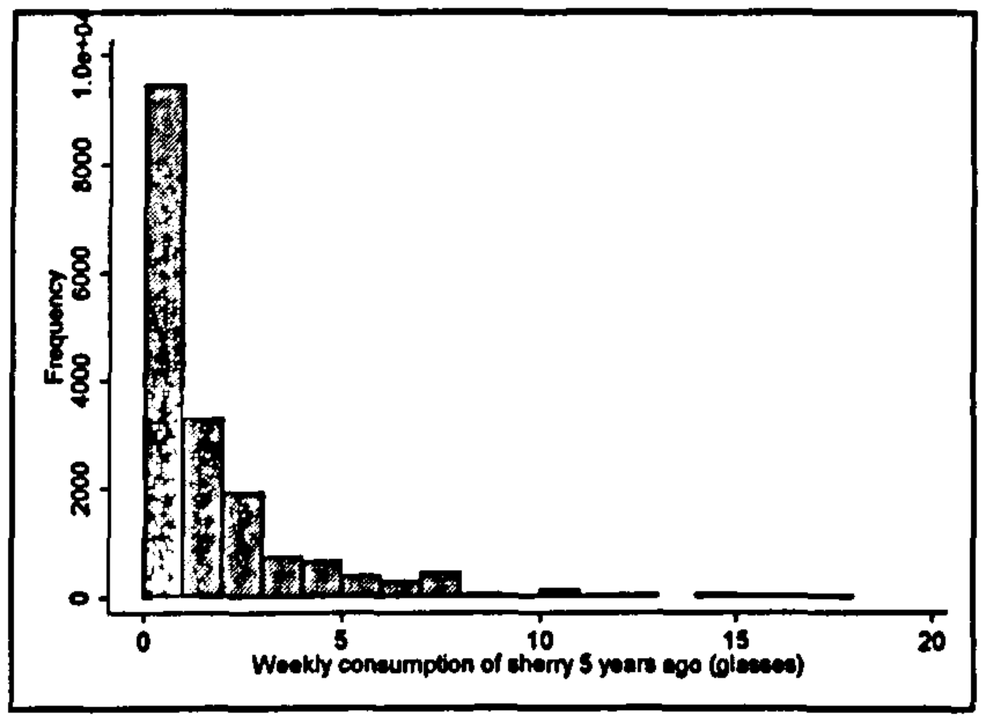
- First, I applied the row borrowing and column borrowing described in Sections 3.3.1 and 3.3.2.
- Second, an improvement to imputing means of the cells was done by fitting a regression model. An ordinary least square model was fitted and missing values were predicted from the regression model, these regression imputations was applied to the four types of alcohol.
- Regression imputation requires the same assumption as the ordinary least square regression; normality. All the alcohol variables were skewed, however normal model was assumed. This variables could have been modelled under a logarithmic transformation and then transformed back after imputation. Following Schafer (1997) these variables were imputed under normality assumptions without transformation. This regression imputation was described in Section 2.7.2.
- For example, for the pair of variables *B-beer and C-beer,* values were imputed for *B-beer,* by regressing B-*beer* on *C-beer* for the cases with complete data, and then the resulting regression equation was used to generate predicted values for the cases that were missing for B-*beer.*
- The roles were next exchanged and values for *C-beer* were imputed by regressing *C-beer* on *B-beer* for the cases with complete data where the imputed values of *B-beer* imputed in the previous step are also accounted for. The resulting regression equation was then used to generate predicted values for the cases that were missing on *C-beer.* Percentages of missing values on the two blocks of alcohol after imputation were presented in Table 3.7.

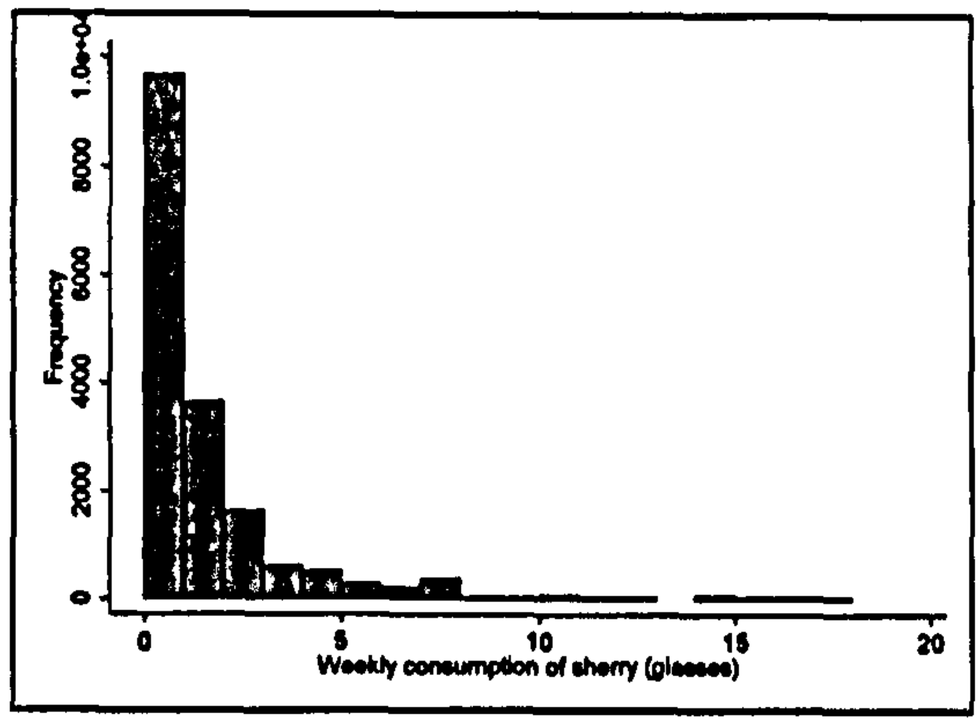


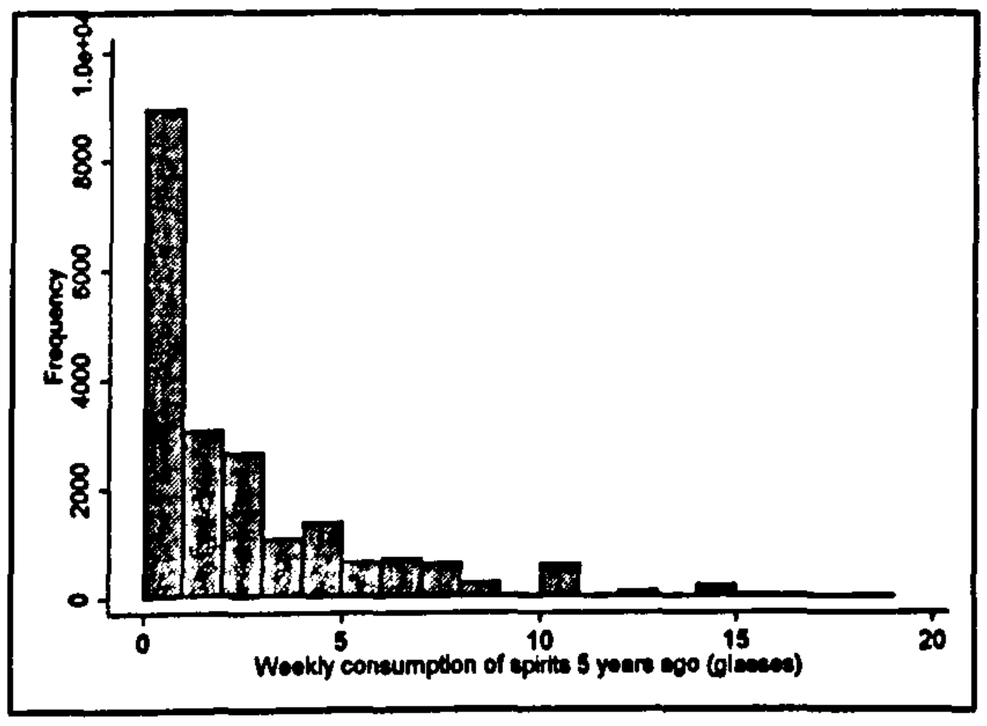












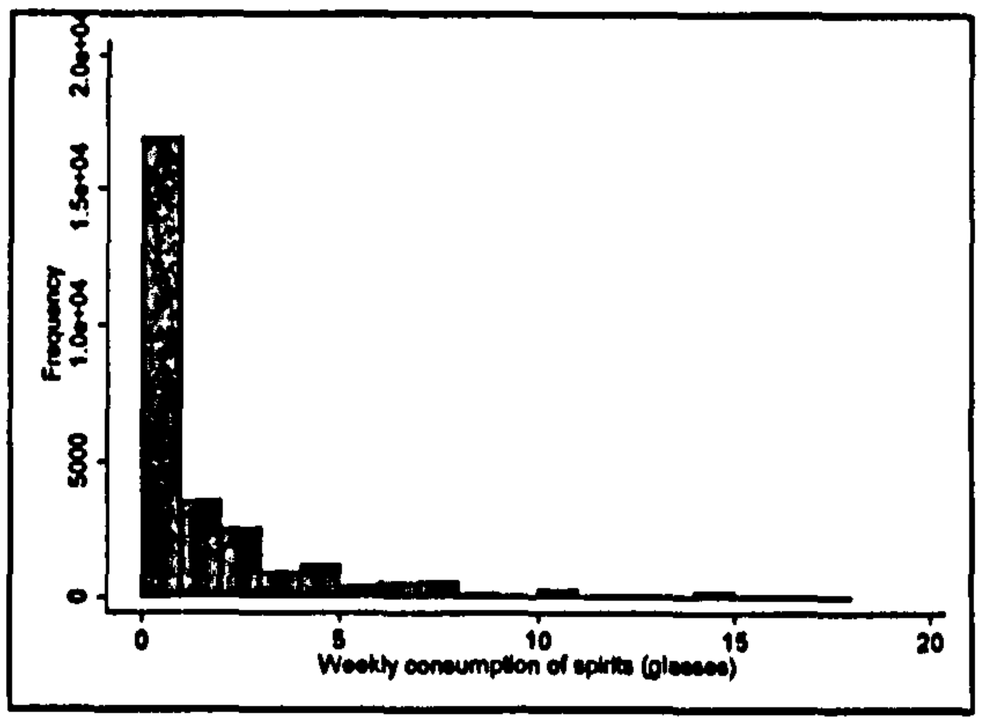
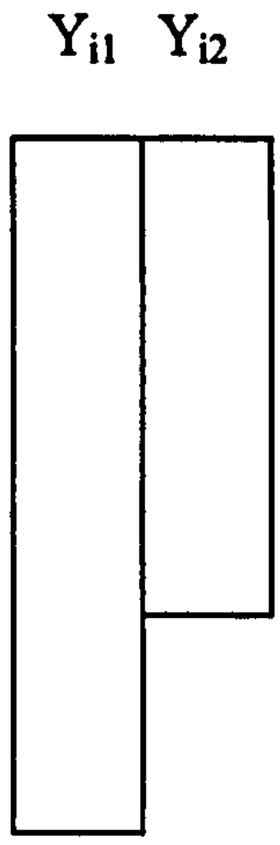


Figure 3.2: Histograms of weekly consumption of Beer, Wine, Sherry and Spirits 5 years ago and now.

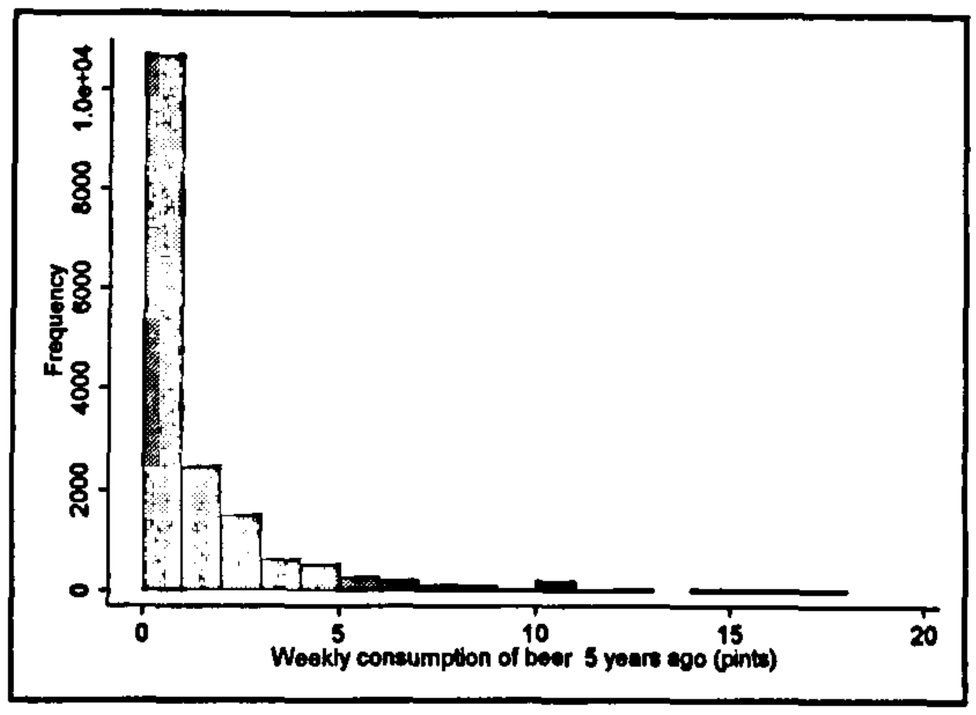


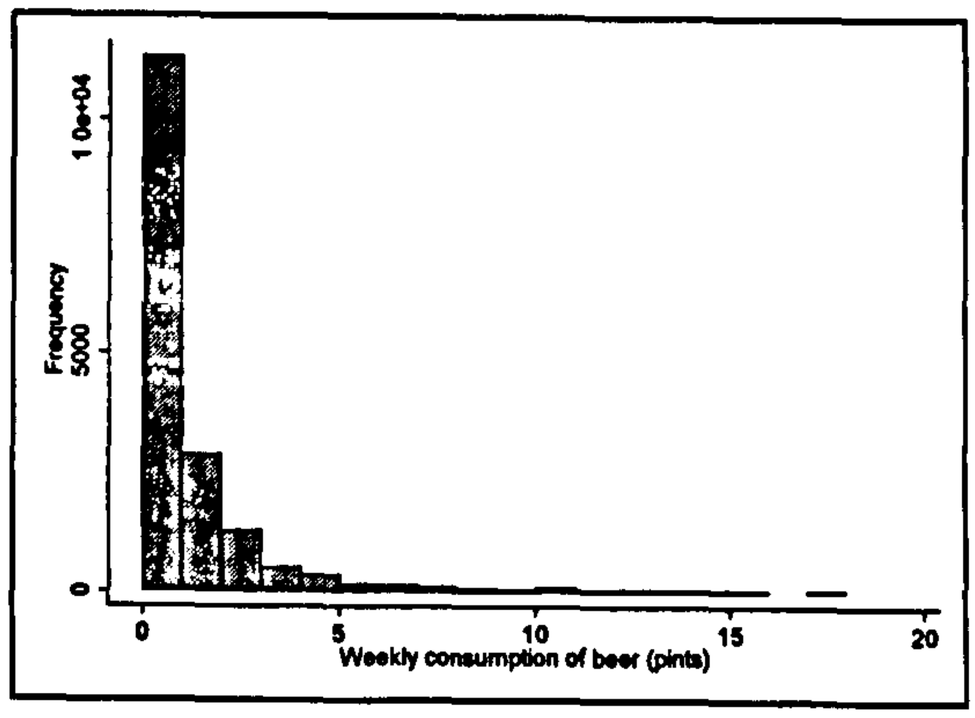
This method was an improvement to imputation of the means because the distributional characteristics of the pairs of consumption variables Y_{i1} and Y_{i2} were maintained (assuming MAR).

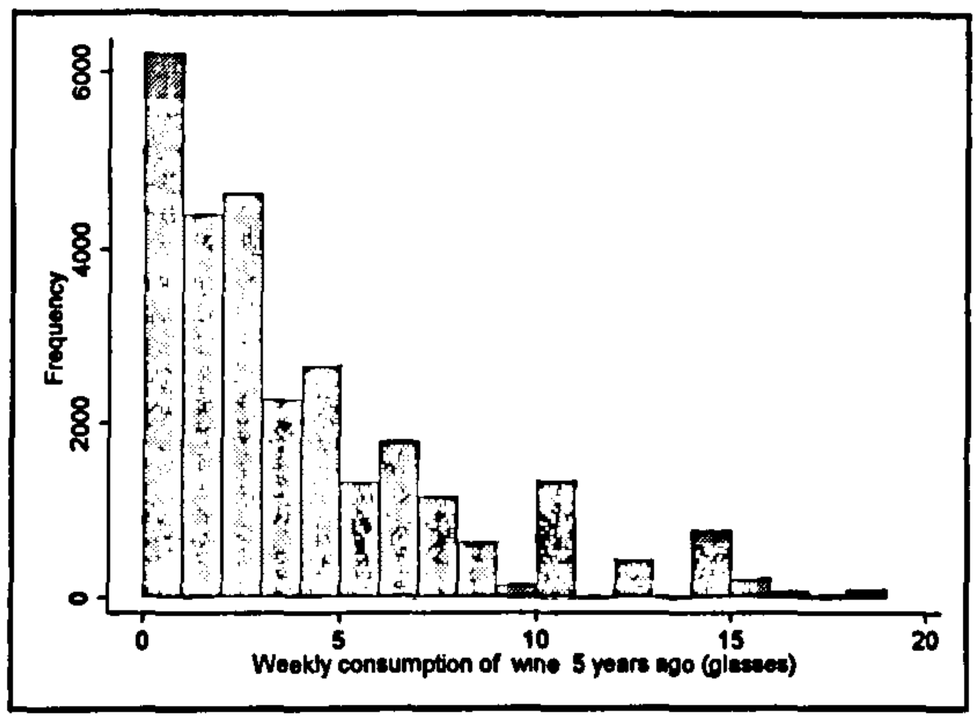
Although these single imputation procedures sound logical and were expected to lead to logical imputations in most of the cases, they still treat missing values as if they were known. The method did not reflect the uncertainty about the prediction of the unknown missing values, and therefore underestimate the variance of the parameter estimates.

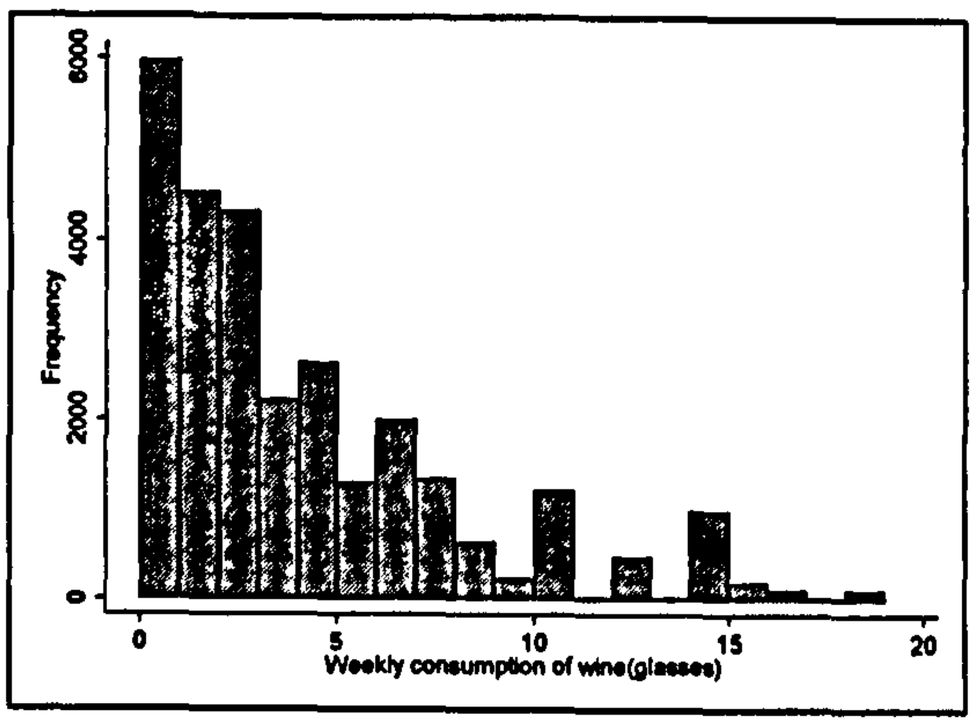
In this Section single imputation methods were used to handle missing data. The aim was to investigate these methods and no specific analysis was applied to check results after imputation. All the applied methods were capable of filling in the missing values and were easy to implement, however all had serious deficiencies.

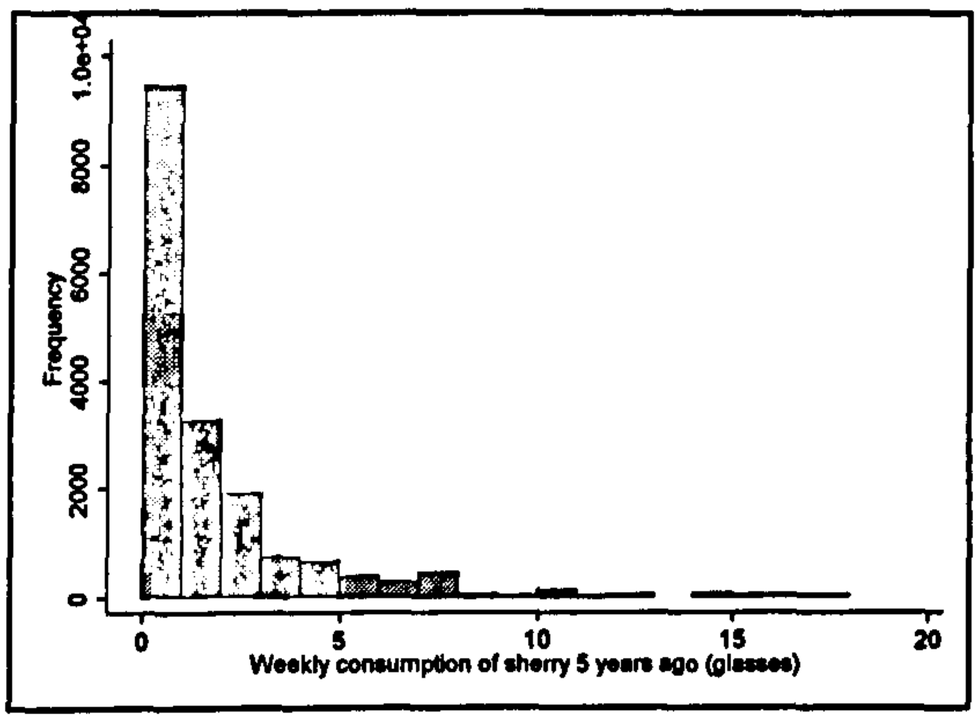
In the next Section single and multiple imputation will be compared and its impact on the amount of alcohol nutrients computed from the alcohol consumption questionnaire will be investigated.

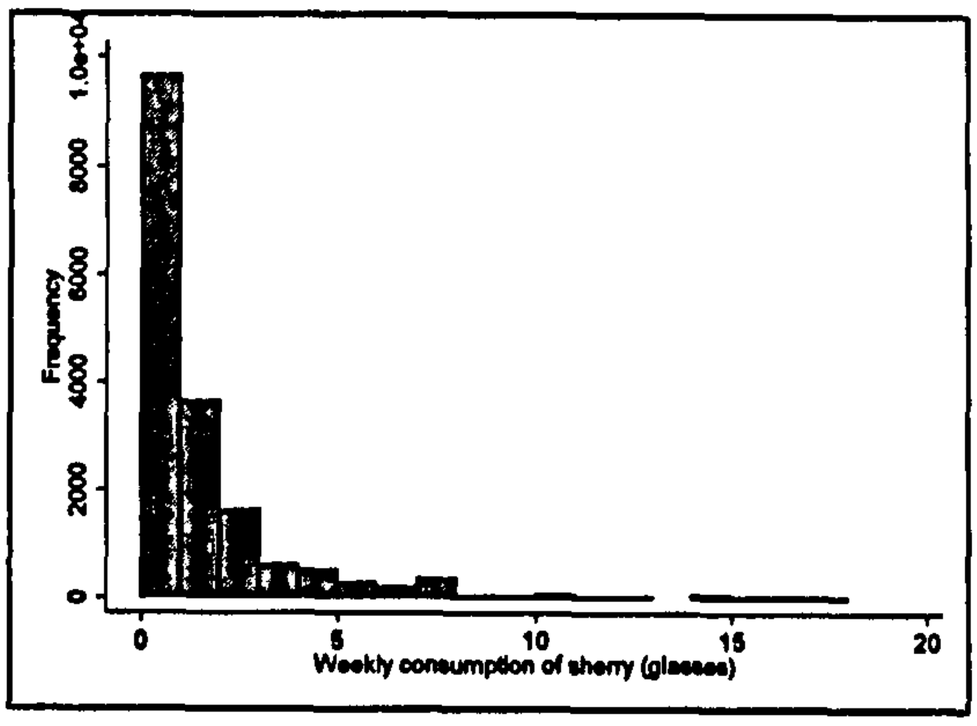


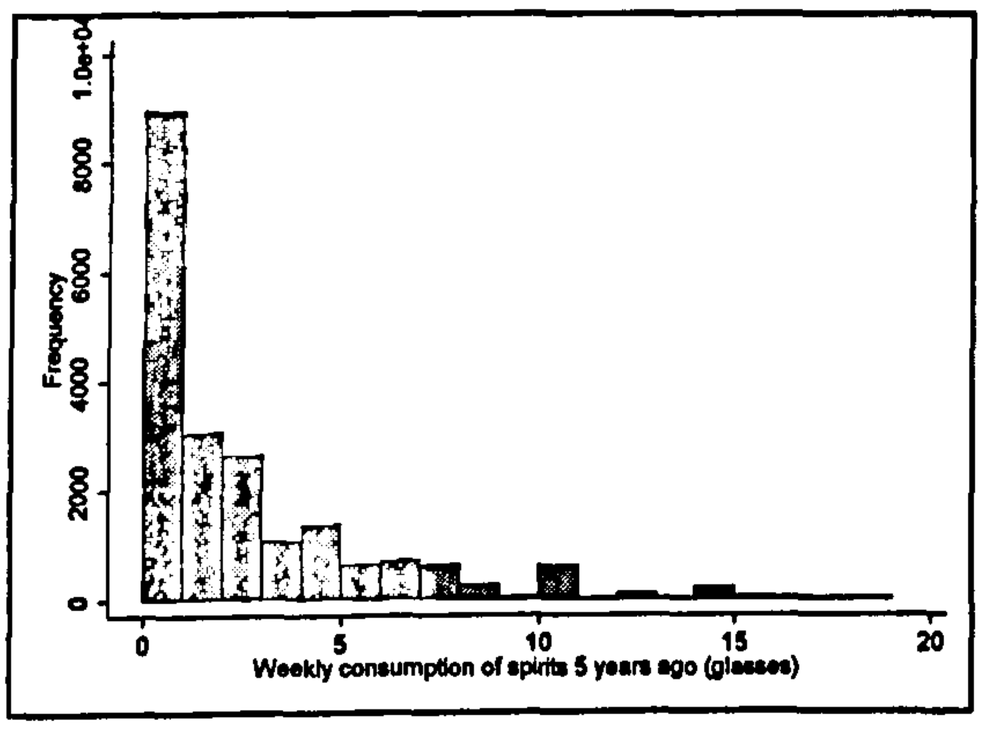












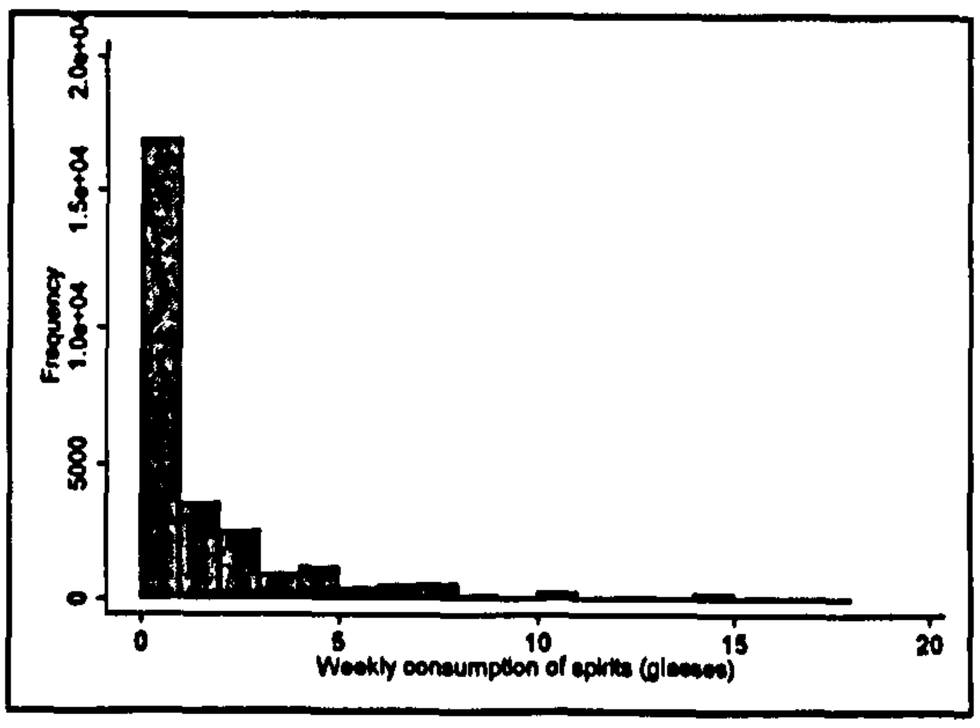
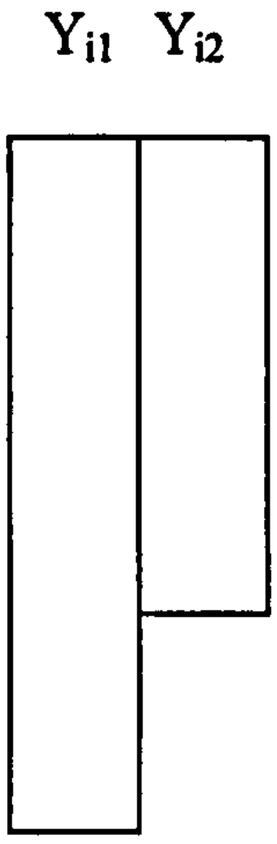


Figure 3.2: Histograms of weekly consumption of Beer, Wine, Sherry and Spirits 5 years ago and now.



This method was an improvement to imputation of the means because the distributional characteristics of the pairs of consumption variables Y_{i1} and Y_{i2} were maintained (assuming MAR).

Although these single imputation procedures sound logical and were expected to lead to logical imputations in most of the cases, they still treat missing values as if they were known. The method did not reflect the uncertainty about the prediction of the unknown missing values, and therefore underestimate the variance of the parameter estimates.

In this Section single imputation methods were used to handle missing data. The aim was to investigate these methods and no specific analysis was applied to check results after imputation. All the applied methods were capable of filling in the missing values and were easy to implement, however all had serious deficiencies.

In the next Section single and multiple imputation will be compared and its impact on the amount of alcohol nutrients computed from the alcohol consumption questionnaire will be investigated.

3.4 Alcohol nutrients

Estimates of overall nutrient intake were found by summing the product of the reported frequency of each food by the amount of nutrient in a specified portion of that food.

In this Section, the effect of missing data on the total nutrient intake for alcohol consumption was illustrated with simple calculations. To find the total alcohol nutrient intake, only the four types of alcohol were considered. However, to evaluate the effect of missing data on carbohydrates, for example, most of the FFQ variables should be considered as most of the food items contained carbohydrates.

To clarify the missing data problem, the effects of different methods of handling missing data on nutrient intake were compared. These methods were applied to the responses of question B that was referred to in Section 3.1.

3.4.1 Complete case analysis

To calculate total nutrient alcohol consumption by the complete-case analysis, it was found that only 12,571 (36%) had complete data, see Table 3.10. Such a large reduction in the data raises two issues. First, having collected less data than anticipated, we have less information than planned. Second, the subjects who fail to respond (to an item or a block of items) may tend to differ (systematically) from subjects whose records are complete. A naïve analysis based on just one third of the data would definitely be biased, as the analysis of the complete

records may yield inferences substantially different from those that would be obtained had no data been missing.

Missing data can therefore lead not only to incorrect p-values but also to incorrect estimates of relationships such as that between diet and cancers.

3.4.2 Imputing a default value

In most nutrition surveys, zeros would be imputed for every missing value. Analysts assume that respondents skip items when they do not consume the specific food concerned; therefore, imputing zeros would be the obvious choice for the missing values. Although logical in many cases, such imputation might not be appropriate for every missing value.

A typical criticism to imputing zeros for missing alcohol items could be *"What if binge drinkers were the ones who didn't report their alcohol consumption ?"*, i.e. what if the respondents who left the alcohol items blank were the binge drinkers who did not want to reveal their alcohol problem, and this could be an example of MNAR, missing data mechanism.

By imputing zeros for all missing items of question B, Table 3.10, all subjects could be included in the analysis. The mean alcohol nutrient intake increased from 7.75g/day in the complete case analysis to 8.60g/day. This increase in mean resulted from the fact that only records with complete four items of alcohol were included in the complete case analysis. For example, if a subject reported consuming beer and wine, and left spirit and sherry as missing, all four items of alcohol would be omitted from the total alcohol nutrient intake. These findings suggest that subjects, who left part of the four items of alcohol consumption

missing, drank more than those who filled in the four items of alcohol. Therefore, when those answers with parts missing were included in the final total, even by imputing zeros for missing values the mean alcohol nutrient intake was found to be higher than the complete case analysis.

3.4.3 Multiple imputation

In this Section multiple imputation was applied to handle the missing data of question B of the alcohol consumption. The method has great advantages over single imputation and the complete case analysis. Multiple imputation helps the researcher to make inferences about the population, rather than the subset of the smaller and possibly different population who had complete records. This method was described fully in Section 2.8.

The most difficult part of the multiple imputation is how to generate the values to be imputed and the choice of the best imputation model. In this Section, the aim was to impute for the four types of alcohol, which were collected in question B. There was the advantage of similar information being collected in question C (alcohol consumption five years ago) and the alcohol consumption of the FFQ.

Multiple imputation for the alcohol nutrient intake was applied by improving the method of regression imputation described in Section 3.3.4.

For a dataset with a monotone pattern of missing data Rubin (1987) described a typical regression imputation model for a variable Y_j with missing values by

$$Y_{ij} = \beta_{i0} + \beta_{i1}X_{i1} + \beta_{i2}X_{i2} + \ldots + \beta_{ij-1}X_{ij-1} + \varepsilon_{ij} \qquad (3.1)$$

where j=1,…,p is equal to the number of variables in the dataset and i=1,…,n is the number of subjects. The residual for the complete cases is then computed by

$$\hat{\varepsilon}_{ij} = \mathrm{Y}_{ij} - (\hat{\beta}_{i0} + \hat{\beta}_{i1}\mathrm{X}_{i1} + \hat{\beta}_{i2}\mathrm{X}_{i2} + \ldots + \hat{\beta}_{ij-1}\mathrm{X}_{ij-1}) \qquad (3.2)$$

for each missing value of Y_{ij} the fitted value are then computed

$$\hat{\mathrm{Y}}_{ij} = \hat{\beta}_{i0} + \hat{\beta}_{i1}X_{i1} + \hat{\beta}_{i2}X_{i2} + \ldots + \hat{\beta}_{ij-1}X_{ij-1} \qquad (3.3)$$

In the last step impute

$$\tilde{\mathrm{Y}}_{ij} = \hat{\mathrm{Y}}_{ij} + \varepsilon_{ij}^{*}$$

where ε_{ij}^{*} is the residual selected at random from a normally distributed random variables with mean 0 and variance $\hat{s}_{ij}^{\,2}$, where $\hat{s}_{ij}^{\,2}$ is an estimate of the variance of $(\hat{\varepsilon}_{ij})$.

An application of this ideal model in the multiple imputation for alcohol nutrient intake was not feasible for two reasons.

1- The large number of variables to condition on would make the model too complicated.
2- The dataset does not have a monotone pattern, because all X's or predictor variables are incomplete.

To get around these difficulties, imputed values were generated for each alcohol type by regression imputation conditioned on its counterpart from the other questions, with the assumption of normality and MAR mechanism of missing data. Correlation coefficients of each alcohol type in question B and its counterparts of question C and the FFQ were presented in Table 3.9. These correlations were from all available data. The table presented strong associations

between alcohol types. If each alcohol type was to be imputed by conditioning on its counterparts, the main setback would be that the predictors in the imputation models were also not complete; therefore not all the missing values could be recovered. However, recovery of missing data by this method had the following advantages: -

1- A large amount of observed information was used when compared to the complete case analysis
2- The generation of imputed values was based on real information from the dataset.
3- Random variation was taken into consideration when compared to single imputation by a default value.

The multiply imputed datasets were then analysed by using standard methods for complete data, and then the results were combined by Rubin's rules described in Section 2.8 and equations (2.1) – (2.4).

Correlation coefficients	Now	Five years ago	FFQ
Beer			
Now	1.00		
Five years ago	0.61	1.00	
FFQ	0.76	0.55	1.00
Wine			
Now	1.00		
Five years ago	0.80	1.00	
FFQ	0.78	0.65	1.00
Sherry			
Now	1.00		
Five years ago	0.77	1.00	
FFQ	0.73	0.61	1.00
Spirits			
Now	1.00		
Five years ago	0.71	1.00	
FFQ	0.75	0.57	1.00

Table 3.9: Correlation coefficients (all available data), of each alcohol type of question B and its counterparts from question C and the FFQ

The analysis using multiple imputation was based on information from 35,055 records, compared to 12,571 records in the complete-case analysis. The bias of underestimating the variance, which results from deterministic imputation was corrected for by the addition of the random variability and then combining the results. That is, the total variance was made up of two components: -

- The natural variability among the five imputed datasets, known as the within imputation variance. This part of the variance is similar to the variance we would produce if we did not account for the missing data, and is simply found by averaging the variance estimates from each imputed dataset.

- The second part of the variance known as the between imputation variance, which measures the amount of variability of estimates from dataset to dataset. In other words, if estimates vary greatly then this amount is expected to be large and uncertainty due to imputation is high.

The first obvious difference among the three methods of handling missing data was the large variation in mean alcohol intake, which varied greatly by the three methods, see Table 3.10. The larger mean nutrient intake from both the multiple imputation (11.30g/day) and zero imputation (8.60 g/day) compared to the complete case analysis (7.75 g/day), resulted from the inclusion of more subjects, suggesting that the excluded subjects with missingness, drank more than those who completed all the questions. This was even true when zeros were imputed — although the lowest possible value was imputed, the release of new information from subjects previously excluded, still led to an increase in estimated mean intake. The standard errors decreased slightly due to the larger sample size included.

Multiple imputation in this particular application had its own deficiencies. The first and most serious was that imputation models were generated conditioning each type of alcohol on its counterpart. Although correlation coefficients showed strong association between each type of alcohol, it is important to ensure that the imputation model is more general i.e. includes at least as many covariates as the analysis model.

In the next Section generation of multiple imputation will be improved by the use of a Markov Chain Monte Carlo method.

Alcohol g/day	Complete Case Analysis			Imputing Zero			Multiple Imputation			Multiple Imputation by MCMC		
	Obs.	Mean	S.E.	Obs.	Mean	S.E.	Obs.	Mean	S.E.	Obs.	Mean	S.E.
Wine	28,937	6.72	0.047	35,367	5.50	0.041	35,298	5.95	0.041	34,061	6.92	0.049
Beer	16,877	3.11	0.055	35,367	1.48	0.028	35,175	2.93	0.030	34,061	4.02	0.057
Spirits	27,620	1.45	0.019	35,367	1.13	0.015	35,236	1.58	0.016	34,061	1.53	0.020
Sherry	17,122	1.02	0.015	35,367	0.49	0.008	35,169	0.88	0.008	34,061	1.25	0.016
Total alcohol Intake g/week	12,571	7.75	0.098	35,367	8.60	0.056	35,055	11.30	0.061	34,061	13.74	0.083

Table 3.10: The impact of handling missing data by the complete case analysis, imputing zeros and multiple imputation, on alcohol nutrient intake.

3.4.4 Multiple imputation by MCMC

The multilevel modeling software *MLwiN* introduced multiple imputation by Markov Chain Monte Carlo methods in its latest version. *MLwiN* uses Gibbs sampling in a Bayesian framework for the generation of imputed values. The Bayesian approach to multiple imputation and Gibbs sampling was described in Section 2.10.

MLwiN generates initial values for the missing data by the convergence of a method known as the Iterative Generalized Least Squares (IGLS), which is equivalent to finding the maximum likelihood for the unknown parameters. IGLS finds point estimates through an iterative procedure, which involves iterating between two deterministic steps. Convergence is achieved when two consecutive estimates for a specific parameter are close together. *MLwiN* adapts this method, which was originally designed for hierarchical models to fit all models.

Imputed values for missing data of the four types of alcohol consumption were generated using the above-mentioned method and software. Initial values of the means of the four types of alcohol were generated before the start of the MCMC, by IGLS to ensure good starting values (maximum likelihood estimates). The MCMC next generated the first imputed dataset by the Gibbs sampling method through 1000 iterations after a "burn-in" of 500 iterations. As the chain started at a particular starting point, it would actually take time for the chain to settle down and sample from the posterior distribution. The "burn-in" is the initial iterations in which the chain settles down (converge) that are discarded and not used to describe the final parameter distribution, i.e. they are discarded and are used only to initialize the Markov chain; see Section 2.10.1 for a description of the sequence of this method. As the chain started

with maximum likelihood estimates the burn-in of 500 iterations was considered adequate for the chain to settle in. Further four datasets were then generated through 2000, 3000, 4000 and 5000 iterations. Estimates and standard errors of the five imputed datasets are presented in Figure 3.3. The Figure shows that after 1000 iterations the chain has converged to its equilibrium distribution.

Table of imputed values generated by *MLwiN* are presented in Figure 3.4. The five columns labelled dataset 1 – dataset 5, represent the five completed datasets. It is clear from the figure that imputed values included a lot of negative numbers; this was the result of treating variables as continuously normal distributed.

Minimum and maximum of the four types of alcohol as well as the total alcohol nutrient intake, in the observed and completed datasets are presented in Table 3.11. The minimum values in the completed dataset were as low as –32g of alcohol nutrient per week. The mean alcohol nutrients in th four types of alcohol was almost double the mean alcohol nutrient from the observed datasets.

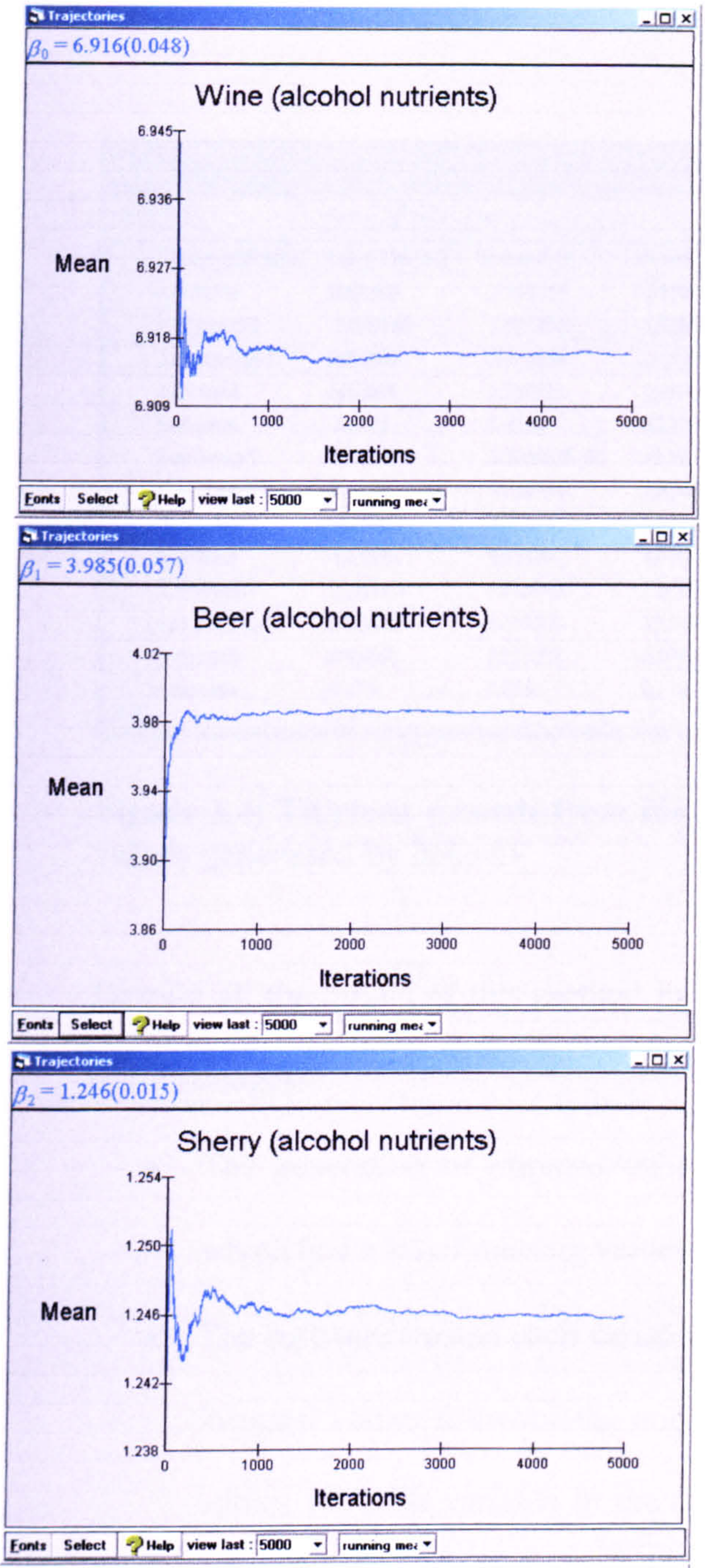


Figure 3.3:Estimates(standard errors) and iterations of alcohol nutients in the five completed variables Wine, Beer, Sheery and Spirits

Data

goto line 1 view Help Font

	resp_indicator(	resp(136244)	dataset 1(	dataset 2(	dataset 3(	dataset 4(	dataset 5(
1	alcwine	MISSING	-.7394316	-.8239453	15.84055	.4152407	-5.336477
2	alcbeerp1	1.820154	1.820154	1.820154	1.820154	1.820154	1.820154
3	alcsherryp1	MISSING	-.5968936	-.1163952	2.294325	-.2650845	-.5891918
4	alcspirit	MISSING	3.887722	-3.898413	4.537008	-1.857958	8.441582
5	alcwine	3.2375	3.2375	3.2375	3.2375	3.2375	3.2375
6	alcbeerp1	MISSING	3.329463E-02	17.18288	5.287097	2.52176	.9989982
7	alcsherryp1	MISSING	1.783701	4.031929	3.351338	1.567732	.665386
8	alcspirit	MISSING	-.2338743	6.807915	3.996006	.0285239	1.374399
9	alcwine	1.61875	1.61875	1.61875	1.61875	1.61875	1.61875
10	alcbeerp1	7.280616	7.280616	7.280616	7.280616	7.280616	7.280616
11	alcsherryp1	MISSING	2.608336	3.594672	.8484795	.2884406	-1.206147
12	alcspirit	MISSING	1.047958	-2.696742	-.5377663	-.1870835	-4.701867
13	alcwine	6.475	6.475	6.475	6.475	6.475	6.475

Figure 3.4: Thirteen records from the observed data and its relevant five imputed values generated by *MLwiN*

Despite all the merits of this method in filling the missing values it had the following disadvantages:-

- The generation of imputed values was slow especially in variables like beer, which had a lot of missing values.
- The software treated each variable as normally distributed, which meant that the imputed values followed the normal distribution rather than the observed set of values for each variable in the dataset, this lead to the imputation of negative values.
- *MLwiN* is not capable of imputing binary or categorical variables, which is the case of the majority of the variables in the UKWCS.

Having negative values for alcohol consumption is not acceptable, as the information on total alcohol intake per subject is vital to predict the association between alcohol consumption and illness for example cancer, among the women taking part in the

cohort. Transformation for example using log transformation, of the continuous variables before imputation can help in avoiding the generation of negative imputed values, the imputed values can be transformed back after imputation.

MLwiN at this stage is not recommended for the imputation of missing data in a complex datasets such as the UKWCS. An improvement in the software is still needed for it to cope with the imputation of categorical and binary data.

Alcohol g/day	Observed	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Wine						
N	28,937	34,061	34,061	34,061	34,061	34,061
Mean	6.72	6.96	6.88	6.92	6.93	6.92
Min	0	-20.75	-23.98	-26.07	-21.05	-23.63
Max	113.31	113.31	113.31	113.31	113.31	113.31
Beer						
N	16,877	34,061	34,061	34,061	34,061	34,061
Mean	3.11	4.02	3.98	4.05	3.98	3.98
Min	0	-21.86	-23.87	-25.75	-22.51	-27.03
Max	182.02	182.02	182.02	182.02	182.02	182.02
Sherry						
N	17,122	34,061	34,061	34,061	34,061	34,061
Mean	1.02	1.23	1.24	1.25	1.23	1.24
Min	0	-6.77	-6.42	-6.19	-6.19	-6.46
Max	46.62	46.62	46.62	46.62	46.62	46.62
Spirit						
N	27,620	34,061	34,061	34,061	34,061	34,061
Mean	1.45	1.53	1.52	1.51	1.53	1.52
Min	0	-10.07	-8.85	-11.04	-11.88	-9.31
Max	71.45	71.45	71.45	71.45	71.45	71.45
Total alcohol intake						
N	12,571	34,061	34,061	34,061	34,061	34,061
Mean	7.75	13.73	13.66	13.73	13.67	1 3.66
Min	0	-31.83	-32.17	-32.17	-32.42	-32.18
Max	233.45	233.45	233.45	233.45	233.45	233.45

Table 3.11: Alcohol nutrients in the four types of alcohol and total alcohol nutrients in g/day, in the observed and the five completed datasets

3.5 Discussion

The questions on alcohol consumption being in two parts of the questionnaire, and in two different formats, helped to understand how the format of the question had great impact on subjects' responses and how different methods of handling missing data had great effect on results.

The response rate for FFQ questions of alcohol consumption was quite satisfactory. Imputing zeros led to reasonable results, and no complicated imputation was needed for the missing values. A motivation which supported the imputation of zeros in the FFQ question of alcohol, is that it was by far the most frequent category among the respondents to the alcohol consumption question in FFQ format, and the assumption of subjects skipping this question to reveal that they don't consume that type of alcohol was found realistic. This was also supported in my earlier discussion that when the amount of missing information was 5% or less, the method used for handling missing data did not have great effect on the final results. With this small percentage of missing values even analysing only the complete cases had no great effect on the overall final results.

The response rate for the crosscheck questions of alcohol consumption (questions B and C) was poor. It was clear that the women who took part in the study were happy to tick the questions, which were in FFQ format, but did not want to report or were not able to report their actual alcohol consumption in number of pints of beer and glasses of wine, sherry and spirits.

This chapter discussed the impact of different methods of handling missing data on questions B and C as well as on alcohol nutrients computed from these questions.

Single imputation methods applied to questions B and C, which were named as column borrowing, row borrowing and imputation of the means, recovered part of the information that would have been lost by the complete case analysis; however all the methods underestimated the variability of missing data, not all the records with missing values could be imputed.

Comparison of the three methods of handling missing data, in the total alcohol nutrient intake (crosscheck question), showed that ignoring missingness by analysing only complete cases underestimated the mean. A lot of information contained in the non-empty records was discarded. Imputing extreme values, in this case zero, also led to biased results. This would pretend that one knew exactly what the missing values were if they had been observed, but also incorrectly increased the apparent precision of estimation (i.e., inappropriately small estimated standard errors).

The two applied methods of multiple imputation had the following strengths: -

- Information on most of the incomplete records was used.
- The algorithm intended for analysing the complete data was applied, although several times, without any alterations.

Multiple imputation was better even if it was imperfectly executed. Multiple imputation by the MCMC using *MLwiN* converged well, the software was not straightforward to use but had great advantages of good graphical presentation. However, it had the disadvantage of imputing negative values. The software was also not capable of dealing with categorical and binary variables. This serious disadvantage made it not a favourable solution for handling missing data in the UKWCS. Nonetheless, future modification of the software and the possibility of it imputing for binary and categorical data would be of great help.

The imperfection of the first applied method of multiple imputation, was on the imputation models used for the generation of plausible values. This values were generated by conditioning each type of alcohol on its counterpart, although there was no evidence against the MAR assumption, however the inclusion of as many predictors as possible tends to make this assumption more plausible. This method also suffered from the fact that predictor variables in imputation models were also incomplete, and this led to the disadvantage that not all of the incomplete records could be retrieved. Nevertheless, by multiple imputation the uncertainty about the missing data was still taken into account. This method will be further modified in Chapter four to account for some of the mentioned disadvantages.

Chapter 4

Imputation of missing predictors in a regression model: a comparison of methods

4.1 Introduction

The consumption of fruit and vegetables has been found to be beneficial for general health, and great interest in recent years has focused on the possible protective effect of eating fruit and vegetables against certain types of cancer (Block *et al.*, 1992; Gandini *et al.*, 2000). WHO (1990) recommends that the total intake of fruit and vegetable of at least 400g per person per day provide an effective protection level against diseases. This protective level was put in a simpler form of five portions of fruit and vegetables per person per day (COMA, 1994).

Pollard *et al.* (2001) investigated the health and lifestyle factors that affect fruit and vegetable consumption in the UKWCS. A logistic regression model was fitted to predict the association between high and low levels of fruit and vegetables consumption with a number of lifestyle factors. Among these factors were 'physical exercise', 'alcohol consumption', 'smoking', 'having children under 16 years of age', being 'vegetarian or vegan' and whether or not the women take any 'vitamin' supplements. Seven records from the UKWCS were excluded from the analyses as they were found to have extreme

outliers in one or two variables; the final sample size analysed was 35,367 records. As listwise deletion was applied in the multi-variable logistic regression analyses, the sample size was reduced to 10,313 instead of 35,367 of the women within the cohort.

The aim of this chapter is to improve the standard analysis by exploiting the information in the incomplete records. The limitations and bias that can result from analysing only complete records has already been explored in Section 1.1. The impact of missing data and the difference between results on complete and imputed datasets was also compared in the alcohol nutrient intake in Chapter 3. In this chapter methods of handling missing data were applied to a different type of analysis, to understand that missing data does not just affect estimates like mean alcohol intake, but also odds ratios in logistic regression, and the precision of these estimates. By ignoring this information the inferences would be exposed to possible bias due to poor representation of the population, which one assumes to be represented well by the sample of 35,367 subjects. Two methods of dealing with complete records (multiple imputation conditioning on one variable, as well as multiple imputation conditioning on all covariates in the model by chained equations) were applied. Differences in the results following these three methods to handle incomplete records were then compared. These investigations would help to give a better insight on how to handle the problem of missing data, and whether more complex methods are worth the programming effort or if simpler methods could handle the problem adequately.

4.2 The variables

The number of portions of fruit and vegetables consumed per day was calculated from the FFQ data. Pollard *et al.* (2001) showed that the daily figure included fruit and vegetables from composite dishes and allowed for one portion of fruit juice per day but excluded potatoes. Standard portion sizes were employed for calculation of amounts consumed using Helen Crawley's book of food portion sizes, Crawley (1994). The subjects were then divided into tertiles (referred to as T1, T2 and T3) according to their fruit and vegetable consumption. *T1* comprised those with the lowest intake and T3 represented the consumers with the highest intake. The binary outcome variable used in the logistic regression compared the highest tertile T3 with T1 excluding the middle tertile T2.

The aim of this chapter is not to comment on the choice of outcome predictor variables. Instead I will investigate the effect of missing data and compare a number of approaches to handle it using these analyses to illustrate it.

All the covariates included in the logistic regression model were derived from the long *questions of Section 2 in the questionnaire; these variables were as follows:-*

a- Categorical variables
   - Vitamin intake (two categories)
   - Alcohol consumption (four categories)
   - Smoking (four categories)
   - Socio-economic group (ten categories)
   - Marital status (five categories)
   - Education level (four categories)
   - Vegetarian/Vegan status (two categories)
   - Employment status (five categories)
   - Region (thirteen categories)
   - Women with children under 16 years (two categories)

- Illness (nine categories)

b- Continuous variables

- Age (measured to the nearest year at the time of completing the questionnaire)
- 'Physical exercise' in hours per day (this variable was computed as the overall time spent on housework, do-it-yourself jobs, walking, cycling, and other 'physical exercise' such as keep fit, aerobics, jogging, tennis and swimming)

Some of the variables were originally reported in the questionnaire; these were illness, alcohol consumption, smoking, education level, employment status and socio-economic group. The remaining variables were computed from one or more variables in the questionnaire. For example the vegetarian/vegan variable was constructed from a combined response to two questions of the questionnaire.

5a- would you describe yourself as a vegetarian? Yes ☐ No ☐

5b- would you describe yourself as a vegan? Yes ☐ No ☐

Do you describe yourself as vegetarian?	**Do you describe yourself as vegan?**			
	Yes	**No**	**Missing**	**Total**
Yes	233	8,686	672	9,591
No	122	23,578	929	24,629
Missing	77	266	804	1,147
Total	432	32,530	2405	35,367

Table 4.1: Cross-tabulation of the responses to being vegetarian or vegan

By definition, a vegetarian is a person who does not eat meat or fish, but may eat other animal products such as eggs, milk, or cheese. On the other hand, a vegan is defined as a person who does not eat any animal products at all, including meat, fish, seafood and dairy products. The difference between a vegan and a vegetarian is sometimes not clear in the public's perception. The cross-tabulation in Table 4.1 shows that 122 respondents claimed to be vegans but not vegetarians and 8,686 claimed to be vegetarians but not vegans. Although the latter situation could be true, the former is not realistic. It is likely that this question was interpreted differently by different people with some considering vegans as a subset of vegetarians, and others considering them a separate group.

A binary variable was constructed by combining the responses to 5a and 5b. The 23,578 subjects who claimed to be neither vegetarians nor vegans were classified as non-vegetarians and non-vegans. All respondents who filled in "yes" for being vegetarian, or "yes" for being vegan, were classified as vegetarians and vegans. Those who filled in "no" for being vegetarian and left vegan missing were considered as neither vegetarian nor vegans, the same was done for those who filled in "no" for vegans and left vegetarian missing. The new generated variable, had 804 (2.3%) missing values, see Table 4.2.

The history of having any type of illness was recorded from a question, which read: -

Have you been told by a doctor that you have or had any of the listed illnesses?

Yes ☐

No ☐

The aim of the inclusion of this variable in the model was to assess the history of getting any of the listed illnesses, and how that can influence the consumption of fruit and

vegetables. In general, people tend to remember major illnesses like cancer, diabetes, high blood pressure in the past if they had any, and would skip the question if they didn't. For this reason, the variable was completed, by filling in zeros for every missing value.

The dataset also included a variable on whether or not the subject had children less than 16 years of age. This variable was generated from the question

Have you had any children? Yes ☐ No ☐

Respondents were then asked about the sex and date of birth of the children. The reported date of birth for each reported child was then computed to find out if the child was younger than 16 years of age.

4.3 Missing data in the covariates

The extent of missingness in the thirteen covariates included in the logistic regression model was very uneven; this was summarized in Table 4.2. The percentage of missing values ranged from around 44% for 'physical exercise' to 1.2% cases for age. The variable 'job' had no missing values as the result of having an additional category 'other'. Missing values in this variable were filled in as 'other' for this variable. This could be an alternative way of dealing with missing data in categorical explanatory variables, by having missing as a separate category with its own dummy variable. The 'region' variable, was computed from postcodes of the address of the respondents, this was almost complete with only 3 values missing, see Table 4.2.

Zeros were imputed for missing values in the variable on whether or not the subject had children less than 16 years of age. It was assumed that those who did not have children left the question blank. For the remaining variables in the logistic regression model there were no default values that could be substituted for the missing items.

The 'physical exercise' variable had the highest percentage of missing values. There was no obvious explanation for these missing values, as subjects must have been practicing if not all of the listed exercises then at least one or two of them.

Variable	Observations	Missing (%)
Vitamin	32,113	3,254 (9.2%)
Alcohol	34,563	804 (2.3%)
Smoking	34,315	1,052 (3.0%)
Physical exercise	19,844	15,523 (43.9%)
Socio-economic group	34,733	634 (1.8%)
Marital status	34,813	554 (1.6%)
Age	34,940	427 (1.2%)
Education level	32,315	3,052 (8.6%)
Vegetarian /vegan status	34,563	804 (2.3%)
Job	35,367	0 (0.0%)
Region	35,364	3 (0.0%)
Women with children under 16 years	35,367	0 (0.0%)
Illness	35,367	0 (0.0%)

Table 4.2: Numbers and percentages of missing values in variables included in the logistic regression model.

In the following Section the logistic regression model was fitted with the available information or incomplete data. This reflects the actual analysis presented in Pollard *et al.* (2001).

4.4 Complete case analysis

Pollard *et al.* (2001), compared the highest consumers of fruit and vegetables (T3) with the lowest consumers (T1), adjusting for the covariates. Having a binary outcome a multiple logistic regression model was used to determine predictors of high fruit and vegetable consumption. Only cases with complete information were included in the analysis, see Table 4.9 for the results. The sample size was reduced substantially for two reasons: -

- Most of the covariates included in the model, had missing values, these ranged from 1.2% for age to more than 44% for 'physical exercise', see Table 4.2.
- Exclusion of subjects who were in the mid-tertile of fruit and vegetable consumers.

The logistic regression model included only about ten thousand subjects, instead of about 23,000. Most analysts choose complete case analysis (listwise deletion) for its simplicity, and also since this is the default method incorporated in most statistical packages.

The loss of over a half of the subjects results in inefficient inferences. The 13,000 records that were not used in the analysis contain a lot of information; their rejection seemed an unnecessary waste of useful information. Most of the records were not completely empty. In fact, many had only one item missing.

Complete case analysis requires the strong assumption of MCAR, see Section 2.2.1, to generate consistent estimates. No attempt was made in Pollard *et al.* (2001) to investigate the relation between missing and observed values to the outcome variable.

4.5 Single variable analyses

A dummy category for missing values was constructed in the categorical variables; single variable logistic regression models on all available data were then fitted to compare difference in associations between observed and missing values to the outcome variables, see Table 4.3.

The single variable logistic regression with complete-record and incomplete-record strata within each of the categorical variables, showed that missing vitamin, alcohol, vegetarian/vegan, smoking, highest educational level all significantly predicted higher consumption of fruit and vegetables than the reference category. Missing vitamin had a 55% increase in the odds of being amongst the high consumers of fruit and vegetables while consumption of vitamins had 61% increase in the odds. In the alcohol consumption variable, subjects with missing values had 25% increase in the odds of being high fruit and vegetables consumers. Relative to those who smoke every day, missing smoking showed a 2.38 fold increase in the odds. In the education level variable, subjects who had their qualification missing had 24% increase in the odds for being high consumers of fruit and vegetables. There was no evidence in the single variable analyses that subjects with missing marital status were significantly different from the other marital status categories. Although no conclusions could be based on these single variable analyses, these results suggested that eliminating subjects with missing values could bias the estimation of the true impact of these factors on subjects being high or low consumers of fruit and vegetables.

Variable	Odds ratio (95% CI)
Vitamin	
No	1.00
Yes	1.61(1.52-1.70)
Missing	1.55(1.42-1.71)
Drink alcohol	
More than once per week	1.00
Once per week	0.93(0.86-1.01)
Less than once per week	0.84(0.79-0.89)
Never	0.85(0.78-0.92)
Missing	1.25(1.05-1.48)
Vegan/vegetarian	
No	1.00
Yes	2.28 (2.15-2.41)
Missing	1.51 (1.28-1.78)
Smoking	
Smoke everyday	1.00
Smoke occasionally	1.89(1.58-2.56)
Used to smoke daily	2.52(2.26-2.79)
Never smoked	2.36(2.13-2.61)
Missing	2.38(1.20-2.83)
Marital Status	
Married	1.00
Divorced	0.89(0.81-0.97)
Widowed	0.87(0.79-0.97)
Separated	0.72(0.65-0.79)
Single	0.91(0.76-1.07)
Missing	0.88(0.71-1.09)
Education Level	
No qualification	1.00
O-level	1.32(1.22-1.42)
A-level	1.77(1.63-1.93)
Degree	1.84(1.69-1.99)
Missing	1.24(1.11-1.38)

Table 4.3: Single variable logistic regression on all available data, with missing values in categorical covariates replaced by a dummy category for each variable.

An improvement to the complete case analysis, and to be able to include more records in the logistic regression analysis, would be to exclude the variable with the greatest amount of missing values. This could be achieved by excluding the 'physical exercise' variable. The same logistic regression model without the 'physical exercise' variable would have included, 17,999 subjects, instead of 10,126. However, not only is this an

important lifestyle factor to investigate its relation to fruit and vegetable intake, but the confounding effect of 'physical exercise' variable was essential to assess the effect of other lifestyle factors on the consumption of fruit and vegetables. It would therefore not be appropriate to leave it out of the regression model.

Results from the complete case analyses were presented in Table 4.9; this default analysis was used as a baseline method of comparison.

A more useful approach would be to incorporate all information in the incomplete records. In this Chapter two approaches to multiple imputation were explored, which are multiple imputation conditioning on one variable, and multiple imputation by chained equations, which conditions on all the variables in the model. The final results from each analysis were compared to results presented by the complete case analysis in Pollard *et al.* (2001).

4.6 Multiple imputation conditioning on one variable

4.6.1 Plausible values

The plausible values were generated from a model for missing values as described in Section 2.8.1. A realistic proposition was that the associations of the variables among the complete records were the same as among the incomplete records, possibly after conditioning on one or several variables, this is the MAR assumption. For categorical variables, such models could be fitted by cross-tabulating the variables, see Section 2.8.1. When the variables involve many categories, multi-way cross-tabulation would not be feasible because the resulting Table could have many sparse cells.

With the released data, a secondary analyst would complete the dataset with the first set of plausible values, carry out the planned analyses (in effect, oblivious to all issues related to missing values) and store the results. The analyses could then be repeated with the dataset completed by the second set of plausible values, and so on. The results, for each completion, would then be summarized in a straightforward manner, see Section 2.8. These procedures would then yield inferences that use all the available information in the incomplete records and appropriately reflect the information lost due to missing values.

Although generating the plausible values is a complex process, the other elements of the analyses were not affected by the extent or pattern of missing values. Simply, the

intended analysis was carried out several times, using the same programme that handled the problem for the complete dataset. Even though the computing was more extensive (a few times), the programming effort for any analyses was not any greater than if there were no missing values. Summarizing the results from the different completions was straightforward; see Section 2.8. Thus, for the initial investment of generating the plausible values higher quality inferences were gained at a marginal additional cost of analyst's time and effort.

4.6.2 Categorical variables

The approach to categorical variables differs from that to continuous variable. The ideal approach of generating the plausible values for the remaining seven categorical variables included in the logistic regression model would be to condition each variable on all variables in the model. In this Section a simpler approach was preferred of conditioning only on one variable. Since this is a large and complex dataset, and although I was trying to improve the methods, that had already been used, I did not want to employ a very complicated computational routine that might discourage its application.

The variable 'region', had only three values missing as it was computed from the postal code. As it was almost complete, no imputation was needed. In addition no imputation was needed for the complete variables 'illness', 'women with children under 16 years' and 'job'. Each of the remaining seven categorical variables was conditioned on the strongly associated variables.

The correlation matrix, calculated from all available data, was presented in Table 4.4. Most correlations were very small. The strongest correlation was found between highest

educational qualification and the socio-economic class –0.45, but all the other correlations were much smaller. In this Section, every variable was conditioned with the most strongly associated variable i.e. on the one with the highest estimated correlation coefficient.

Based on the Table of correlations, 'smoking' was conditioned on 'alcohol', 'alcohol' on 'highest educational qualification', 'vitamin' on 'highest educational qualification', 'vegetarian/vegan' on 'highest educational qualification', 'socio economic class' on 'highest educational qualification', 'highest educational qualification' on 'social class' and 'marital status' on 'highest educational qualification'.

	Smoking	Alcohol	Vitamin	Vegetarian/ Vegan	Social-class	Highest educational qualification	Marital status
Smoking	**1.000**	0.084	-0.021	0.005	-0.032	0.075	-0.052
Alcohol	0.084	**1.000**	0.004	0.081	0.136	-0.161	0.079
Vitamin	-0.021	0.004	**1.000**	-0.083	0.001	0.101	-0.030
Vegetarian /Vegan	0.005	0.081	-0.083	**1.000**	-0.056	0.133	0.081
Social class	-0.032	0.136	0.001	-0.056	**1.000**	-0.455	-0.071
Higher education	0.075	-0.161	0.10	0.133	-0.455	**1.000**	0.087
Marital status	-0.052	0.079	-0.030	0.081	-0.071	0.087	**1.000**

Table 4.4: * Pearson's correlation coefficients of categorical variables (all available data) included in the logistic regression model

*The degree of associations between categorical variables was measured using Pearson's correlations. Although this measure requires the data to be normally distributed it was used to give a rough guide on the strength of association between variables.

For example the generation of plausible values for the variable 'socio-economic class', conditioning on the 'highest educational qualification' was as follows: -

- Values were imputed for missing items, by computing the conditional distribution of the missing part of the record, given the observed part, and normalizing their probabilities, so that they add up to unity.
- The unit total was found by dividing each entry by the number of complete pairs of values of class and highest educational qualification, equal to 31,873.
- The joint probabilities of (class, highest educational qualification) were estimated from their cross-tabulation shown in Table 4.5, and the normalized probabilities of these two variables were shown in Table 4.6.

Socio-economic class	Highest educational qualification				
	No education	O-level	A-level	Degree	Missing
Never had paid job	65	42	25	17	19
Managers/Administrators	554	1,480	1,108	1,356	299
Professional	80	382	3,094	4,978	143
Associate professional	333	1,720	1,365	1,353	546
Clerical / secretarial	2,043	4,081	1,508	652	1,230
Craft / skilled	186	132	58	45	37
Personal and protective	851	1,242	454	216	296
Sales	654	512	176	72	160
Plant/ machine operative	184	93	23	14	38
Other	415	230	60	20	92
Missing	159	144	75	64	192

Table 4.5: Cross-tabulation of the two variables socio-economic class and highest educational qualification.

Socio-economic class	Highest educational qualification			
	No education	O-level	A-level	Degree
Never had paid job	0.002	0.001	0.001	0.001
Managers/administrators	0.017	0.046	0.035	0.043
Professional	0.003	0.012	0.097	0.156
Associate professional	0.010	0.054	0.043	0.042
Clerical/secretarial	0.064	0.128	0.047	0.020
Craft/ skilled	0.006	0.004	0.002	0.001
Personal and protective	0.027	0.039	0.014	0.007
Sales	0.021	0.016	0.006	0.002
Plant/ machine operative	0.006	0.003	0.001	0.001
Other	0.013	0.007	0.002	0.001
				1.00

Table 4.6: Joint distribution of socio-economic class given highest educational qualification

For missing 'highest educational qualification' given 'socio-economic class' =(*Professional*), the probability of 'highest educational qualification' (*A-level*) = 0.097/ (0.003+0.012+0.097+0.156) =0.36. A plausible conditional distribution was then drawn for the missing part, given the observed part of the record. The standard error associated with this probability is $\sqrt{(0.36\times 0.64)/31{,}873}$ =0.0027, since this was based on a large enough sample for the normal approximation to apply, the percentage of (socio economic class=*professional* | highest educational qualification=*A-level)* was estimated to have a normal distribution with mean 36% and standard deviation 0.26%. A set of five imputations based on the plausible vector of probability ***p***, drawn from the estimated sampling distribution of the estimator ***p*** of the probabilities was then computed.

The same procedure was carried out for "Do you take vitamins, minerals or food supplements" conditioning on "Are you vegetarian or vegan". The joint probabilities of

vitamin (yes, no), vegan (yes, no) were estimated from their cross-tabulation shown in Table 4.7.

Do you take vitamins, minerals or food supplements	**Are you vegetarian or vegan?**		
	No	**Yes**	**Missing**
No	10,119	3,169	260
Yes	12,478	5,701	386
Missing	2,176	920	158

Table 4.7: Cross-tabulation of the variable vitamin and vegan

The complete Table of estimated probabilities for these two variables is shown in Table 4.8.

The estimated joint distribution was obtained after removing the row and columns of missing observations, and normalising it to have a unit total. The unit total was found by dividing each entry by the number of complete pairs of values on 31,467. For example for missing vegan, given vitamin = no, the probability of vegan (no) = 0.32/ (0.32+0.10) = 0.76 and vegan (yes) is drawn with probability = 0.10/(0.32+0.10)= 0.24.

Do you take vitamins, minerals or food supplements	**Are you vegetarian or vegan?**		
	No	Yes	
No	0.32	0.10	
Yes	0.40	0.18	
			1.00

Table 4.8: Conditional distribution of vitamin and vegan

Missing values were imputed for the remaining categorical variables following the same procedure of generating plausible values.

4.6.3 Regression methods for continuous predictors

The independent variables of the logistic regression included two continuous variables, 'age' and 'physical exercise'. A typical imputation for continuous variables would be by regression. In this method a regression equation based on complete case data for a given variable Y could be developed, treating Y as the outcome and using all other relevant variables as predictors. Then data are imputed for the missing values of Y, using values predicted by the regression equation. An improvement on this method involves adding uncertainty to the imputation of Y so that a plausible value is imputed instead of the predicted one. This method was described in Section 2.8.2.

This method was applied to the two continuous variables, 'age' and 'physical exercise'. First, taking the variable 'age' as an outcome and 'physical exercise' as the independent variable, imputing the predicted values of the 'age' variable for missing 'age' from linear regression of observed 'age' on observed 'physical exercise'. An error term was then added, drawn as a random number with mean zero and variance equal to the estimated variance of the residual of the regression equation. In the second step, I imputed the predicted values of 'physical exercise' for missing 'physical exercise' from linear regression of observed 'physical exercise' on observed 'age', and added an error

generated as a random variable from a normal distribution with mean zero and variance equal to the estimated variance of the residual of the regression equation. Five sets of imputations were computed for each variable.

This imputation was kept very simple, i.e. each of the above continuous variables was imputed by conditioning on one variable. That was mainly because all variables were incomplete and the more variables I included in the imputation regression model the less values I could impute. For example let us assume that there were four variables A, B, C and D with pattern of missing values as follows: -

A	**B**	**C**	**D**	**Observations**
.	X	X	X	10
.	X	X	.	10
.	X	.	X	10
.	.	X	X	10
X	X	X	X	50

where X indicate observed values and . represent missing values. Let us assume that the aim was to impute only for missing values **A**. Using regression imputation model with **A** as the outcome and **B, C** and **D** as predictors, it would be possible to fill in only for 10 observations with pattern (. X X X) by imputation. On the other hand if I imputed for **A** conditioning on variable **B**, it would be possible to fill in 30 records which are records with patterns (. X X X), (. X X .) and (. X . X). The same would apply if I conditioned on **C** alone or **D** alone. Therefore, a compromise had to be reached, between finding the most suitable variable or set of variables to condition on. In these selections one has to decide between the variable with the strongest association or correlation, and the one with the least number of missing values to be used as a predictor.

4.6.4 Multiple imputation analysis

Logistic regression with the binary variable (high, low consumption of fruit and vegetable) was applied to the five completed datasets generated through multiple imputation, see Table 4.9. Vegetarians and vegans had almost two and half times higher odds of being high fruit and vegetable consumers when compared to non-vegetarians or non-vegans with (OR= 2.36, 95% CI = 2.21–2.52). Occasional smokers had 61% higher odds for being high fruit and vegetable consumers when compared to daily smokers (OR=1.61, 95% CI = 1.34–1.94). The consumption was even higher for women who used to smoke and women who never smoked, with (OR=2.06, 95% CI 1.65–2.58) and (OR= 2.02, 95% CI 1.84–2.28) respectively. It was found that women who never drink alcohol had 22% lower odds for being high fruit and vegetable consumers (OR=0.78, 95% CI= 0.71–0.86), (OR=0.96, 95% CI 0.88–1.04) for women who drink once per week and (OR=0.84, CI=0.79–0.89) for women who drink alcohol less than once per week, compared to those who drink once per week.

For highest educational qualification, it was found that women with O-levels had 31% increase in the odds for high fruit and vegetable consumption (OR=1.31, 95% CI 1.19–1.42) when compared to women with no qualifications, consumption of fruit and vegetable was even higher among women with A- level (OR=1.52, 95% CI =1.37–1.67) and women with degree (OR=1.49, 95 % CI= 1.31–1.63).

Divorced, widowed, single and separated women were found less likely to consume higher amounts of fruit and vegetables compared to married women, Table 4.9.

4.6.5 Complete case analysis and multiple imputation (conditioning on one variable)

The main aim of Pollard *et al.* (2001) analysis was to identify health and lifestyle factors that contribute to or are associated with higher fruit and vegetables consumption. The conclusion was based on the 10,126 subjects who had complete records and fell into the first and third tertile. To demonstrate the benefits gained by multiple imputation, the complete case analysis was compared with the analysis based on multiple imputation (conditioning on one variable). Table 4.9 presents the results of the two analyses in the form of odds ratios, confidence intervals and standard errors.

The mid-tertile T2 of fruit and vegetable consumers was first removed from the dataset, because this category was not considered in the original analyses of (Pollard *et al.*, 2001). As a result the data were reduced to 23,579 records, and this subset of the data formed the basis of the imputations carried out in this chapter.

The first obvious gain was the number of records included in the logistic regression model, which increased from 10,126 to 23,166 subjects, more than doubling the number of records analysed. It should be noted that multiple imputation conditioning on one variable did not retrieve all records with missing values. This was the result of some records that had missing values in the variable to be imputed for as well as the conditioning variable.

There were small changes in associations in part of the variables before and after imputation, however there were also considerable differences in some of the variables. For example for 'physical exercise' the strength of association changed from 1.28 to

1.12. Occasional smokers had 64% increases in the odds after imputation compared to 80% increase in the odds in the complete case analyses. In the socio-economic group variable the results showed reverse association in the category of managers, this was 0.83 in the complete case analyses compared to 1.05 after imputation, for technical and associate professionals the odds ratios changed from 1.12 to 1.44, and from 0.77 to 0.97 for the clerical and secretarial category.

The table showed smaller standard errors in all the variables suggesting greater precision, as well as shorter confidence intervals.

In the method of multiple imputation applied in this Section, a compromise of conditioning on one variable was reached in order to recover more records for the analysis, that was achieved by conditioning on one variable that was found to have the greatest association with the variable to be imputed. The method was capable of more than doubling the number of records analysed when compared to the complete case analysis. This suggests that with multiple imputation information available in the dataset is used more efficiently, since the recovered information include high percentage of observed information that was excluded from the analysis because of missing values. However, for the assumption of MAR to hold, one should include as many covariates in the imputation model as possible that predict missingness. It is also desirable that the imputation model preserves the structure of the dataset, i.e. the relation between variables as well as uncertainty.

To overcome the mentioned limitations this method was further improved in the next Section. The next Section describes multiple imputation by chained equation in which each variable was imputed conditioning on all the variables in the model.

Variables	n=10,126 Complete case analyses, odds ratio, (95% CI) [s.e.]	n=23,166 MI conditioning on one variable odds ratio, (95% CI) [s.e.]	n=23,575 MI by chained equation, odds ratio, (95% CI) [s.e.]
Age	1.04 (1.03–1.04)[0.004]	1.02(1.01–1.02)[0.002]	1.02(1.02–1.03)[0.002]
Physical exercise	1.28 (1.18–1.39)[0.054]	1.12(1.06–1.18)[0.027]	1.04(0.97–1.11)[0.036]
Vegetarian status			
Vegetarian / vegan	2.24 (2.03–2.46)[0.108]	2.36(2.21–2.50)[0.032]	2.38(2.23–2.53)[0.032]
Women who			
Take vitamin or mineral supplements	1.52 (1.40–1.65)[0.065]	1.45(1.36–1.54)[0.032]	1.51(1.42–1.60)[0.029]
Illnesses			
Women who have had any of the defined illnesses*	0.97 (0.88–1.08)[0.049]	1.08(1.01–1.14)[0.032]	1.07(1.00–1.15)[0.031]
Drink alcohol			
More than once per week	1.00	1.00	1.00
Once per week	0.97 (0.86–1.10)[0.062]	0.96(0.88–1.04)[0.041]	0.97(0.89–1.05)[0.041]
Less than once per week	0.82 (0.74–0.92)[0.045]	0.84(0.78–0.90)[0.035]	0.84(0.80–0.89)[0.035]
Never	0.87 (0.75–1.01)[0.068]	0.78(0.71–0.86)[0.046]	0.78(0.72–0.84)[0.047]
Smoking habit			
Smoke daily	1.00	1.00	1.00
Smoke occasionally	1.80 (1.37–2.38)[0.256]	1.64(1.37–1.98)[0.093]	1.64(1.35–1.96)[0.094]
Used to smoke	2.08 (1.75–2.48)[0.184]	2.06(1.84–2.29)[0.055]	2.11(1.88–2.35)[0.056]
Never smoked	2.00 (1.69–2.35)[0.170]	2.02(1.82–2.24)[0.052]	2.07(1.86–2.30)[0.054]
Marital Status			
Married	1.00	1.00	1.00
Divorced	0.78 (0.67–0.90)[0.058]	0.83(0.75–0.91)[0.049]	0.82(0.75–0.90)[0.049]
Widowed	0.72 (0.58–0.89)[0.080]	0.81(0.71–0.89)[0.059]	0.79(0.70–0.87)[0.059]
Separated	0.62 (0.53–0.73)[0.052]	0.63(0.57–0.70)[0.054]	0.62(0.58–0.69)[0.055]
Single	0.66 (0.51–0.85)[0.088]	0.80(0.67–0.97)[0.092]	0.79(0.68–0.95)[0.093]
Education Level			
No qualifications	1.00	1.00	1.00
O-level	1.46 (1.26–1.69)[0.109]	1.30(1.19–1.42)[0.043]	1.37(1.25–1.49)[0.044]
A-Level	1.66 (1.41–1.95)[0.135]	1.52(1.37–1.67)[0.050]	1.64(1.49–1.79)[0.048]
Degree	1.61 (1.36–1.91)[0.141]	1.49(1.33–1.64)[0.053]	1.61(1.44–1.79)[0.054]

Table 4.9: Results from the logistic regression model describing the relative probabilities of being a high fruit and vegetable consumer

Table 4.9 (continued)

Variables	n=10,126 Complete case analyses, Odds ratio, (95% CI)[s.e.]	n=23,166 MI conditioning on one variable odds ratio, (95% CI)[s.e.]	n=23,575 MI by chained equation, odds ratio, (95% CI)[s.e.]
Employment Status			
Employed	1.00	1.00	1.00
Housewives	1.18 (1.04–1.35)[0.079]	1.17(1.07–1.27)[0.043]	1.17(1.06–1.28)[0.042]
Unemployed	0.83 (0.56–1.24)[0.169]	0.93(0.71–1.19)[0.131]	0.93(0.74–1.21)[0.130]
Retired	1.01 (0.86–1.18)[0.082]	1.00(0.90–1.09)[0.049]	1.00(0.91–1.09)[0.048]
Students	0.92 (0.60–1.39)[0.197]	1.01(0.70–1.32)[0.156]	1.01(0.74–1.36)[0.157]
Unknown	1.14 (0.97–1.35)[0.098]	1.04(0.94–1.15)[0.052]	1.04(0.94–1.15)[0.049]
Socio-economic group:			
Never had a paid job	1.00	1.00	1.00
Managers/admin.	0.83(0.35–1.84)[0.337]	1.05(0.72–1.52)[0.199]	1.02(0.68–1.51)[0.200]
Professional	1.02 (0.46–2.25)[0.412]	1.32(0.90–1.92)[0.190]	1.25(0.84–1.85)[0.251]
Technical & associate professional	1.12 (0.51–2.47)[0.453]	1.44(0.99–2.07)[0.188]	1.40(0.94–2.06)[0.198]
Clerical & secretarial	0.77 (0.35–1.69)[0.309]	0.97(0.67–1.40)[0.186]	0.95(0.65–1.40)[0.198]
Craft & skilled	0.78 (0.33–1.83)[0.340]	1.05(0.68–1.60)[0.216]	1.05(0.66–1.64)[0.230]
Personal/protective	0.86 (0.39–1.91)[0.350]	1.11(0.77–1.62)[0.191]	1.11(0.75–1.65)[0.201]
Sales	0.64 (0.29–1.44)[0.264]	0.85(0.58–1.23)[0.194]	0.85(0.57–1.27)[0.204]
Plant/machine oper.	0.73 (0.30–1.81)[0.340]	0.87(0.53–1.40)[0.245]	0.90(0.56–1.44)[0.240]
Other	0.67 (0.29–1.23)[0.286]	0.77(0.52–1.16)[0.205]	0.79(0.51–1.20)[0.216]
Women living in:			
North East	1.00	1.00	1.00
North West	0.73 (0.56–0.95)[0.099]	0.74(0.63–0.89)[0.087]	0.75(0.63–0.89)[0.087]
Yorkshire & the	0.85 (0.65–1.11)[0.118]	0.93(0.76–1.09)[0.090]	0.93(0.79–1.11)[0.089]
Humber	0.77 (0.58–1.02)[0.110]	0.97(0.80–1.16)[0.094]	0.97(0.81–1.17)[0.094]
East Midlands	0.77 (0.59–1.01)[0.107]	0.89(0.74–1.06)[0.090]	0.90(0.75–1.07)[0.089]
West Midlands	0.89 (0.67–1.15)[0.120]	0.97(0.80–1.16)[0.088]	0.99(0.84–1.18)[0.088]
East of England	0.89 (0.69–1.16)[0.119]	0.91(0.76–1.07)[0.086]	0.93(0.79–1.08)[0.084]
Greater London	0.87 (0.68–1.12)[0.110]	0.94(0.79–1.09)[0.082]	0.95(0.82–1.11)[0.081]
South East	1.02 (0.79–1.32)[0.134]	1.10(0.91–1.28)[0.085]	1.10(0.92–1.30)[0.085]
South West	0.80 (0.61–1.06)[0.114]	0.92(0.77–1.09)[0.085]	0.92(0.78–1.09)[0.091]
Scotland	0.88 (0.63–1.21)[0.145]	0.93(0.75–1.13)[0.104]	0.93(0.77–1.14)[0.103]
Wales	2.16 (0.65–7.25)[1.336]	0.81(0.43–1.51)[0.317]	0.88(0.48–1.61)[0.310]
Northern Ireland	0.93 (0.20–4.20)[0.717]	1.21(0.14–2.29)[0.548]	1.09(0.46–2.57)[0.437]
Channel Isles			
Women with:			
Children under the age of 16 years	1.09 (0.98–1.21)[0.161]	1.02(0.94–1.00)[0.040]	1.02(0.95-1.11)[0.040]

Table 4.9: Results from the logistic regression model describing the relative probabilities of being a high fruit and vegetable consumer (n=23,166)
***Defined illnesses were self-reported heart attack, coronary thrombosis, myocardial infarction, angina, stroke, hypertension, hyperlipidaemia, diabetes or cancer.**

4.7 Multiple imputation by chained equations

In the previous Section, missing data in the variables of the logistic regression model were imputed using multiple imputation, conditioning on one variable. Table 4.4 showed that with the exception of two variables correlations between most of the variables in the model were rather weak. Generating imputations by conditioning each incomplete variable on one variable helped to make use of a lot of information in the incomplete records, which would have been wasted otherwise. Researchers have developed several approaches since the first book on multiple imputation was written (Rubin, 1987). This book however, did not cover methods for imputing multi-variable data. A few years later, Li (1988) and Rubin and Schafer (1990) presented multivariate multiple imputation, using Bayesian simulation. The method assumes that the data are missing at random and it has a multivariate normal distribution. In this Section a different approach will be used for multiple imputation. This approach is similar to that described by Van Buuren *et al.* (1999). The method assumes that a multivariate distribution exists, without specifying a specific form for it, and that draws from it can be generated by Gibbs sampling the conditional distributions, i.e. the multivariate problem is split into a number of univariate problems. Let us assume that the dataset $Y=(Y_1, Y_2,\ldots,Y_p)$ is a set of p variables, and that each variable Y is partially observed ($Y=(Y_{obs}, Y_{mis})$), and let t denote an iteration counter. Assuming MAR the Gibbs sampler draws imputations by following a sequence of iterations:-

For Y_1, draw imputations for Y_1^{t+1} from $P(Y_1 | Y_2^t, Y_3^t, \ldots\ldots Y_p^t)$

For Y_2, draw imputations for Y_2^{t+1} from $P(Y_2 | Y_1^{t+1}, Y_3^t, \ldots\ldots Y_p^t)$

.

.

.

For Y_p, draw imputations for Y_p^{t+1} from $P(Y_p | Y_1^{t+1}, Y_2^{t+1}, \ldots\ldots Y_{p-1}^{t+1})$

One conditions each time on the most recently drawn values of all other variables.

Rubin and Schafer (1990) presented that if P(Y) is multivariate normal then iterating linear regression models like

$$Y_1 = \beta_{12} Y_2^t + \beta_{13} Y_3^t + \ldots \beta_{1p} Y_p^t + \varepsilon_1 \quad \text{with } \varepsilon_1 \sim N(0, \sigma_1^2)$$

produce a random draw from the wanted distribution.

Multiple imputation by chained equations differs slightly from this approach, this method which is sometimes referred to as variable by variable multiple imputation proceeds as follows: -

- Fill in missing values for each incomplete variable by a starting value, this can be chosen as a mode for categorical variables and the mean for continuous variables.
- Discard the filled-in values from the first variable leaving the original missing values. The missing values are then imputed using linear regression, polytomous regression or logistic regression, conditioning on other variables as described below.
- The filled-in values are discarded from the second variable. These missing values are then imputed using a form of regression imputation.

- The procedure is repeated for each variable in turn. Once each variable has been imputed, we have then completed one iteration.
- The same procedure is repeated for several (in this case 10) iterations. This generates one complete dataset.
- For m completed datasets, repeat the procedure m times independently.

This imputation procedure was programmed using STATA 8, which is included as Appendix D. The algorithm imputed missing values in an iterative way for each variable in the model one at a time, conditioning on all remaining variables in the logistic regression model.

The imputation methods were chosen as linear regression for the continuous variable 'age' and 'physical exercise', logistic regression for the binary variables, vitamin consumption, vegetarian status and polytomous logistic regression for the categorical variables, alcohol consumption, smoking, socio-economic group, marital status, education level. Since most predictor variables have missing data, an initial starting value was imputed for each variable; the starting value was set as the mode for categorical variables and the mean for continuous variables. Plausible values were generated through a sequence of ten iterations. Five completed datasets were then generated, and for each completed dataset a logistic regression model was fitted similar to the model fitted in (Pollard *et al.* 2001). The final results were finally combined using multiple imputation combining rules 2.3-2.6 described in Chapter 2.

The binary outcome variable, high or low consumers of fruit and vegetables included 23, 579 observations rather than 35,367 as the result of the exclusion of the mid-tertile, see Section 4.2 and 4.4.

4.7.1 The quality of the imputed values

The only way to check the quality of the imputed values is to compare them with the actual data, which were not available. Therefore, one can often only hope that by choosing the most sensible imputation model, given the observed values of the donors, with the assumption of the missing data being MAR, and then pooling the results of inferences of the completed datasets that would yield valid results.

To check whether the completed datasets were in line with the observed data, I checked some characteristics of the completed datasets. For the two continuous variables, 'physical exercise' and 'age', histograms of the observed values and the imputed values in the five completed datasets were shown in Figure 4.1 and 4.2. For 'physical exercise', the Figure 4.1 shows that the distribution of the observed values and the five imputed datasets are positively skewed, however the distribution appears shifted towards higher values as a result of imputation. The observed values of 'physical exercise' show signs of digit preference. The mean age of imputed values, appear slightly higher than the observed age.

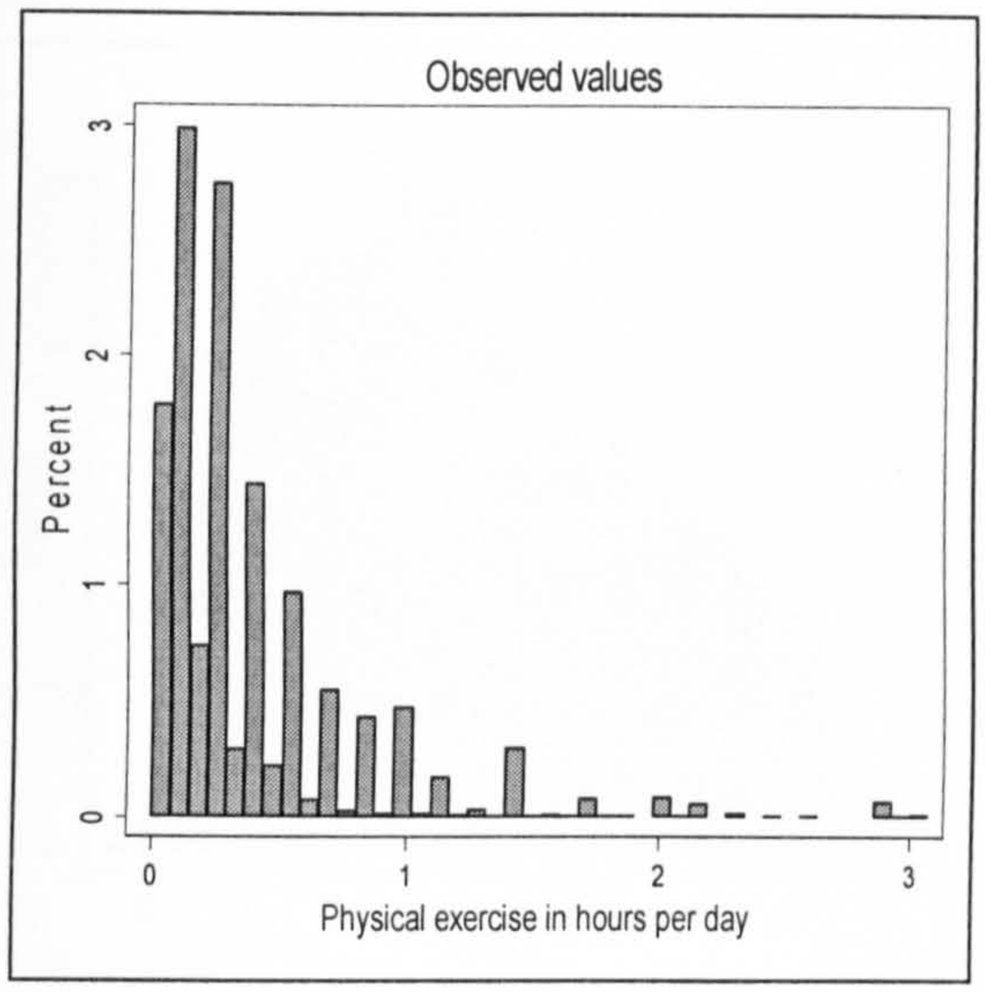
The association between variables with strongest correlations between observed values and the five completed datasets were also compared. Table 4.10 showed very slight changes in the correlations between variables in the observed dataset when compared to the completed dataset. This suggested that after the imputation associations between variables remained the same.

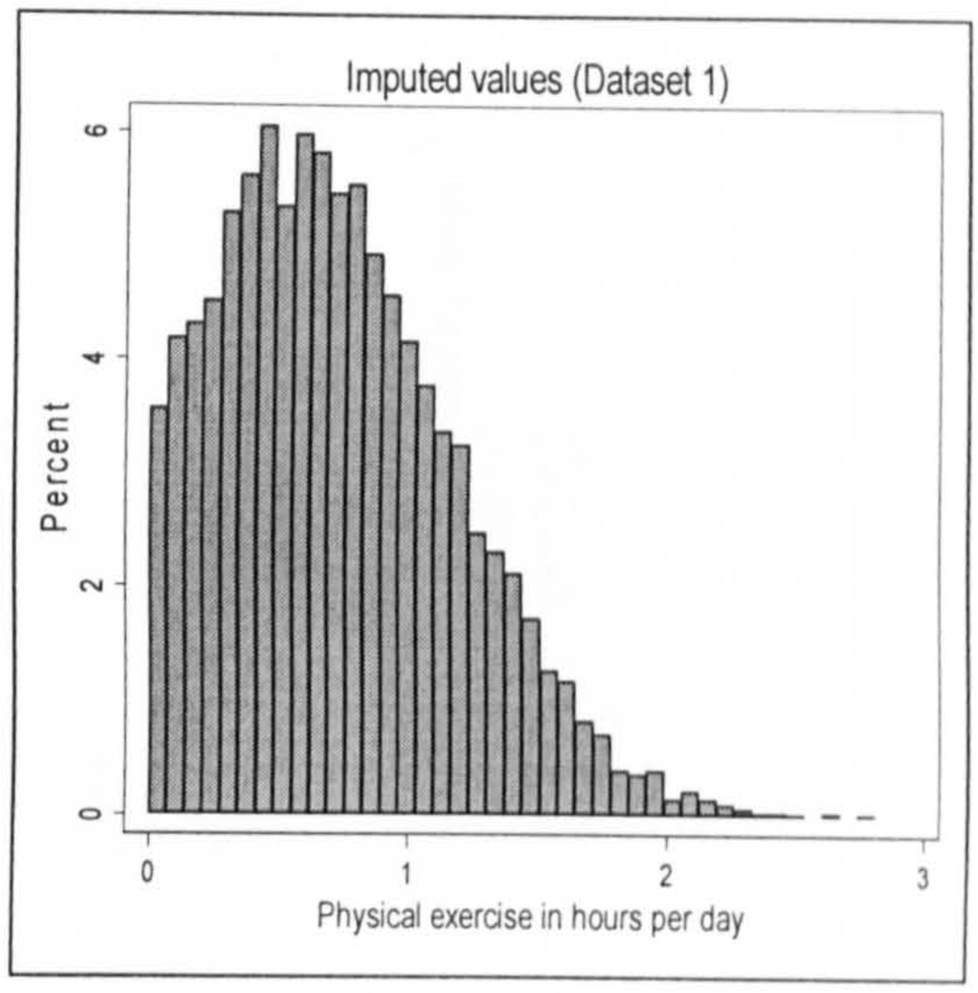
Correlations variables	Observed values	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Smoking/ Alcohol	0.0844	0.0818	0.0822	0.0808	0.0842	0.0857
Alcohol/ Social class	0.1359	0.1365	0.1355	0.1373	0.1354	0.1359
Vegetarian/ Highest educational qualification	0.1340	0.1419	0.1383	0.1378	0.1408	0.1430
Social class/Highest educational qualification	-0.4533	-0.4584	-0.4575	-0.4588	-0.4573	-0.4571

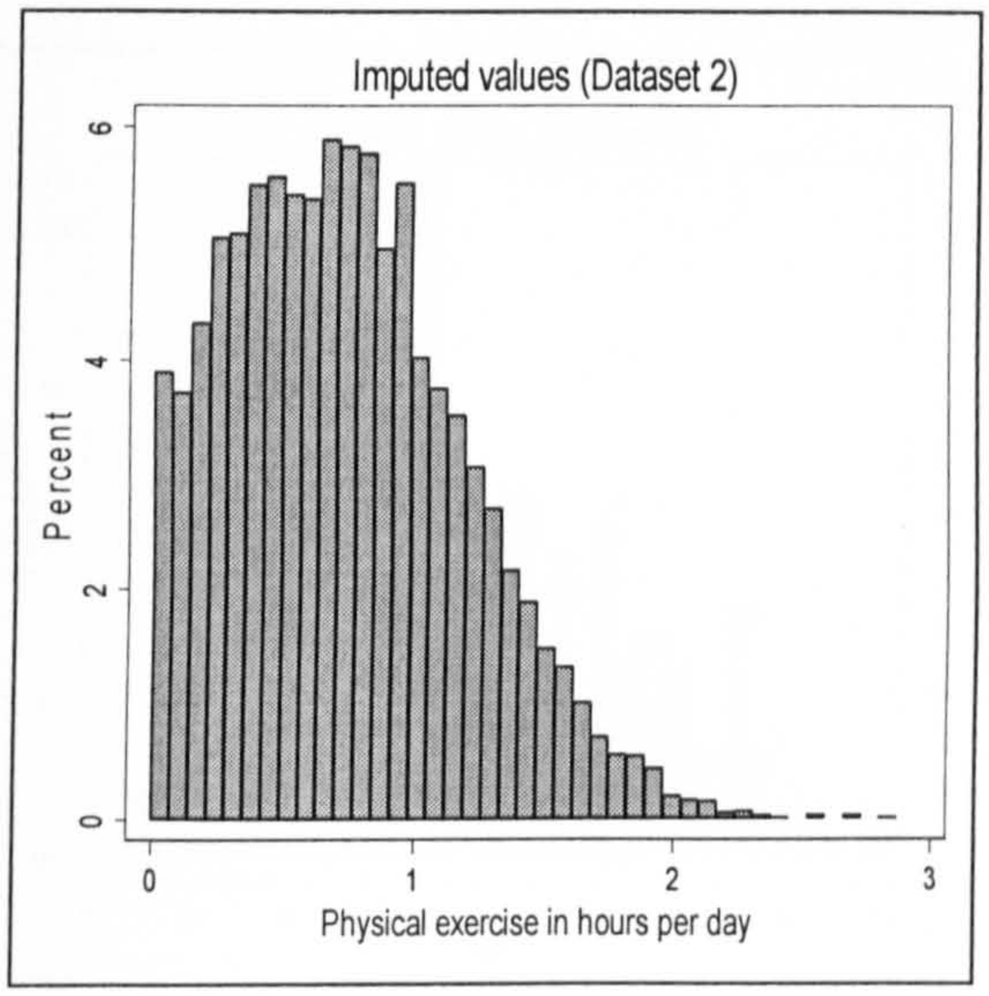
Table 4.10: Correlations between variables with the strongest association, in observed dataset and the five completed datasets.

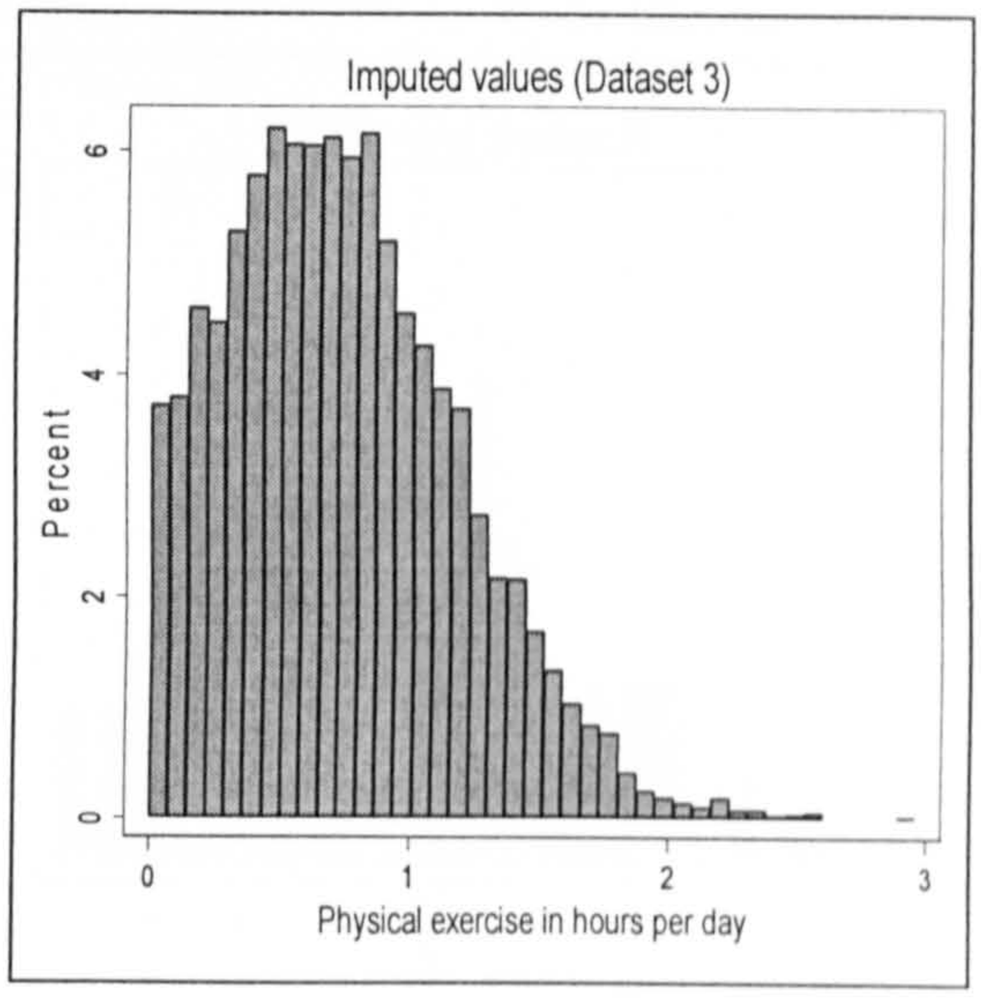
4.7.2 Monitoring convergence

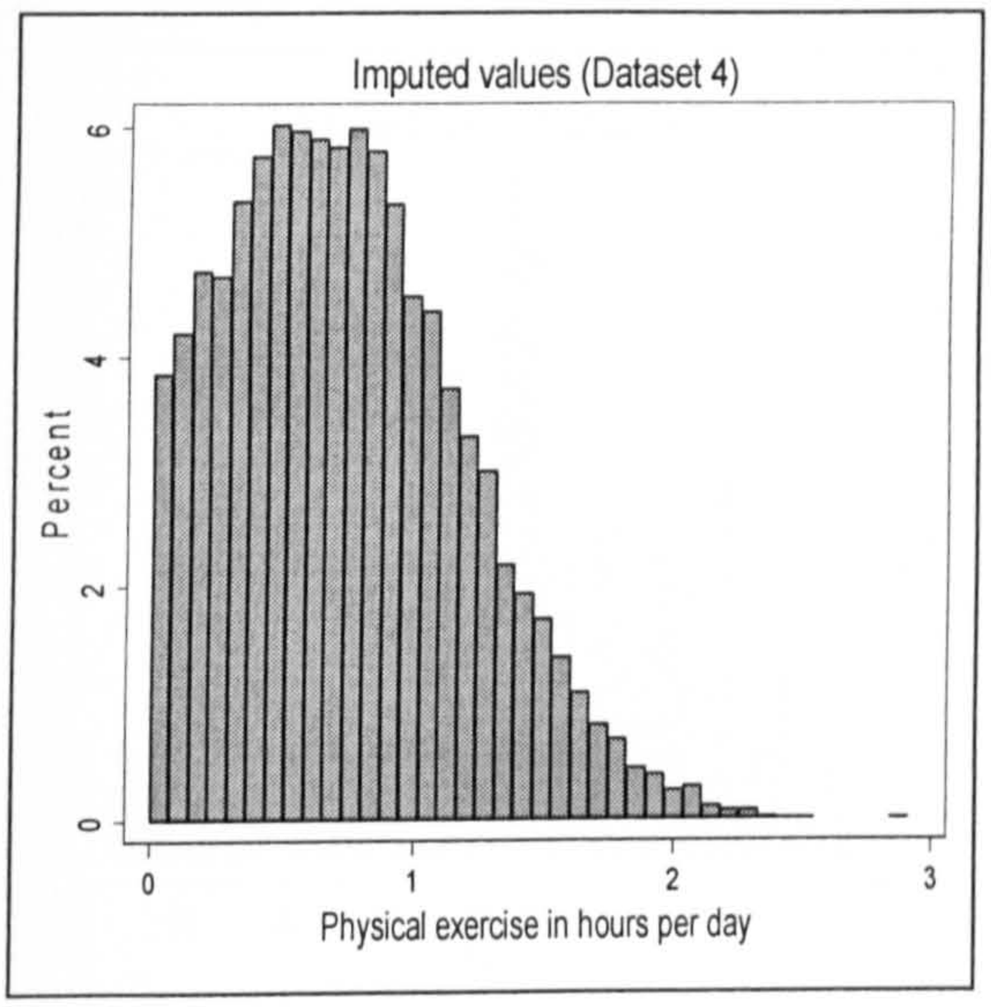
The chain of the Gibbs sampling should be iterated until it reaches convergence or when the chain reaches equilibrium. There is no definite method to assess that the algorithm has converged. The main aim would be to choose sufficient number of iterations to stabilize the distribution of the parameters. Brand (1999) applied the variable-by-variable Gibbs sampling algorithm and reached satisfactory results with 5 iterations. However, Brand's results were based on a moderate amount of missing values in variables. In my case, the amount of missing values varied between variables and the worst scenario was that of the variable 'physical exercise', where missing values were around 44%, see Table 4.2. A plot of the mean and standard deviations of the 'physical exercise' variable, through the











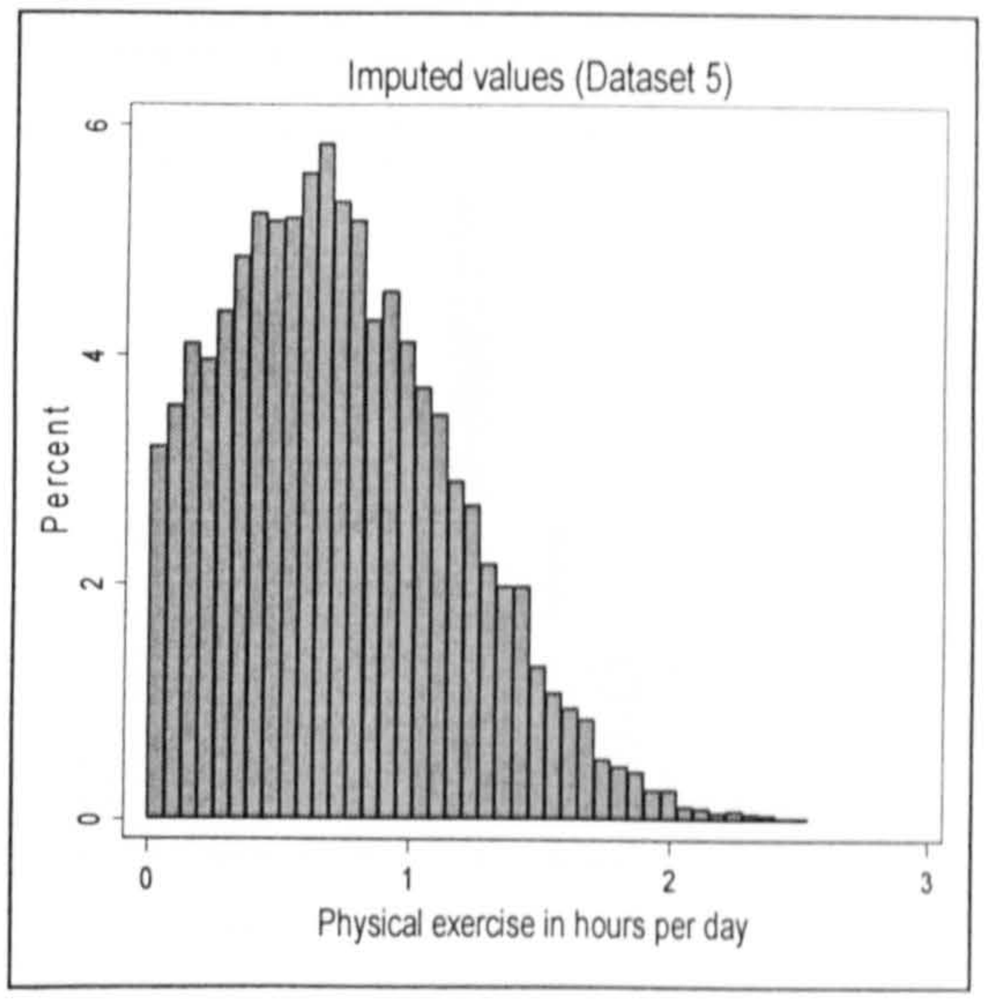


Figure 4.1: 'Physical exercise', the observed values and the imputed values in the five completed datasets

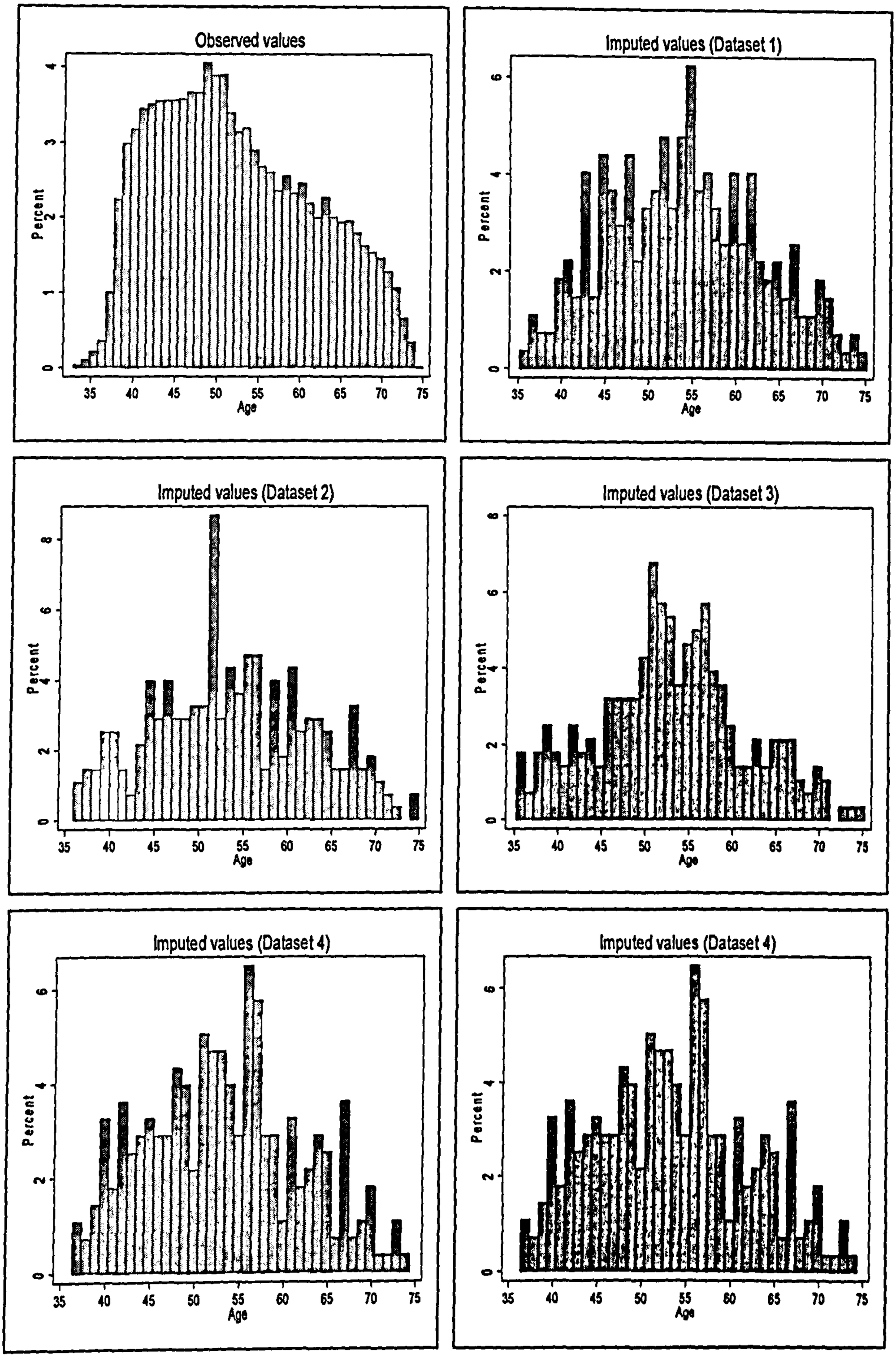


Figure 4.2: Age, the observed values and the imputed values in the five completed datasets

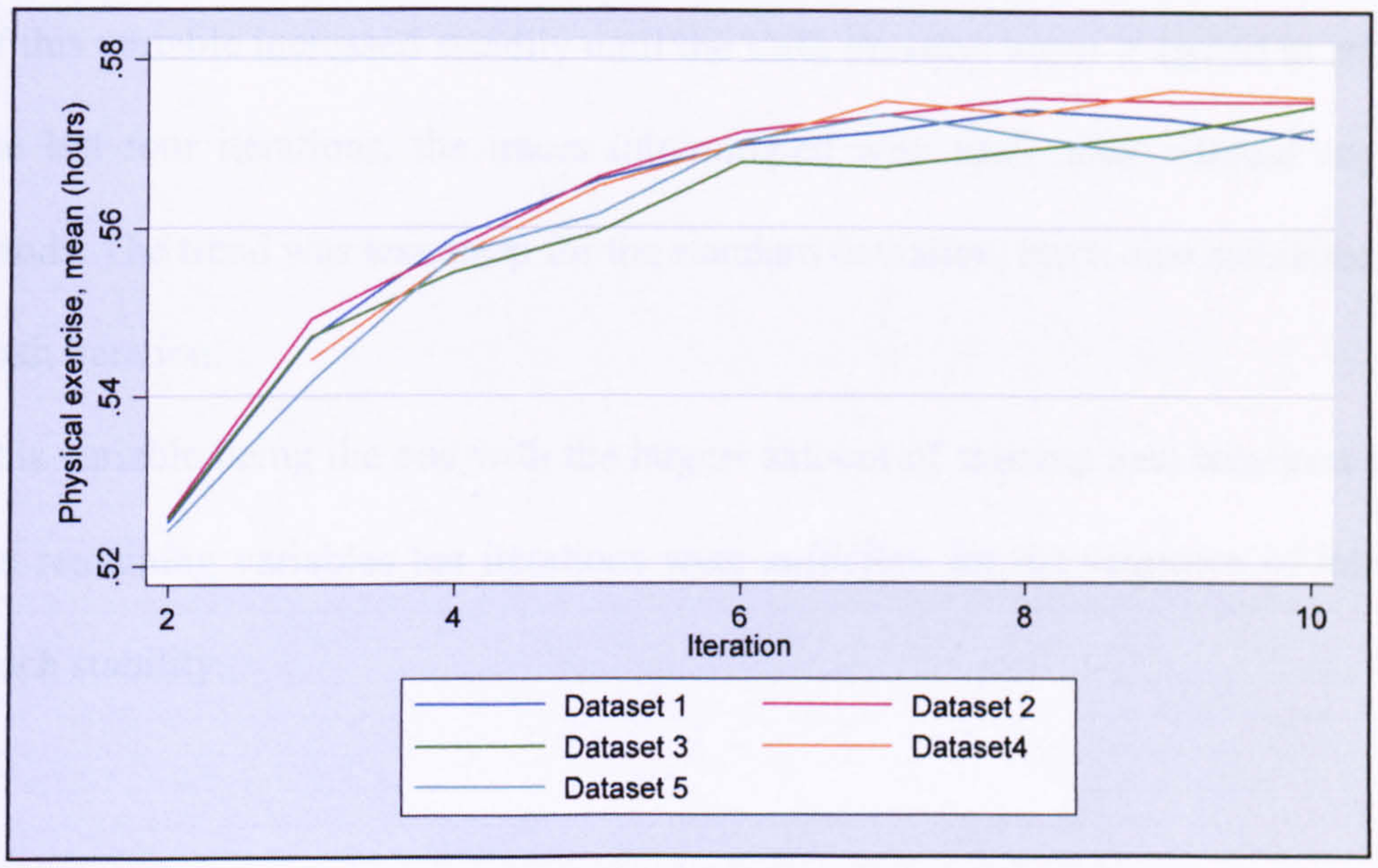


Figure 4.3: 'Physical exercise' (mean) within ten iterations in the five imputed datasets

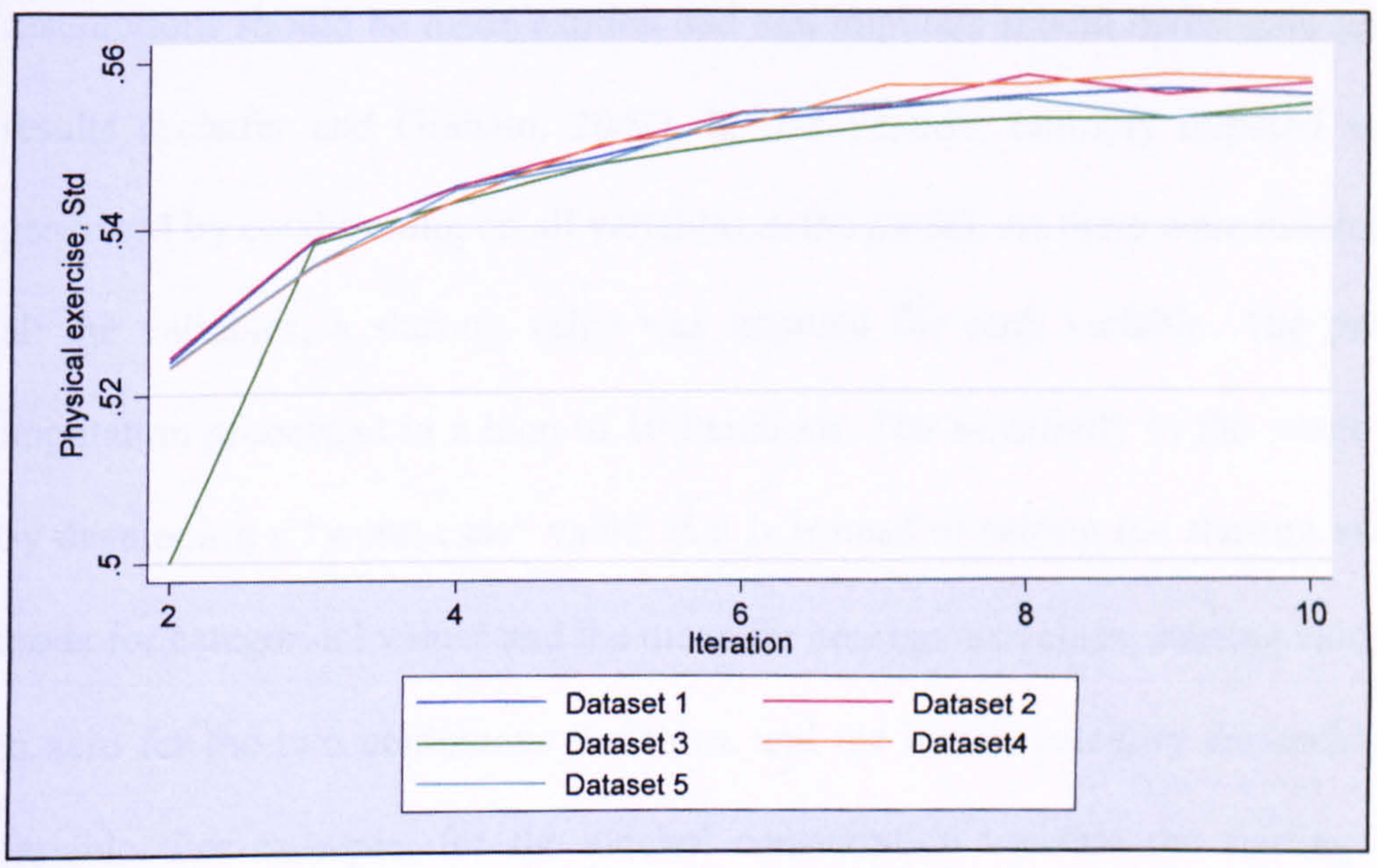


Figure 4.4: 'Physical exercise' (Std deviation) within the ten iterations in the five imputed datasets

sequence of iterations in the five datasets, see Figure 4.3 and 4.4, showed that the mean of this variable increased steadily until the sixth iteration when it started to stabilize. In the last four iterations, the traces intermingled with each other without any definite trends. The trend was less steep for the standard deviation, but it also stabilized after the sixth iteration.

This variable being the one with the largest amount of missing data reassures us that in the remaining variables ten iterations were sufficient for the sequence of iterations to reach stability.

4.7.3 Sensitivity analysis

In the process of imputing missing values, certain assumptions are frequently made, and in most of the cases, they are un-testable. Researchers therefore recommend that assumptions should be made explicit and that imputers should investigate sensitivity of results (Schafer and Graham, 2002). In this Section, multiply imputed values were generated by conditioning on all variables in the model. As there were missing values in all the variables, a starting value was imputed for each variable. The procedure of imputation proceeded in a loop of 10 iterations. The sensitivity of the result was tested by developing a "worst-case" value, that is instead of setting the starting values to the mode for categorical values and the mean for continuous values, starting values were set to zero for the two continuous variables, and the lowest category for each categorical variable. For example, for the alcohol consumption variable the starting value was imputed as the lowest category 1 representing "Drinking alcohol more than once a

week". This would then help to test the sensitivity of the results to starting values, and whether more iterations were needed for the model to reach convergence.

The mean values of the continuous variable 'physical exercise' was plotted against iteration number for the five imputed datasets, see Figure 4.5. The graph showed that means increased steeply until the sixth iteration and then they became stable. The plot for the standard deviations of this variable displayed a similar pattern, with stability achieved after six or seven iterations, see Figure 4.6. Pairwise associations among the categorical variables were presented in Table 4.11. The strength of the association was assessed between the completed variables after 10 iterations, when the starting values were set to the lowest category rather than the mode of the variable. It was clear from the Table that all the completed variables retained their original associations with only slight changes in correlations.

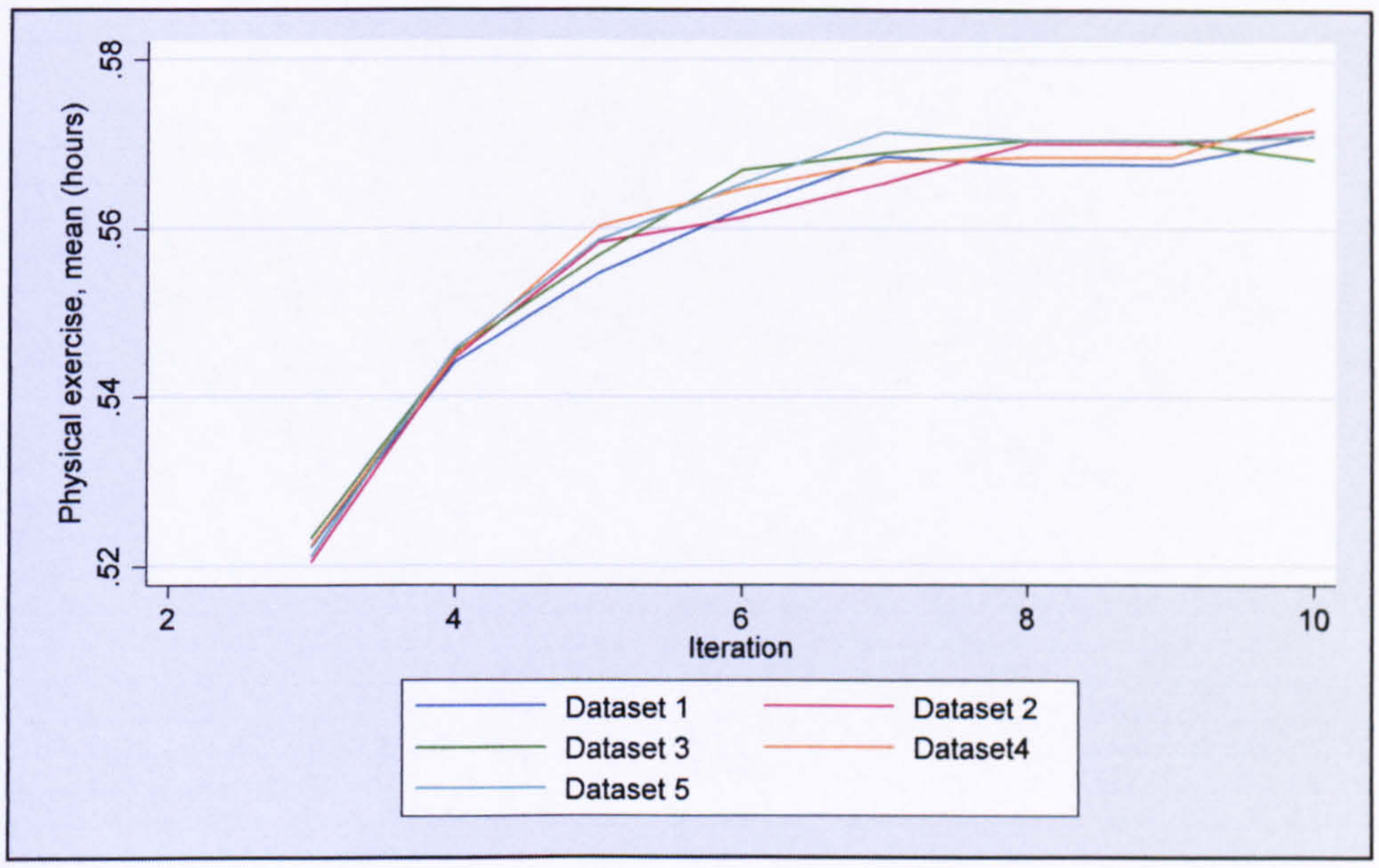


Figure 4.5: 'Physical exercise' (mean) within the ten iterations in the five imputed datasets, with starting value set to zero

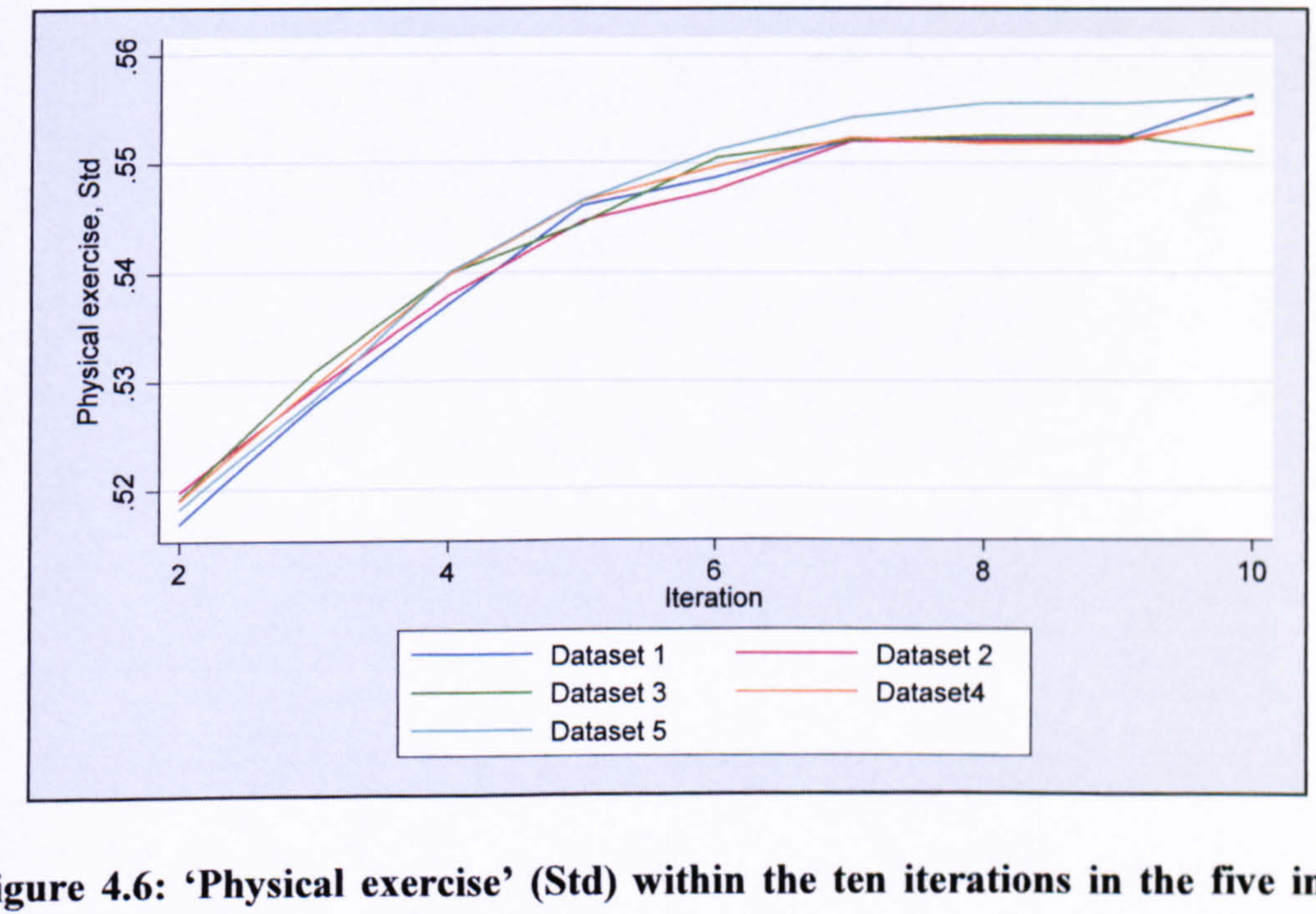


Figure 4.6: 'Physical exercise' (Std) within the ten iterations in the five imputed datasets, with starting value set to zero

Correlations variables	Observed values	Dataset				
		1	2	3	4	5
Smoking/ Alcohol	0.0844	0.0827	0.0841	0.0792	0.0813	0.0838
Alcohol/Social class	0.1359	0.1380	0.1365	0.1363	0.1356	0.1389
Vegetarian/Highest educational qualification	0.1340	0.1384	0.1417	0.1379	0.1421	0.1402
Social class/Highest educational qualification	-0.4533	-0.4582	-0.4577	-0.4596	-0.4548	-0.4598

Table 4.11: Correlations between variables with the strongest association, in observed dataset and the five completed datasets, with the lowest category set as starting value

4.7.4 Comparing complete case analysis and multiple imputation by chained equations

The number of records analysed in the logistic regression model after multiple imputation were 23,575. A gain of more than 300 records, compared to imputation conditioning on one variable and a gain of 13, 262 compared to the complete case analyses. The results presented in Table 4.9 showed narrower confidence intervals and smaller standard errors, compared to the complete case analyses. However, standard errors remained the same as the results after multiple imputation conditioning on a single variable. 'Physical exercise' showed 4% increase in the odds of being high fruit and vegetable consumer compared to 28% increase in the odds in the complete case analyses. This is a substantial reduction in the estimate compared to the complete case analyses. It is also lower than after conditioning on just one variable. For the other

predictors imputation by chained equations yielded similar results to multiple imputation conditioning on one variable, though different from the complete case analyses. Vegetarians and vegans had 38% increases in the odds in the analyses after imputation by the chained equations compared to 24% increase in the odds in the complete case analyses. Occasional smokers relative to those who smoke daily had 64% increase in the odds being high fruit and vegetable consumers in this analyses after imputation compared to 80% increase in the odds in the complete case analyses. Managers and administrators, compared to those who never had a paid job showed 2% increase in the odds in the analyses after multiple imputation. However this category showed decrease in the odds of 17% in the complete case analyses, the category of technical and associate professionals in the same variable had 40% increase in the odds of being high fruit and vegetable consumers, yet this association was only 12% increase in the odds in the complete case analyses. Smaller changes were seen in the other variables.

4.8 Discussion

This chapter replicated Pollard *et al.* (2001) analysis, and compared the results after handling missing data by multiple imputation. A standard and widely applicable logistic regression analysis was used. The first criticism of the analysis was that by dividing the data into three tertiles, and then comparing the highest consumers of fruit and vegetables (T3) with the lowest consumers (T1), almost one third (11,789) of the subjects who were in the middle category (T2) were not considered.

In this logistic regression model, after the exclusion of the mid-tertile, the analyses should have been fitted to 23,579 records. However, only 10,126 records were analysed as the result of missing data.

The chapter discussed imputation of missing data by two methods: -

1 - multiple imputation by conditioning on one variable:

Each variable was conditioned on the most strongly associated variable in the model. Using this method the number of records analysed increased to 23,166. Associations between most variables in the model were rather weak. An argument could be made that because of these weak associations imputed values were nearer to imputing random numbers. However, such imputation facilitates the use of the incomplete records that would otherwise be wasted. The utilization of the wasted recorded information was more important than the specific imputed values. The logistic regression model included 13 variables (outcome variable and twelve predictors). There were 17,705 missing items, but the actual recorded data, which were previously wasted was 172,458 items of information. This was the difference between the total number of records and the analysed number of records (excluding the mid tertile), multiplied by the number of variables. In other words the wasted

records were 10% missing and 90% recorded. The results obtained after multiple imputation were more precise, with smaller standard errors and shorter confidence intervals.

2- Multiple imputation by chained equations:

In this method, multiple imputation was further improved by an approach in which a conditional distribution was specified for each incomplete variable in the model in terms of all the other variables. This method used a series of variable-by-variable Gibbs sampling algorithm. This algorithm is suitable for using existing relationships between variables, when the dataset is large with many incomplete variables. Another advantage of this technique is its flexibility in selecting imputation models, which adequately fits the variable to be imputed. Programming was rather complex, however it was achievable. The quality of the imputed values was found to be similar to the recorded values and the number of iterations used proved to be quite satisfactory for convergence to be reached. This method of conditioning on all the variables proved to be more appropriate than straightforward multiple imputation methods, in this complicated dataset where most of the variables were incomplete.

Pollard *et al.* (2001) stated that although the conclusions of the multi-variable analysis was based on 10,316 participants because of the missing data problems, no changes in the pattern of results were observed in the single variable analyses, and therefore the missing data had no effect on the results. However, the results of the same analyses in this chapter were slightly different in the strength of the effect when odds ratios were compared after handling missing data by multiple imputation, particularly with 'physical exercise', were a large proportion were missing. Nevertheless, the significant effect of most of the variables remained the same. The fact that the difference in the results between multiple imputation and the complete case analyses were close does not

necessarily imply that complete case analyses had been appropriate. The results after multiple imputation had narrower confidence intervals and more precision.

The chapter was not meant to imply that the data collection methods in the survey were fundamentally deficient. The view was that the established methods of analyses ignoring the data imperfections were deficient.

The final results backed up the conclusions of a previous study on diet and lifestyle characteristics associated with dietary supplement use which concluded that the use of supplements (e.g. vitamins) was associated with higher consumption of fruit and vegetables (Kirk *et al.*, 1999).

A criticism of the validity of the results after multiple imputation would be, that the assumption of the data being MAR may not hold, as it could not be tested. However, there was no contradicting proof that the mechanism of missing data was MNAR, which is the only situation when multiple imputation would lead to invalid results this is explored more in Chapter 6. Carrying out the complete case analyses as in Pollard *et al.* (2001), for the results to be valid the mechanism of missing data has to follow a stronger assumption of MCAR. Therefore, results of the logistic regression model after multiple imputation with conditioning on all variables in the model, although not ideal, it is likely to give results closer to the true answers that would be given had a complete dataset been collected.

Chapter 5

Handling missing data in survival analysis of cancer incidence in the UKWCS

5.1 Introduction

There are around 200 different types of cancer and one in three people in the UK will get it at some time in their lives. Cancer is considered the cause of 25% of all deaths in the UK (ONS, 2001).

At the first follow-up of the UKWCS, 2,445 women reported having one or more types of cancer at some point in their life. The age at which they were diagnosed with these cancers was also recorded. However, self-reported cancer is not the most reliable assessment.

The National Cancer Registry also flags incidence of cancers among women taking part in the study, together with the dates these cancers have been registered. According to the information flagged by the National Cancer Registry, there were 2,715 reported cases of cancer among these women before the start of the cohort study. This information compared to the questionnaires shows that around 270 cases of cancer among the respondents were not reported in the questionnaire. The difference may have resulted from lumps that were technically cancers, but which a layperson does not count as such.

A valuable and informative analysis would be to look at the prognostic factors, which significantly predict incidence of breast cancer, the most common type of cancer among women. There is a three-year time gap from first diagnosis of cancer to registration by the cancer registry; therefore, information on incidence of cancer was not complete at the time of this analysis. However, 1,064 cases of breast cancer were flagged to the cohort. Almost 50% of these cases were pre-existing cases, i.e. women who already had the disease at the beginning of the study, 481 cases were diagnosed with breast cancer after taking part in the cohort.

This chapter will illustrate the idea of multiple imputation and its impact on results in a different area of statistics, survival analysis. The imputation model developed for incidence of cancer in this chapter can be used in a later stage of the UKWCS when numbers of incidences are complete.

The Cox proportional hazard model is the most frequently used regression model for the analysis of survival-time data (Hosmer and Lemeshow, 1999), particularly within health science. This model was fitted to incidence of breast cancer; survival time was from the date the questionnaire was received to the reported date of having the cancer. Women who were not flagged by the Cancer Registry were not included in the analysis. This multi-variable model was expected to suffer from missing data, with all variables in the UKWCS being incomplete.

The aim of this chapter is to assess the impact of different methods of handling missing data in the statistical analysis – investigating possible links between incidence of breast cancer and a number of prognostic factors.

Three different procedures for handling missing data were compared and their impact on the results assessed.

5.2 Cancer in women

The classification of factors that predict cancer is of great importance for medical research and clinical practice. Breast cancer is the second most common cancer in the world and the most common cancer among women. Over 41,500 women are diagnosed with the illness every year in the UK, (Cancer Research UK, 2003). Breast cancer develops in the milk-producing glands in the breast or in the ducts that deliver milk to the nipples, it can then spread to the surrounding tissues, and to other parts of the body. The lifetime risk of cancer is one in nine. The incident rates of breast cancer are much higher in industrialized countries than the developing countries, (WHO/FAO, 2003). This report also stated that the only dietary factors that were found to increase the risk of breast cancer are obesity and alcohol consumption.

In this chapter breast cancer cases reported among women who took part in the UKWCS are considered. The chapter investigates the effect of 'smoking', body mass index (BMI), 'age', 'alcohol consumption', having 'children' and 'menopausal status' on getting breast cancer. The literature was searched for evidence of association between these factors and breast cancer.

Smoking

Many studies reported a strong link between smoking and different types of cancer (Miller and Fain, 2003; Haverkos *et al.*, 2003). In a recent study Ghadirian *et al.* (2003), reported a strong association between smoking and pancreatic cancer. Lung cancer, the second commonest cancer in the UK was also found to have strong association with smoking. Smokers as well as ex-smokers were found to have increased risk of lung

cancer, when compared to never-smokers (Ebbert *et al.*, 2003). The literature showed evidence of the link between breast cancer and smoking in women. A cohort study on 90,000 Canadian women which concluded that positive association between breast cancer and smoking, was only found on women who smoked for 40 years, and especially those who smoked 20 cigarettes or more per day, (Dobson, 2002). Smoking can also have strong effect on survival after breast cancer (Manjer *et al.*, 2000). However, many studies show no or little effect of smoking on breast cancer, (Ghadrian *et al.*, 2004; Collaborative group on hormonal cancer, 2002; CGHFBC, 2002).

Alcohol

There is a lot of evidence in the literature showing the association between alcohol consumption and various cancers, hypertension and liver disease (Rehm and Bondy, 1998). Liver cancer is thought to be strongly related to alcohol consumption on epidemiological grounds. Heavy alcohol consumption accounted for 76% of liver cancer deaths in Japan (Makimoto *et al.*, 2000). A large number of studies shows increased risk of breast cancer with the increase in alcohol consumption, with around 10% increase of risk for an average of one alcoholic drink everyday, (Smith-Warner and Spiegelman, 1998). Evidence of a link between alcohol consumption and breast cancer was reported by Lenz *et al.* (2002) and Tjonneland *et al.* (2003). A recent study showed that up to 4% of breast cancer in the developed world can be attributed to alcohol, and the beneficial effect of moderate drinking on the heart and circulation outweigh the increased risk of breast cancer, (CGHFBC, 2002).

Body mass index (BMI)

The relationship between cancer and BMI has been investigated in many studies. Obesity has been found to have a major effect on incidence of breast cancer; heavier women also have a higher mortality due to breast cancer. Daling *et al.* (2001) reported that being in the highest quartile of BMI is a strong predictor of mortality in women with breast carcinoma diagnosed at a young age. It is also estimated that obesity and lack of exercise 'contribute to up to a third of cancers of the colon, breast, kidney and digestive tract', (Josefson, 2001). Obesity was also reported to increase the risk of breast cancer in postmenopausal women by around 50%. Obesity was not found to increase the risk in pre-menopausal women; however, obesity in pre-menopausal women in most of the cases leads to obesity throughout the woman's life and eventually to breast cancer risk, (WHO/FAO, 2003).

5.3 Survival analysis (Complete case analysis)

The Cox proportional hazards model is widely used because it allows for the baseline hazard to be modeled non-parametrically i.e. very flexibly and the ratio of the hazard parametrically. The main assumptions are that the hazards of the groups being compared are proportional so that the hazard ratio is a constant over the whole length of the study.

Cox proportional hazards regression was used to assess disease-free survival and hazard ratios for the prognostic factors in the development of breast cancer in the UKWCS, as

in Hosmer and Lemeshow (1999). Survival time was calculated from the date the questionnaire was received to the date the cancer was diagnosed, and for those who died from the date the form was received to date of death. Survival time for those who were not diagnosed with cancer was calculated from the day the form was received until end of 2002, the last day the cohort was flagged with cancer cases.

5.3.1 Prognostic factors

The prognostic factors, 'smoking', current 'BMI', 'alcohol consumption', 'age', 'menopausal status', and whether the woman had any 'children', were chosen either because they were medically important or because they significantly improved the fit of the model.

To simplify the interpretation, it is always helpful to group any continuous quantity into discrete categories. The body mass index (BMI) is defined as weight/(height)^2, given weight in kilograms and height in meters. The computed BMI was then grouped into four categories: -

1 "Under-weight" BMI<18.5 kg/m^2

2 "Average weight" 18.5≤ BMI<24.9 kg/m^2

3 "Over-weight" 24.9 ≤ BMI<29.9 kg/m^2

4 "Obese" ≥ 29.9 kg/m^2

For example a woman with height of 1.75 m and weight 75 kg has BMI = 75/1.75^2 = 24.49 kg/m^2

The continuous variable age in years was also grouped into 5 categories,

1. 30-40 years
2. 41-50 years
3. 51-60 years
4. 61-70 years
5. 71-75 years

5.3.2 Single variable survival analyses

Single variable analyses were performed on complete cases using Cox regression. Their purpose was mainly to find variables individually predictive of the development of breast cancer; see Table 5.1. The single variable analyses show no significant association between alcohol consumption, BMI, having children and smoking in getting breast cancer. However, there were two factors which were found to increase risk of getting breast cancer, these were being post-menopausal increased the risk by 27% when compared to pre-menopausal women. There was greater risk of getting breast cancer associated with older age, with 81% increase in the risk for women in the age band 41-50 when compared to age-group 30-40, this risk was even higher in the age-group 51-60 and 61-75 with hazard ratios 2.24 and 2.39 respectively. These results should be interpreted with great caution, as this single variable analyses suffers from the following limitations: -

- Single variable analysis does not take into account the confounding effect by other prognostic factors which might change the results

- Missing data can play a major role in the analysis, as one would never be able to know how much these data would have contributed to the analysis if it were observed.

Variable	Hazard ratio	Std. error	P-value	95% C.I.
Alcohol				
Never drink alcohol	1.00			
Less than once a week	1.17	0.19	0.34	0.85–1.62
Once a week	0.94	0.18	0.76	0.65–1.36
More than once a week	0.96	0.15	0.79	0.71–1.30
BMI				
Average weight	1.00			
Under weight	0.68	0.26	0.31	0.32–1.44
Over weight	0.97	0.11	0.82	0.78–1.22
Obese	1.26	0.19	0.12	0.94–1.69
Smoking				
Smoker	1.00			
Used to smoke	1.12	0.19	0.48	0.75–0.97
Never smoked	1.15	0.19	0.37	0.64–0.82
Children				
Yes	1.00			
No	1.21	0.16	0.14	0.93-1.58
Menopausal status				
Pre-menopausal	1.00			
Post-menopausal	1.27	0.12	0.01	1.06-1.53
Age group (years)				
30 - 40	1.00			
41 - 50	1.81	0.42	0.01	1.15-2.86
51- 60	2.24	0.52	<0.01	1.41-3.52
61- 75	2.39	0.56	<0.01	1.50-3.79

Table 5.1: Single variable Cox regression, of the potential risk factors on incidence of breast cancer, on all available data.

5.3.2 The proportional hazards analysis

A multi-variable analysis using Cox regression (Hosmer and Lemeshow, 1999) was secondly fitted. In this type of analysis the hazard for a particular event is found by dividing the number of people who experience the event of interest, in our case Breast cancer, by the number of people known to be at risk of experiencing the event in the same time interval.

When modeling time independent covariates, the only assumption the proportional hazard makes is that the effect of each covariate is the same at all time points. This therefore, calls for testing of the assumption in any model fitted.

The two known procedures of assessing the proportional hazard assumption is either by using a statistical test or graphically. Although there are a number of applied tests for assessing the proportional hazard assumption in the literature, a method originally proposed by Schoenfeld (1982) is incorporated into STATA. This test is based on the analysis of residuals after the Cox proportional hazard model is fitted. Grambsch and Therneau (1994) demonstrated that testing the time dependent covariates is equivalent to testing a nonzero slope in a generalized linear regression of the scaled residuals on functions of time. A non zero slope is an indication of the violation of the proportional hazard assumption. The scaled Schoenfeld residuals are the difference between the covariate at the failure time and the expected value of the covariate at this time. A graph of the scaled Schoenfeld residuals versus a function of time provides a graphical assessment for the proportional hazards assumption. The graph should give a random scatter around the zero line if the proportional hazard assumption is satisfied. The results of fitting the proportional hazard model are shown in Table 5.2. Following the fit

of the model, the proportional hazard assumption was tested using the Schoenfeld residuals, see Table 5.3. The test of proportional hazard assumption on the time-scale was not significant overall for the categories of 'smoking', 'BMI', 'children', 'menopausal status' and 'age' with $p>0.05$, Table 5.3. However, the test showed significant violation for the third category of 'alcohol', *drink alcohol once a week*, with $p<0.05$. The plot of the Scoenfeld residuals on the time scale was then presented graphically, for this factor, see Figure 5.1.

The line in the plot for, *drink alcohol once a week* has slope approximately equal to zero, suggesting that there many be no time-varying effect, which does not agree with the Schoenfeld residuals test.

The scaled Schoenfeld residuals appear to be very sensitive to minor violations of assumption, which the test proved to be statistically significant. This was expected because of the large number of records of the cohort, leading to even small departures from the proportional hazard assumption becoming statistically significant. Since they had only negligible effect on the hazard ratios, this minor violation was ignored in the subsequent analysis.

Variable	Hazard ratio	Std. err.	P-value	95% C.I.
Alcohol				
Never drink	1.00			
More than Once a week	1.06	0.19	0.74	0.74-1.52
Once a week	1.04	0.23	0.84	0.68-1.61
Less than Once a week	1.43	0.27	0.06	0.98-2.08
BMI				
Average	1.00			
Underweight	0.39	0.23	0.11	0.12-1.21
Overweight	0.88	0.11	0.32	0.69-1.13
Obese	1.06	0.18	0.73	0.76-1.49
Smoking				
Never Smoked	1.00			
Smoker	0.91	0.17	0.61	0.64-1.61
Used to Smoke	1.08	0.12	0.48	0.87-1.86
Children				
Yes	1.00			
No	1.39	0.19	0.01	1.06-1.82
Menopausal status				
Pre-menopausal	1.00			
Post-menopausal	0.98	0.14	0.90	0.75-1.29
Age group (years)				
30-40	1.00			
41-50	1.94	0.53	0.01	1.13-3.32
51-60	2.55	0.73	<0.01	1.45-4.50
61-75	2.46	0.74	<0.01	1.36-4.45

Table 5.2: Cox proportional hazard model fitted with 28,166 (81%) observations as the result of missing data in all prognostic factors.

Variable	χ^2	Degrees of freedom	P-value
Alcohol			
Never drink			
More than Once a week	0.63	1	0.43
Once a week	4.62	1	0.03
Less than Once a week	2.13	1	0.14
BMI			
Average			
Underweight	0.14	1	0.71
Overweight	0.29	1	0.59
Obese	0.07	1	0.80
Smoking			
Never smoked			
Smoker	1.31	1	0.25
Used to Smoke	0.04	1	0.85
Children			
Yes			
No	0.32	1	0.57
Menopausal status			
Pre-menopausal	1.68	1	0.19
Post-menopausal			
Age			
30-40			
41-50	0.01	1	0.92
51-60	0.56	1	0.45
61-75	0.29	1	0.59

Table 5.3: Test of the proportional hazard assumption based on Schoenfeld residuals, performed on time scale.

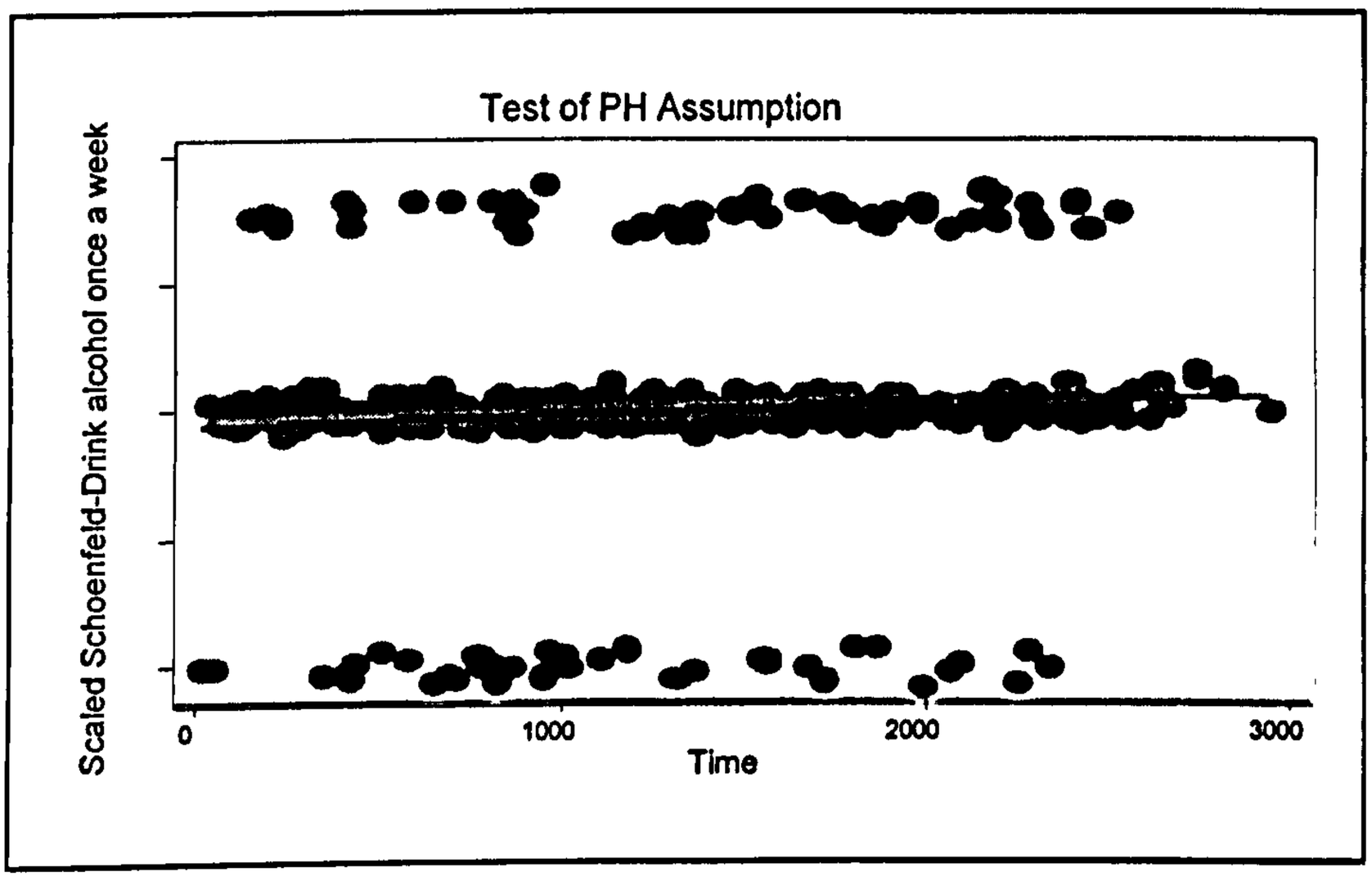


Figure 5.1: Plot of the raw and smoothed scaled Schoenfeld residuals for drink alcohol once a week

5.4 Missing data

In survival analysis, like most other analyses with several covariates, a problem occurs when data are missing on one or more prognostic factors included in the analysis. The easiest and standard response to this type of problem is simply to exclude all incomplete records, which has severe implications for the conclusions as shown below.

When the Cox proportional hazard model was applied to the complete cases, see Table 5.2, only 28,166 observations were included in the model due to missing values, i.e. 19% of the records were excluded from the analysis because their records for the variables used were not complete, see Table 5.4 for the amount of missing values in each variable.

Such loss, as discussed in earlier Chapters 3 and 4 leads to inefficiency, which can be alleviated by a proper handling method for the missing data.

Variable	Observed (%)	Missing (%)
Alcohol		784 (2.3)
Never drink	3,771 (10.9)	
More than once a week	17,695 (51.0)	
Once a week	4,856 (14.0)	
Less than once a week	7,573 (21.8)	
BMI		1,288 (3.7)
Average	20,716 (5.8)	
Underweight	730 (0.1)	
Overweight	8,605 (24.8)	
Obese	3,340 (9.6)	
Smoking		1,007 (2.9)
Smoker	3,725 (10.7)	
Used to smoke	10,434 (30.1)	
Never smoked	19,513 (56.3)	
Children		3,856 (11.1)
Yes	26,561(76.5)	
No	4,262(12.3)	
Menopausal status		2 (0.1)
Pre-menopausal	15,326 (44.2)	
Post-menopausal	19,351 (55.8)	
Age		171 (0.5)
30-40	2,801 (8 .1)	
41-50	13,056 (37.6)	
51-60	10,496 (30.2)	
61-75	8,155 (23.5)	

Table 5.4: Number observed and missing cases in variables included in the survival analysis model

5.4.1 Investigating missing data

One of the assumptions of multiple imputation is missing data being missing at random, see Section 2.2.2 for definition. This mechanism assumes that the missing data depends on the values that are observed, and not on those missing. Imputation models should then take account of the variables that predict missing values. In Chapter 4, missing values were generated by conditioning on all the variables that were included in the logistic regression model. In this Section, evidence of missing data being MAR was investigated by assessing association between missing data and prognostic factors

included in the survival analysis model, as well as other observed auxiliary variables in the dataset that were not included in the model, Table 5.5.

For each variable included in the survival analysis model the relationship between missing values (yes, no) and other variables was investigated using single variable logistic regression models.

The relationship between survival time and the prognostic factors was explored using Kaplan Meier curves stratified by missing values indicator for each of the prognostic factors included in the model as well as the other selected variables from the dataset. Next, for each prognostic factor the difference in distribution of survival times between the missing and recorded data were tested using the log-rank method.

5.4.2. Evidence of missing at random

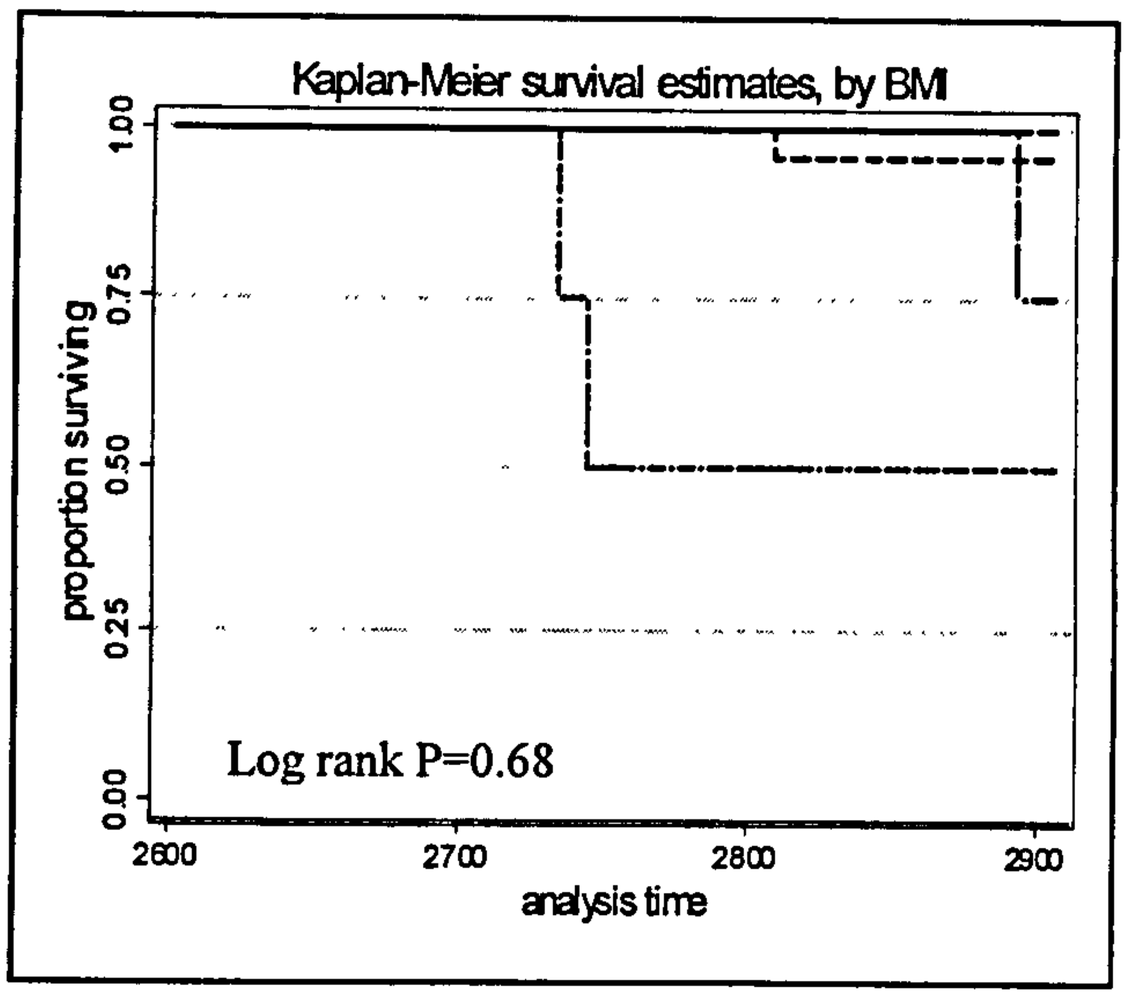
The single variable logistic regression models (missing (yes, no)) for each prognostic factor showed that the missing values were associated with other prognostic factors in the model as well as a number of variables in the dataset which were not included in the survival analysis model, see Table 5.5. For example, missing data in the alcohol variable was predicted by some of the prognostic factors like BMI, children, menopausal status and age, but the missing values in this variable was also predicted by auxiliary variables like marital status, social class, higher educational level and 'physical exercise'. The associations found in these logistic regression models must be interpreted with caution as they were all single variables and were not adjusted for other prognostic factors. By definition, missing values are said to be MAR if they depend on

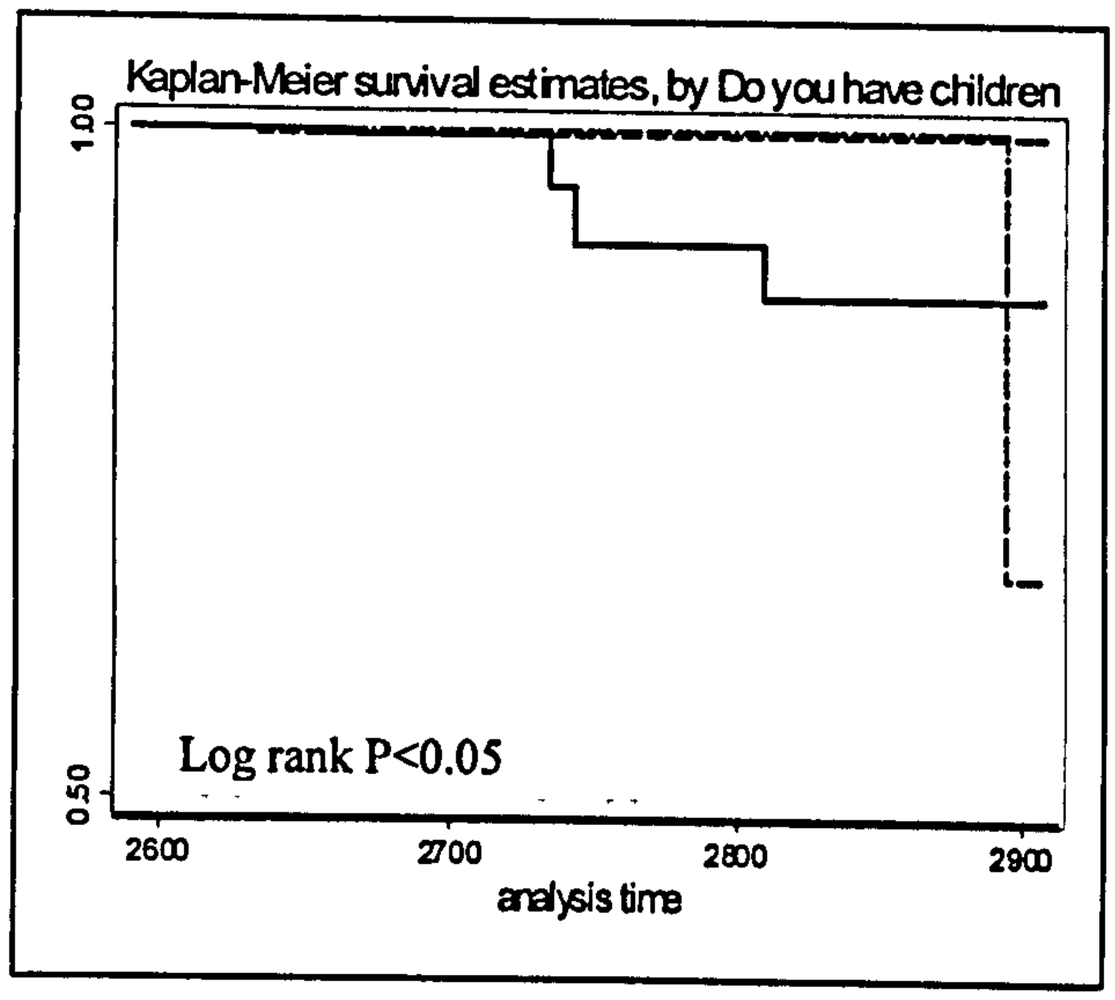
the observed values but not on the missing values. Although this test shows that missing values depend on the variables, which were observed, there is no way to test if the missing values also depend on the variables missing.

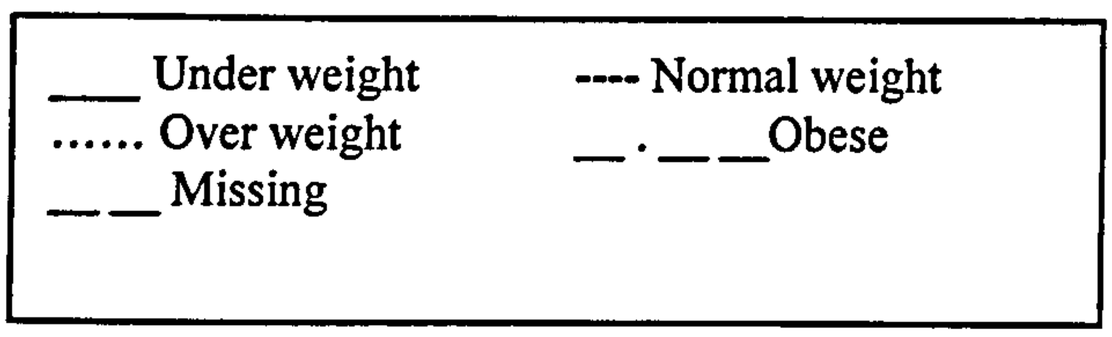
Using the log rank test there was no evidence of difference between the distribution of non-missing and missing strata within each of the prognostic factors alcohol, BMI, age group and smoking with P-values >0.05, this was supported by the Kaplan Meier curves presented in Figure 5.2. However, there was significance difference between survival distributions of observed and missing factors for children (P<0.05), see Figure 5.2. This finding suggests that eliminating the women with missing children can lead to misleading effect of this prognostic factor in the survival analysis model. Similar to the logistic regression models presented above the survival curves presented in Figure 5.2 are not adjusted for other variables; therefore the effect of the prognostic factors might not be real.

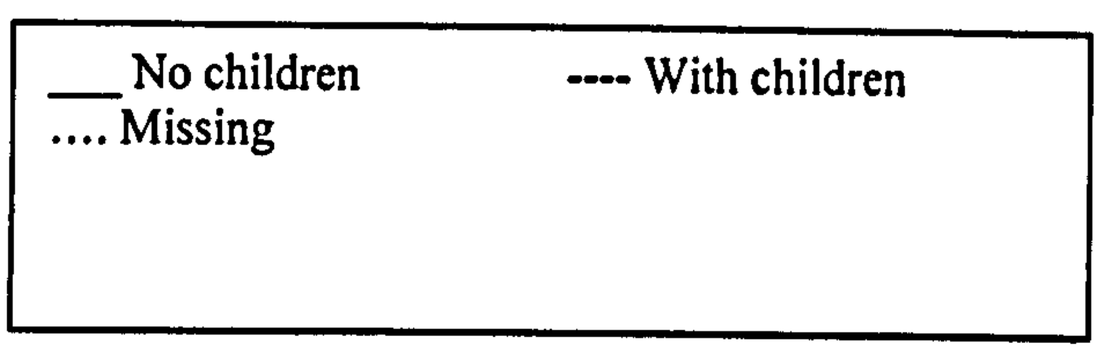
Missing values of	Associated with prognostic variables						Auxiliary variables				Survival time[C]
	Alcohol	BMI	Smoking	Children	Menopausal Status	Age	Marital status	Social class	Higher education level	Physical exercise	
Alcohol		X	_	X	X	X	X	X	X	X	_
BMI	X		_	_	_	X	X	X	X	_	_
Smoking	X	_		_	X	X	_	X	X	_	_
Children	_	X	_		X	_	X	X	X	X	X
Age	_	X	_	_	X		_	X	X	_	_

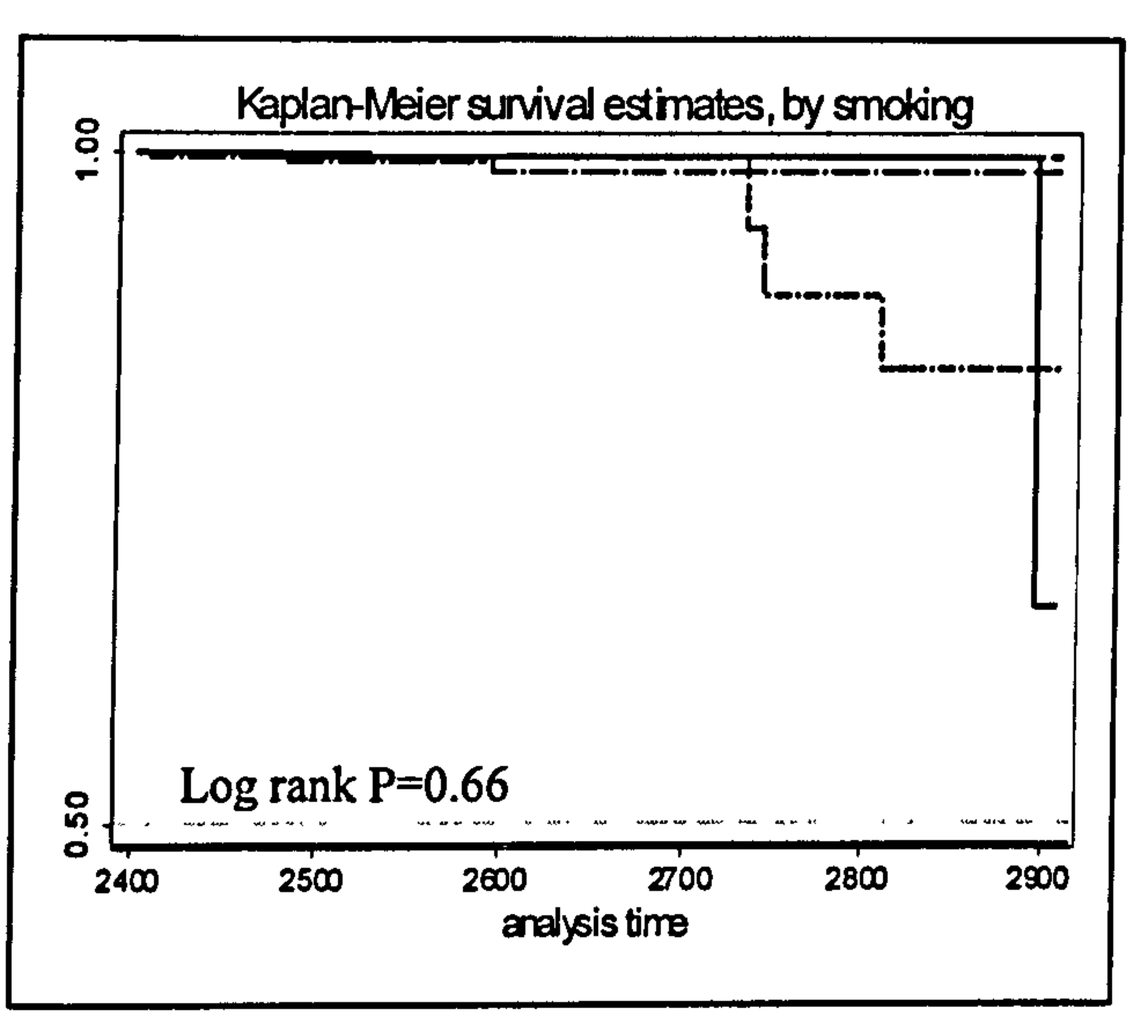
Table 5.5: Associations between missing values and prognostic factors used in the survival analysis model, auxiliary variables and survival time.
X= association using single variable logistic regression (P<0.05), _ = no association (P>0.05) using single variable logistic regression.
[C]= Comparing survival distributions between missing and non-missing groups using log rank test.

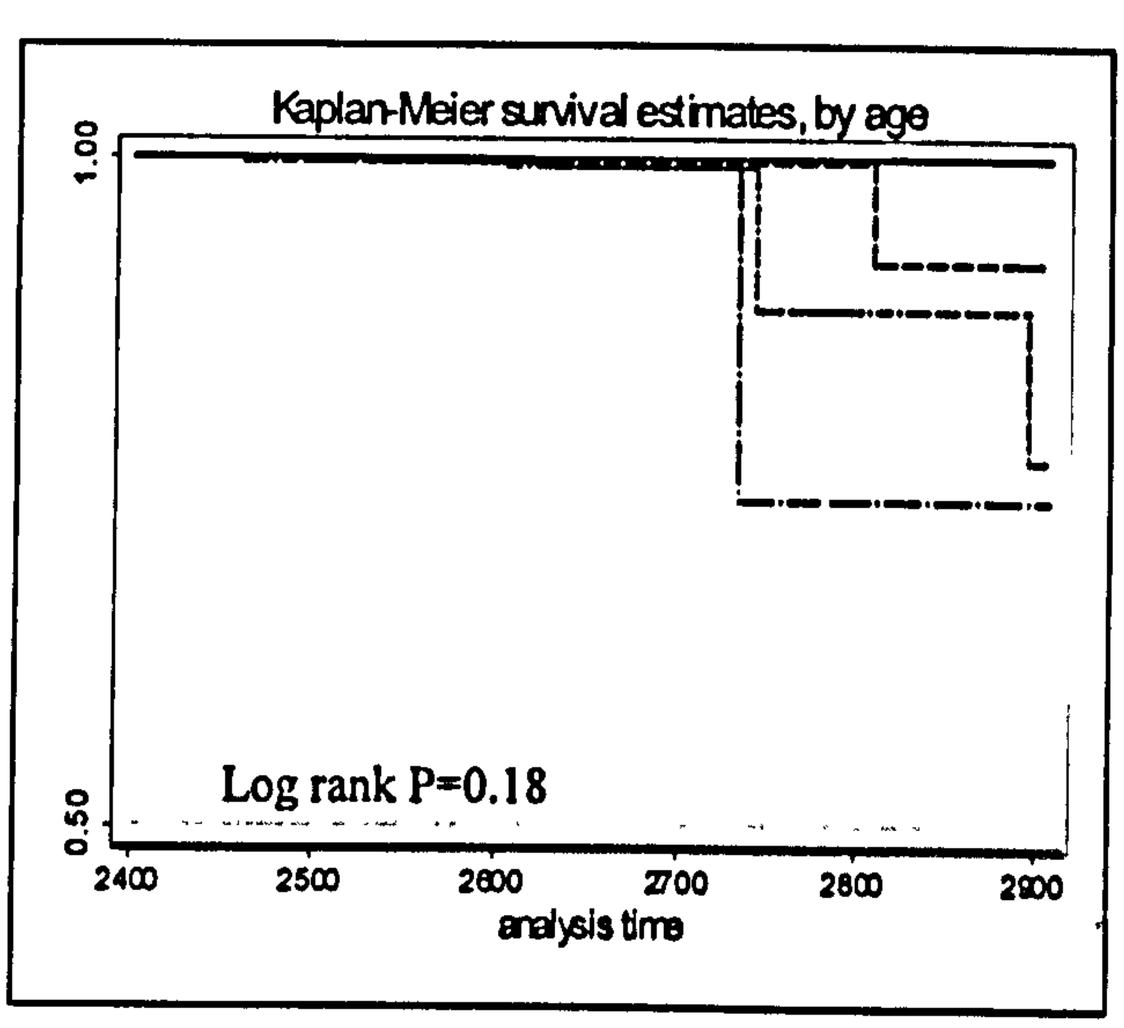


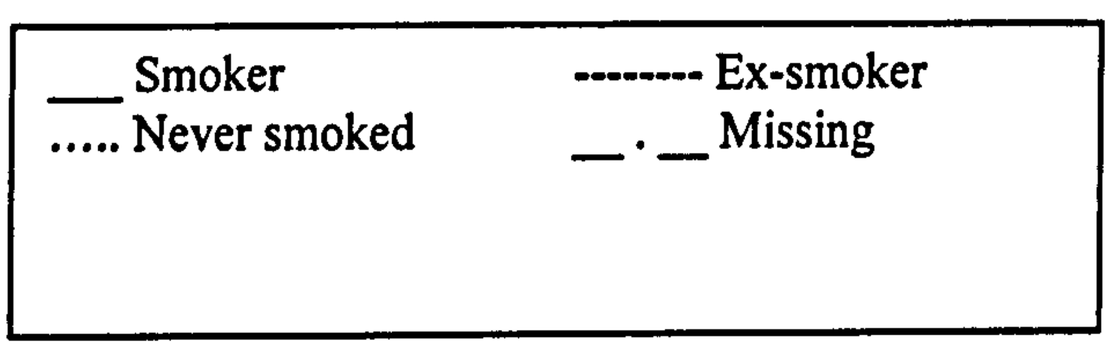












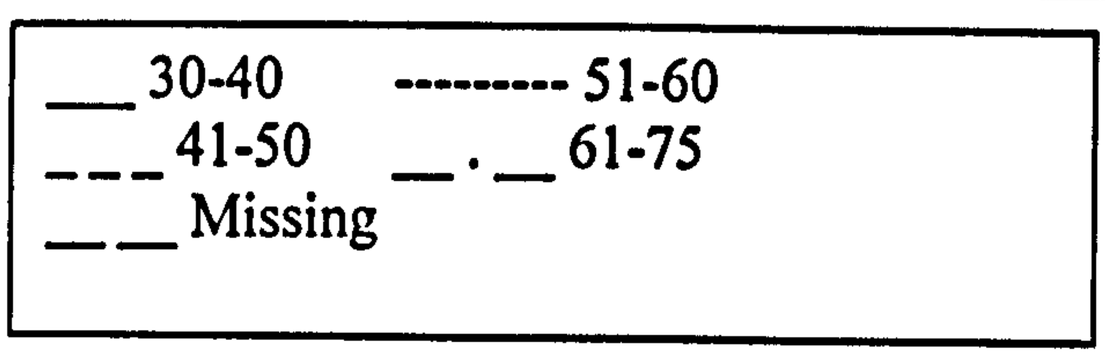


Figure 5.2: Difference in rates of getting cancer with and without recorded prognostic factors, alcohol, BMI, smoking and having children, tested using log rank method

5.5 Handling missing data by multiple imputation

The survival analysis model, like the logistic regression model applied in Chapter 4 suffered from missing data in all prognostic variables. Analysis of complete cases, accounted for the loss of 19% of the recorded observations in the Cox regression model, see Section 5.3.2. This was the result of most records having one or two missing values. In this Section multiple imputation was applied to fill in the missing values. The method of multiple imputation by chained equation described in Section 4.7, was applied. The same sequence was followed to generate imputed values, using starting values as the median for categorical variables, and looping through ten iterations. One of the advantages of the multiple imputation is that it allows the analyst to add in additional information in the imputation model. This information may not be of interest in the analysis model but can help in making the MAR assumption more plausible (Rubin, 1996; Liu *et al.* 2000). However, the UKWCS contains around 600 variables and a large percentage of this can be used to generate imputations. The computational complexity as well as the problem of multicolinearity makes it impossible to use all variables in the dataset. The increase in explained variance in linear regression also becomes small after the best set of 10 variables or slightly more are included in the imputation model.

A number of variables not in the analysis model were tested for the possibility of predicting missing data in the prognostic factors of the survival analysis model.

For each prognostic factor an imputation model was specified. For the binary variables, 'children' and 'menopausal status', logistic regression model was used and for the categorical variables 'alcohol', 'smoking', 'BMI' and 'age group', a

polytomous regression model was specified. The imputation model for each prognostic factor included all the variables in the survival analysis model together with the selected set of variables from the chosen auxiliary variables that were found to predict missing values, see Table 5.5. These imputation models were more general than those applied in Chapter 4, which only included all variables in the analysis model (logistic regression). A set of plausible values was then generated for each prognostic factor. This method of multiple imputation assumes that a multivariate distribution exists and that the Gibbs sampling (see Section 2.10.1 for definition of the Gibbs sampling) based on the specified model for continuous, binary and categorical variables can generate plausible values from it.

For example for 'alcohol' the imputation has a polytomous regression form with 'smoking', 'BMI', 'age', 'children', 'menopausal status', 'marital status', 'social class', 'higher educational level' and 'physical exercise' as predictors. On the other hand the imputation model for the variable 'children' has a logistic regression form with 'alcohol', 'BMI', 'smoking', 'menopausal status', 'age group', 'marital status', 'social class', 'higher educational level', 'physical exercise' and survival time as predictors.

Variable	Original dataset N(%)	Completed datasets 1 N(%)	2 N(%)	3 N(%)	4 N(%)	5 N(%)
Alcohol						
More than once a week	17,695(52.21)	18,101(52.20)	18,083(52.14)	18,072(52.11)	18,100(52.19)	18,075(52.12)
Once a week	4,856(14.33)	4,955(14.29)	4,972(14.34)	4,984(14.37)	4,963(14.31)	4,964(14.31)
Less than once a week	7,573(22.34)	7,762(22.38)	7,771(22.41)	7,756(22.37)	7,754(22.36)	7,753(22.36)
Never drink	3,771(11.13)	3,861(11.13)	3,853(11.11)	3,867(11.15)	3,862(11.14)	3,887(11.21)
BMI						
Underweight	730(2.19)	758(2.19)	755(2.18)	756(2.18)	762(2.20)	762(2.20)
Average	20,716(62.04)	21,505(62.01)	21,491(61.97)	21,514(62.04)	21,489(61.97)	21,505(62.01)
Overweight	8,605(25.77)	8,934(25.76)	8,955(25.82)	8,930(25.75)	8,943(25.79)	8,930(25.75)
Obese	3,340(10.00)	3,482(10.04)	3,478(10.03)	3,479(10.03)	3,485(10.05)	3,482(10.04)
Smoking						
Smoker	3,725(11.06)	3,840(11.07)	3,835(11.06)	3,851(11.10)	3,846(11.09)	3,851(11.10)
Used to smoke	10,434(30.99)	10,755(31.01)	10,745(30.98)	10,753(31.01)	10,744(30.98)	10,733(30.95)
Never smoked	19,513(57.95)	20,084(57.91)	20,099(57.96)	20,075(57.89)		20,095(57.95)
Children						
No	4,262(13.83)	4,952(14.28)	4,926(14.20)	4,902(14.14)	4,924(14.20)	4,947(14.27)
Yes	26,561(86.17)	29,727(85.72)	29,753(85.80)	29,777(85.86)	29,755(85.80)	29,732(85.73)
Menopausal status						
Pre-menopausal	15,326(44.20)	15,328(44.20)	15,327(44.20)	15,328(44.20)	15,326(44.19)	15,326(44.19)
Post-menopausal	19,351(55.80)	19,351(55.80)	19,352(55.80)	19,351(55.80)	19,353(55.81)	19,353(55.81)
Age group (years)						
30-40	2,801(8.12)	2,826(8.15)	2,818(8.13)	2,824(8.14)	2,823(8.14)	2,818(8.13)
41-50	13,056(37.83)	13,113(37.81)	13,112(37.81)	13,112(37.81)	13,116(37.82)	13,117(37.82)
51-60	10,496(30.42)	10,557(30.44)	10,558(30.44)	10,561(30.45)	10,559(30.45)	10,556(30.44)
61-75	8,155(23.63)	8,182(23.59)	8,190(23.62)	8,181(23.59)	8,179(23.58)	8,187(23.61)

Table 5.6: Frequencies and their percentages of prognostic factors in the original and the five imputed datasets

5.6 Multiple imputation analysis

Five completed datasets were generated. The frequency distributions of the completed datasets, were compared to those of the original dataset, see Table 5.6. This was mainly to assess the consistency of the imputed datasets with the original dataset. The Table shows that percentages of categories within prognostic factors in the five completed datasets were similar to the original dataset.

Cox models were fitted to the five completed sets of data. The five results were combined using the rules described in Section 2.8, equations 2.3 - 2.6.

Results of the Cox model on complete cases and the pooled analysis using five completed datasets are presented in Table 5.8.

The results after multiple imputation benefited from almost 6,500 additional records that had been excluded in the complete case analysis. The 39,000 values used in the analysis from 6,500 records and six variables consisted of 7,095 (18%) imputed values and 31,904 (82%) observed values, i.e. for each imputed value almost four observed values are added.

The scaled Schoenfeld residuals test was used to test the assumption of time independent proportional hazards. There was no evidence of violation of the proportional hazard assumption in the five fitted models on the completed datasets.

The first obvious change is the reduction in standard errors and narrower confidence limits in the pooled analysis of the completed datasets.

'Alcohol consumption', 'BMI', 'smoking', 'having children', and 'menopausal status' were not statistically significant prognostic factors in both models. Age group was the only significant prognostic factor in both models, with a greater risk associated with

older age. The risk was found to have steady increase in the pooled analysis model. An interesting finding was in the variable 'having children'. This variable was not significant in the single variable analysis using all available cases (30,823 (89%)). However, it showed significant effect in the complete case analysis when only complete cases 28,166(81%) where analysed due to missing data. Women with no children were at greater risk of getting breast cancer (hazard ratio= 1.39, 95% C.I. 1.06−1.82). Results from the complete case analysis were more rational as not having children is a well-known risk factor for getting breast cancer (Key *et al.* 2001; Ramon *et al.* 1996; Ewertz et *al.* 1990; Layde *et al.* 1989). Results from the complete case analysis supports a study by Sweeney *et al.* (2004) which showed that having children had a protective effect on getting breast cancer (hazard ratio=0.67, 95% C.I. 0.51−0.88). The same variable after multiple imputation had no significant effect on women getting the breast cancer. This shows that this specific factor was more accurate in the complete case analysis and probably the variable of having children had an exaggerated effect in this dataset.

5.7 Sensitivity analysis

The choice of predictors in the imputation models was based on practical considerations. For each incomplete variable, the predictors for missingness were selected as all the variables included in the analysis model together with variables, which were found to predict missingness for that specific incomplete variable. However, there was no straightforward procedure to assess if the choice was

appropriate, i.e. if the list of variables selected were enough for the MAR assumption to hold. Using most of the information in the cohort can lead to multiple imputation with minimal bias and maximal certainty. The cohort is made up of more than six hundred variables all of which can be used to generate imputations. However, including all these variables is not computationally feasible.

The set of predictors included in the imputation model for alcohol consumption was altered to test sensitivity of the results to alterations in this imputation model. The dataset was checked for variables that were correlated with alcohol consumption. As all variables were incomplete, the level of missing data was also considered. The alcohol consumption variable was found to have the following level of association with the four types of alcohol recorded in the dataset. These associations were calculated from all available data. The highest level of association was found with 'wine' and this variable was found to have the least percentage of missing values among the four types of alcohol.

	Correlation coefficient	**(%) of missing values**
Beer	O.30	52.29
Wine	0.61	18.15
Sherry	0.36	51.56
Spirits	0.37	43.66

Plausible values were regenerated using the same procedure described in Section 5.4.2, and the same set of predictors presented in Table 5.5. However, an additional predictor 'wine' was added to the imputation model for the 'alcohol' variable. Sensitivity of the multiple imputation estimates of the survival analysis model 'including wine' and 'without wine' was examined.

Results of the of the Cox proportional hazard model are presented in Table 5.7. There were no real differences between the multiple imputation estimates or standard errors using the imputation models 'without wine' and 'including wine'. Adding wine to the imputation model adds little extra information about alcohol consumption, however the estimates of alcohol consumption were not affected by using this more comprehensive information. These results suggests that the inclusion of predictors for each imputation model should be as large as possible to make the MAR assumption feasible and thus reducing the need to make special adjustments for mechanisms that are not MAR. The selection should also be capable of testing the instability of the imputation models resulting from multicolinearity when too many predictors are included The imputation model should be capable of reflecting the structure of the dataset and association between variables, thus the selection of predictor variables for imputation models can proceed as follows:-

1. Include all the variables in the analysis model.
2. Include all variables that were found to have different distribution between the observed and missing data. This can be achieved by fitting single variable logistic regression models (missing, not missing) for the variable to be imputed and test if missingness is predicted by other variables in the dataset (see Section, 5.4.2).
3. Include variables that are found to have strong correlation with the imputed variables.
4. From steps, two and three drop variables with high levels of missing values so that the final set of predictors are around 10-15 variables.
5. Check the independence of predictor variables to reduce the instability that can be caused by multicolinearity in imputation models.

Variable	Multiple imputation with wine N=34,666 Breast cancer cases= 474			Multiple imputation without wine N=34,666 Breast cancer cases= 474		
	Hazard ratio	Std. Err.	95% C.I.	Hazard ratio	Std. Err.	95% C.I.
Alcohol						
Never drink	1.00			1.00		
More than once a week	1.03	0.16	0.76-1.40	1.04	0.16	0.77-1.41
Once a week	1.01	0.19	0.70-1.46	1.02	0.19	0.71-1.48
Less than once a week	1.21	0.20	0.88-1.68	1.23	0.20	0.89-1.70
BMI						
Average	1.00			1.00		
Underweight	0.70	0.27	0.33-1.47	0.68	0.26	0.33-1.44
Overweight	0.91	0.10	0.73-1.14	0.92	0.11	0.74-1.15
Obese	1.17	0.18	0.87-1.57	1.17	0.18	0.88-1.57
Smoking						
Smoker	1.00			1.00		
Used to smoke	0.89	0.14	0.65-1.22	0.89	0.14	0.65-1.22
Never smoked	0.95	0.10	0.78-1.17	0.97	0.10	0.79-1.19
Children						
Yes	1.00			1.00		
No	1.21	0.17	0.92-1.59	1.19	0.15	0.92-1.51
Menopausal status						
Pre-menopausal	1.00			1.00		
Post-menopausal	1.03	0.13	0.81-1.32	1.03	0.13	0.81-1.31
Age						
30-40	1.00			1.00		
41-50	1.83	0.43	1.16-2.89	1.83	0.42	1.16-2.88
51-60	2.23	0.55	1.37-3.63	2.22	0.55	1.36-3.60
61-75	2.36	0.61	1.42-3.90	2.35	0.60	1.42-3.88

Table 5.7: Hazard ratios, standard errors and 95% C.I. of survival analysis in multiple imputation 'including wine' and 'without wine' in the imputation model.

5.8 Hotdeck using STATA

In hotdeck imputation, the missing case is substituted by identifying the most similar case to the missing and imputing its value. In more sophisticated hotdeck applications, a set of similar donors is identified; the missing item is then substituted by drawing at random from this pool of similar donors or by averaging them. Hotdeck was discussed in detail in Section 2.7.2.

STATA 8 has implemented a hotdeck procedure by tabulating the missing data pattern in a list of variables. A missing line is defined as a record from the list of variables with a missing value in any of its variables; a complete line is a row of the list of variables where all the data are observed. The hotdeck will then replace the variables in the missing line with the corresponding values from the chosen complete line. Hotdeck can be used several times within a multiple imputation procedure. The variable with missing values in each stratum of the data described is replaced by values sampled from variables with complete records in the same stratum, to add variability. A bootstrap sample of complete records is sampled with replacement from the observed values, and the records with missing values are sampled at random (again with replacement) from this bootstrap sample. This procedure can be repeated several times to have a set of plausible values for each missing value.

Hotdeck imputation was applied to five of the prognostic factors, 'alcohol consumption', 'BMI', 'smoking', 'children' and 'age group'. The variable 'menopausal status' was not imputed as it had two missing values only. Each variable was imputed by conditioning on the most complete variable 'menopausal status'. This decision was based on Table 5.5, which shows that 'menopausal status' predicts

missing values for 'alcohol consumption', 'smoking', 'children' and 'age'. The variable 'BMI' was conditioned on 'age', as 'age' consisted of 0.5% missing values and it predicted missingness in 'BMI'. The decision to conditioning on one variable was reached as it was found that the more variable one conditions on the less values would be imputed. To illustrate this, let us take the imputation of the variable 'alcohol consumption' as an example. The pattern of missing values for the variables 'alcohol', 'menopausal status', 'smoking' and 'children' were as shown below

Alcohol	Menopause	Smoking	Children	Frequency
X	X	X	X	29,328
X	X	X	-	3,642
X	X	-	X	812
-	X	X	X	616
X	X	-	-	110
-	X	X	-	83
-	X	-	X	66
-	X	-	-	19
X	-	X	-	2

X=observed value - =missing value

When 'alcohol' was conditioned on menopausal status, 782 values were imputed for 'alcohol'. This was the result of imputing 616, 83, 66 and 19 from the patterns with missing 'alcohol' and observed 'menopausal status', but 2 values with observed 'alcohol' and missing 'menopausal status' were dropped as the result of this type of imputation, the final alcohol variable had 34,677 observations. Values were next imputed for 'alcohol' using hotdeck and conditioning on the variables 'menopausal status' and 'smoking'. Alcohol variables had 33,670 observations after this imputation. This time hotdeck imputed 616 + 73 values for 'alcohol' from patterns with missing 'alcohol' and observed 'menopausal status' and 'smoking', the imputation procedure deleted 110 records with observed 'alcohol' and 'menopausal

status' and missing 'smoking', as well as 2 records with observed 'alcohol' and 'smoking' and missing 'menopausal status'. Following this investigation it was decided that each variable should be imputed on one variable only, to reduce the amount of observed data being lost and to be able to impute as many observations as possible.

As mentioned above, hotdeck is capable of imputing values, within a multiple imputation procedure, by specifying a stratum of donors for each missing value and each time the routine chooses imputed values from the specified stratum in a random manner. To reduce the underestimation of variability of the data, the hotdeck imputation was repeated 5 times and the results were combined using Rubin's rules (equations 2.3-2.6).

Five imputed datasets were generated; survival analysis was applied to each imputed dataset. The pooled results following the hotdeck imputation are presented in Table 5.8

The number of observations in the survival analysis was 34,503 i.e. 6,338 additional records were used when compared to the complete case analysis. These additional records consist of 6,927 imputed values and 38,022 observed values, i.e. the imputed values were only 18% of the gained records. This additional information resulted in smaller standard deviations as well as narrower confidence intervals in the pooled analysis of the survival analysis model after the imputation by hotdeck.

The model shows that 'age' was the only significant prognostic factor of women getting breast cancer. The risk of getting the cancer steadily increases with older age. Hazard ratios for all other factors were slightly different from the results following multiple imputation. This similarity arises from the fact that most of the added

information was actually observed values and the imputed values were less than 25% of the gained 6, 338 records.

Variable	Complete case analysis N=28,166 Breast cancer cases= 368			Multiple imputation N=34,666 Breast cancer cases= 474			Hotdeck analysis N=34,503		
	Hazard ratio	Std. Err.	95% C.I.	Hazard ratio	Std. Err.	95% C.I.	Hazard ratio	Std. Err.	95% C.I.
Alcohol									
Never drink	1.00			1.00			1.00		
More than once a week	1.06	0.19	0.74-1.52	1.04	0.16	0.77-1.41	1.03	0.16	0.77-1.40
Once a week	1.04	0.23	0.68-1.61	1.02	0.19	0.71-1.48	1.03	0.19	0.71-1.48
Less than once a week	1.43	0.27	0.98-2.08	1.23	0.20	0.89-1.70	1.21	0.20	0.88-1.68
BMI									
Average	1.00			1.00			1.00		
Underweight	0.39	0.23	0.12-1.21	0.68	0.26	0.33-1.44	0.73	0.28	0.35-1.49
Overweight	0.88	0.11	0.69-1.13	0.92	0.11	0.74-1.15	0.93	0.10	0.75-1.15
Obese	1.06	0.18	0.76-1.49	1.17	0.18	0.88-1.57	1.20	0.18	0.90-1.59
Smoking									
Smoker	1.00			1.00			1.00		
Used to smoke	0.91	0.17	0.64-1.61	0.89	0.14	0.65-1.22	0.90	0.14	0.65-1.22
Never smoked	1.08	0.12	0.76-1.49	0.97	0.10	0.79-1.19	0.97	0.10	0.79-1.18
Children									
Yes	1.00			1.00			1.00		
No	1.39	0.19	1.06-1.82	1.19	0.15	0.92-1.51	1.23	0.17	0.96-1.58
Menopausal status									
Pre-menopausal	1.00			1.00			1.00		
Post-menopausal	0.98	0.14	0.75-1.29	1.03	0.13	0.81-1.31	1.03	0.13	0.81-1.31
Age									
30-40	1.00			1.00			1.00		
41-50	1.94	0.53	1.13-3.32	1.83	0.42	1.16-2.88	1.83	0.42	1.16-2.89
51-60	2.55	0.73	1.45-4.50	2.22	0.55	1.36-3.60	2.23	0.55	1.37-3.63
61-75	2.46	0.74	1.36-4.45	2.35	0.60	1.42-3.88	2.36	0.61	1.42-3.90

Table 5.8: Hazard ratios, standard errors and 95% C.I. of survival analysis in complete case, multiple imputation and hotdeck analysis.

5.10 Discussion

A problem in survival analysis occurs when data are missing in one or more covariates. The aim of this chapter was to assess the effect of some factors in the incidence of breast cancer for the women who took part in the UKWCS, and to assess the implications of different methods of handling missing data compared to the analysis of the complete cases.

It was found more reasonable to divide some variables into categories, 'BMI' which was computed from two variables 'weight' and 'height' was divided into four categories. 'Age' was also divided into four categories. A Cox proportional hazard model with six prognostic factors using complete case analysis used 28,166 (82%) records, as a result of missing data in every variable.

Hotdeck imputation, in which missing values of the incomplete observations were replaced by some actual values from a similar set of observations in the data, had a long history of use. STATA 8 improved the hotdeck method to a random hotdeck in which for each missing value a donor is selected at random from a stratum sharing common features with the incomplete record. To improve the uncertainity due to missing data, hotdeck was further applied in a multiple imputation routine, i.e. instead of having a single imputed value selected at random from the donor stratum, five versions were generated which was later combined using the multiple imputation combining rules for the estimates. There were two main deficiencies in this implementation: -

- Hotdeck replaces a record with missing data with the complete record chosen at random from a specified stratum. Replacement of an observed value from an incomplete record with another observed value from the

selected complete record is not acceptable. This can lead to alterations in the dataset, which can lead to misleading results.

- This technique is not suitable for a dataset with an arbitrary pattern of missing data, as it was shown earlier in Section 5.7. When an incomplete variable is conditioned on another variable, which was also not complete to generate imputed values, observed values from the record to be imputed can be deleted.

The method of multiple imputation by chained equation was by far superior to the complete case analysis and the hotdeck. Imputation models for each variable involved two modelling choices: -

- The form of the model (logistic for binary variables, regression for continuous and polytomous regression for categorical variable)
- A set of predictors for each imputation model. This included all the prognostic factors of the survival analysis model together with other exogenous variables that were found to predict missing data for that specific variable to be imputed.

This technique was capable of generating plausible values by conditioning on a set of variables that were found to predict missingness. The method was also capable of making use of predictors of missing data even if these predictors were not complete. I presented in Chapter 4 that 6 iterations were enough to reach convergence, however ten iterations were used to generate the plausible values. Computing time needed for these iterations was reasonably fast.

Because many studies are expected to be carried out from this huge dataset, and as the aim of the thesis was to find a universal solution of missing data that can be used by analysts who have no background in handling missing data, multiple

imputation was found to be the most reasonable solution. The final model after imputation included more than 34,000 records. The final results, after imputation, can be reported in exactly the same format as the complete case analysis, with no more effort than running the analysis 5 times and then combining the results. The same computational routines can be used for different variables. The Chapter discussed the selection of predictor variables for imputation models, five steps were developed to help in this choice. Although utilization of all the information in the dataset can help in the generation of multiple imputation that have the least bias, this was found not feasible computationally and might lead to multicolinearity problems. A compromise is to select a suitable set of predictors for each specific imputation model that contains between 10-15 predictors.

The imputed datasets developed in this chapter could be readily adapted to use in subsequent analysis of diet and survival to incidence of cancer when information on incidence of cancer is complete. This analysis, can now also take into account the effect of missing data, and substantially enhance the analysis and robustness of conclusions in this important cohort.

Chapter 6

Assessing the missingness mechanism using a repeated FFQ

6.1 Introduction

In the UKWCS a repeated FFQ, identical to the FFQ used in Phase 1 was mailed mainly to find out to what extent people change their diet. A baseline data collection was mailed to a sample of 2,200 women five years after the first questionnaire was received. A decision was made to mail the repeat FFQ only to the "ideal responders". These women had replied to the UKWCS Phase 1, 2 and 3 mailing by July 2000. This was to ensure a greater chance of achieving the target response rate of 60% from the repeat FFQ, and to form the basis of assessing the consistency of response, as well as testing repeatability of the FFQ. A total of 1,918(87%) subjects responded to the repeated questionnaire, and the responses from this sample will be used for our analysis in this chapter coupled with their responses to the baseline questionnaire.

The three phases of the study were as follows: -

- Phase 1 was the baseline questionnaire for which 35,374 responded
- Phase 2 was the questionnaire about the use of supplements, which followed the baseline questionnaire (this information has not been used in this thesis)

- Phase 3 was the food diaries ;(which to date have not been coded so has not been used in this thesis)

see Section 1.4 for a full description.

In this chapter, the repeated FFQ will be used to find out information about the missing data in the Phase 1 questionnaires. The repeated FFQ provides a unique opportunity to explore the missing data mechanism. There was a gap of about five years between the two sets of questionnaires, which made the task more difficult, as women might have changed their eating habits during this time.

However, given the age range of the women taking part in the cohort, 35-69 years at the start of the cohort and 40-74 years when completing the questionnaires, one would expect that the eating habits, patterns and preferences of the majority of these women were well established. In addition, the gap increases the chance that the reason for any missing data in Phase 1 will have been forgotten, and the item may have been answered the second time round.

To illustrate the usefulness of this information this chapter covers a comparison of consumption of bread in the baseline questionnaire, compared to the repeated questionnaire; this information was collected in the FFQ form.

To assess whether the responses were more consistent when long questions were used instead of the FFQ, a comparison of responses to alcohol consumption and smoking was made between baseline and repeated questionnaires.

In Section 2.2 we discussed how the relationship between the missing data mechanism and the missing and observed values reflects the basis for most types of missing data, and that the most appropriate way to handle missing or

incomplete data will depend upon how data points became missing (De Leeuw *et al.*, 2001).

The goal of this chapter was to study the repeat FFQ responses to get as much information as possible about the non-response mechanism and consistency of the subjects' responses.

6.2 Block I: Bread/Savoury Biscuits

To explore the relation between missing data in the FFQ and its repeated version, I focus on a small number of FFQ items to illustrate the method. Two types of bread from the first block of nine questions of the FFQ, Bread/ Savoury Biscuits were used. This block was chosen because bread is one of the basic food items and is consumed by a high percentage of the population. Therefore, one can assume that respondents would be consistent in the consumption of bread even after five years. The 1,918 repeated FFQ were matched by id-number to the responses on the original set of FFQ of Phase 1. For the purpose of illustration, the first two variables 'white bread' and 'brown bread' were investigated in detail. The tabulation of the first variable 'white bread' against its repeated FFQ responses was given in Table 6.1. One would have expected a lot of agreement, with high figures on the diagonal, that is, many women would still be consuming the same type and amount of bread after 5 years, but that was not the case. Instead, the Table shows that high frequencies were consentrated in the upper left diagonal corner of the Table. Women who first reported never eating White bread gave a range of answers in the repeated questionnaire. Some even reported eating

White bread as frequently as 2-3 times per day. However, mostly women have reduced their consumption of White bread. The repeated questionnaires were filled by a highly selected sample of women who were willing to take part and who completed Phase 1 and Phase 2 properly. Therefore, one would expect them to be the type of cautious women who would change to a healthier diet, i.e. reduce their consumption of White bread. However, most of the women who took part in the cohort were in fact more educated and tended to lead a healthier life style than average. This change in diet could reflect the growing popularity of brown & wholemeal breads in the general population over that period. It should also be noted that the number of missing items for 'white bread' at both time points was very small, which was also shown on Table 6.1.

Original FFQ	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS ?										
	Never	Less than once a month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once per day	2-3 per day	4-5 per day	6+ per day	M
Repeated FFQ											
0	17	40	33	25	55	16	23	34	10	1	0
1	40	85	50	63	79	26	50	50	11	3	0
2	30	42	48	41	45	13	24	35	5	0	0
3	24	52	31	26	37	8	25	23	4	0	0
4	33	48	47	34	51	18	35	36	8	1	0
5	7	14	10	10	12	4	5	7	1	0	0
6	5	22	18	18	22	5	10	20	3	0	1
7	16	28	16	17	27	7	11	17	2	0	1
8	2	6	2	4	6	0	2	3	1	0	0
9	0	0	2	0	0	0	0	1	0	0	0
M	1	8	3	2	0	0	1	2	1	0	0

Table 6.1: Distribution of responses to the first block - White bread and rolls: Phase 1 against the repeated FFQ.

Notes: The frequencies are M = missing values; 0=Never; 1=less than once a month; 2=1-3 times per month; 3=Once a week; 4=2-4 times per week; 5=5-6 times per week; 6=once a day; 7=2-3 times per day; 8=4-5 times per day; 9= 6+ times per day.

White bread original vs. repeated questionnaire	Agreement	Differ by 1 code	Differ by 2 codes	Differ by >2 codes	Missing
Frequency	259	391	365	883	20
Percentage	13.5%	20.4%	19.0%	46.0%	1.0%

Table 6.2: Difference in responses to 'white bread' in original vs. repeated questionnaire. The median intake in the original questionnaire=3 'Once a week'. The median intake in the repeated questionnaire= 2 '1-3 times per month'.

Table 6.2 shows that there was an exact agreement in only 13.5% of the responses to 'white bread' between the original and repeated questionnaire, 20.4% of the responses differed by 1 code, 19.03% of the responses differed by 2 codes and around 46% of the responses had a difference of more than two codes. This amounts to substantial inconsistency between the two questionnaires. The median intake of white bread in the original questionnaire was the code 3 'once a week', while the median intake of white bread in the repeated questionnaire was the code 3 '1-3 times per month'. This finding shows that the sample which filled in the repeated questionnaire consumed smaller amounts of white bread.

Frequencies of consumption of 'brown bread' are presented in Table 6.3. The Table shows that responses have changed considerably within the five years, less consumption of 'brown bread' could be seen, and the decrease in consumption of 'brown bread' was mainly among those who ate 'brown bread' 2-3 times per day or less frequently. Surprisingly, the recorded missing went up to 189 almost 10% of the number taking part in the repeated questionnaire. Women who skipped this question in the repeated questionnaire reported a wide range of answers in the original questionnaire. Of those who left 'brown bread' missing in the repeated questionnaire, 20% reported never eating 'brown bread' in the original questionnaire. 12% reported eating 'brown bread' less than once a month, 16% reported eating 'brown bread' 1-3 times per month, 13% reported eating 'brown bread' once a week, 13% reported eating 'brown bread' 2-4 times per week, 6% reported eating Brown bread 5-6 times per week, 10% reported once per week,

7% reported 2-3 times per day and no one from those who left 'brown bread' missing in the repeated questionnaire left it missing in the original questionnaire. There was no evidence from this tabulation that the missing values were intentional.

Original FFQ	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS?										
	Never	Less than once a month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once per day	2-3 per day	4-5 per day	6+ per day	M
Repeated FFQ											
0	63	55	46	36	46	18	21	22	4	0	1
1	82	55	62	45	69	22	43	37	3	1	0
2	45	35	36	23	39	15	30	21	5	0	0
3	36	30	22	24	24	10	14	16	2	0	0
4	58	37	35	24	35	15	23	23	2	0	0
5	13	16	11	8	7	6	7	10	2	0	0
6	17	12	17	21	16	3	13	12	0	0	0
7	14	18	15	8	19	4	16	16	1	1	0
8	6	3	3	0	1	0	1	1	0	0	0
9	0	1	0	0	0	0	0	0	0	0	0
M	38	23	31	24	24	12	19	14	3	1	0

Table 6.3: Distribution of responses to the first block - Brown bread and rolls: Phase 1 against the repeated FFQ.

Notes: The frequencies are M = missing values; 0=Never; 1=less than once a month; 2=1-3 times per month; 3=Once a week; 4=2-4 times per week; 5=5-6 times per week; 6=once a day; 7=2-3 times per day; 8=4-5 times per day; 9= 6+ times per day.

Table 6.4 shows that for 'brown bread', there was exact agreement for 12.9% responses, difference by one code in 20.3% of the responses, difference of 2 codes in 16.3% of the responses and difference of more than 2 codes in 40.5% of the responses. The median intake of brown bread in the original questionnaire was the

code 3 'once a week', while the median intake of brown bread in the repeated questionnaire was the code 2 '1-3 times per month'. This finding shows that the consumption of brown bread decreased in the sample that filled in the repeated questionnaire.

Each identical pair of bread type was tested for significant difference in responses using the Wilcoxon signed-rank test, a non-parametric test that compares two paired groups. The test makes no assumption about the distribution of the data. Table 6.5 shows significant difference between original and repeated questionnaire in the intake of 'white bread', 'brown bread', 'wholemeal bread', 'chapatis' and 'papudum'. The median intake for 'white bread' and 'brown bread', changed from "Once a week" to "1-3 times per month", on the other hand there was an increase in the consumption of 'wholemeal bread' in which the median intake changed from "2-4 times per week" to "5-6 times per week". The table also shows significant difference between original and repeated questionnaire in the consumption of 'Chapatis', in which the median intake changed from "Never" to "less than once a month", while the median intake remained unchanged for 'papudum'. However, there was no significant difference between the original and repeated questionnaire in the intake of 'pitta bread' and 'crispbread'. This test similar to the cross-tabulation presented in Tables 6.2, 6.3 shows variation in reported consumption of bread in the original and repeated questionnaires. However, this does not fully assess agreement. A test of agreement is covered in Section 6.4

Brown bread original vs. repeated questionnaire	Agreement	Differ by 1 code	Differ By 2 codes	Differ By >2 codes	Missing
Frequency	248	389	313	778	190
Percentage	12.9%	20.3%	16.3%	40.6%	9.9%

Table 6.4: Difference in responses to 'brown bread' in original vs. repeated questionnaire. Median intake in the original questionnaire=3 'Once a week'. Median intake in the repeated questionnaire= 2 '1-3 times per month'

Variable	p-value	Median in Original questionnaire	Median in Repeated questionnaire
White bread	<0.001	3	2
Brown bread	<0.001	3	2
Wholemeal bread	<0.001	4	5
Chapatis	<0.001	0	1
Papudum	0.007	1	1
Pitta	0.554	1	1
Crispbread	0.780	1	1

Table 6.5: Test of significant difference of consumption of bread using Wilcoxon signed rank test.

Notes: The frequencies are 0=Never; 1=less than once a month; 2=1-3 times per month; 3=Once a week; 4=2-4 times per week; 5=5-6 times per week; 6=once a day; 7=2-3 times per day; 8=4-5 times per day; 9= 6+ times per day.

6.3 Plotting data

To assess the extent of discrepancies in the responses to the bread blocks between the two phases of FFQ, scatter plots of the responses of subjects to Phase 1 and the repeated questionnaire were produced. Spherical random noise was added to the points in the graph to keep the categorical data from over-plotting, and make it easier to see plots accumulating at the same point.

In Figure 6.1 'white bread' shows no clear association of responses between the original and the repeated questionnaire. Responses look more consistent and tend to condense in lower consumption codes (0-4). Similar scattered responses with no clear trend, was observed for codes between 0-8 in Figure 6.2 scatterplot of responses to 'brown bread' and Figure 6.3 responses of 'wholemeal bread'. For the less popular types of bread, 'chapati', Figure 6.4, 'papudum', Figure 6.5, 'tortilla', Figure 6.6, 'pitta bread', Figure 6.7 and 'crispbread', Figure 6.8 responses were concentrated at low frequencies of consumption 0-3. This was another evidence of inconsistency of responses, all the nine plots showed poor agreement between responses to the original and repeated questionnaires.

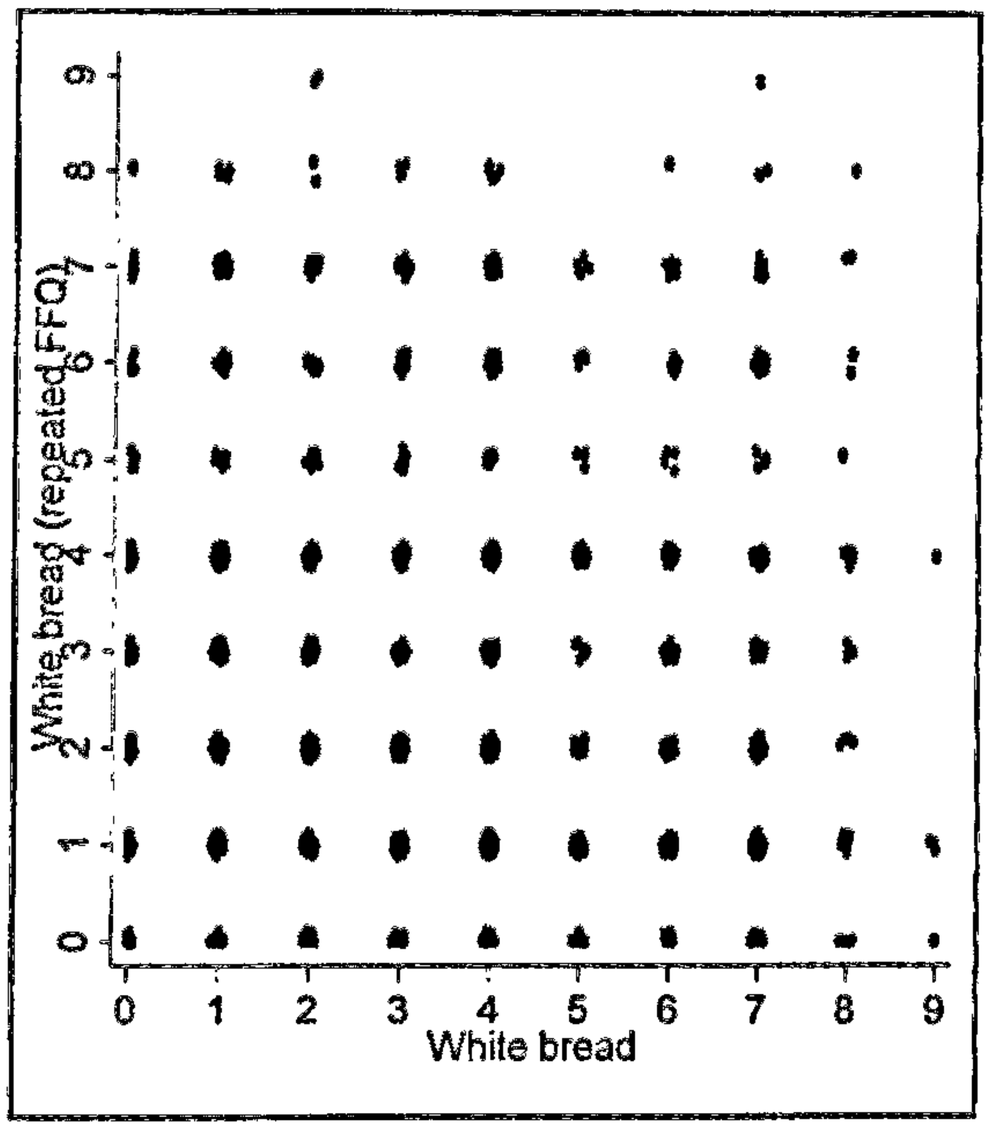


Figure 6.1: Scatter plot of White bread in original questionnaire vs repeated questionnaire

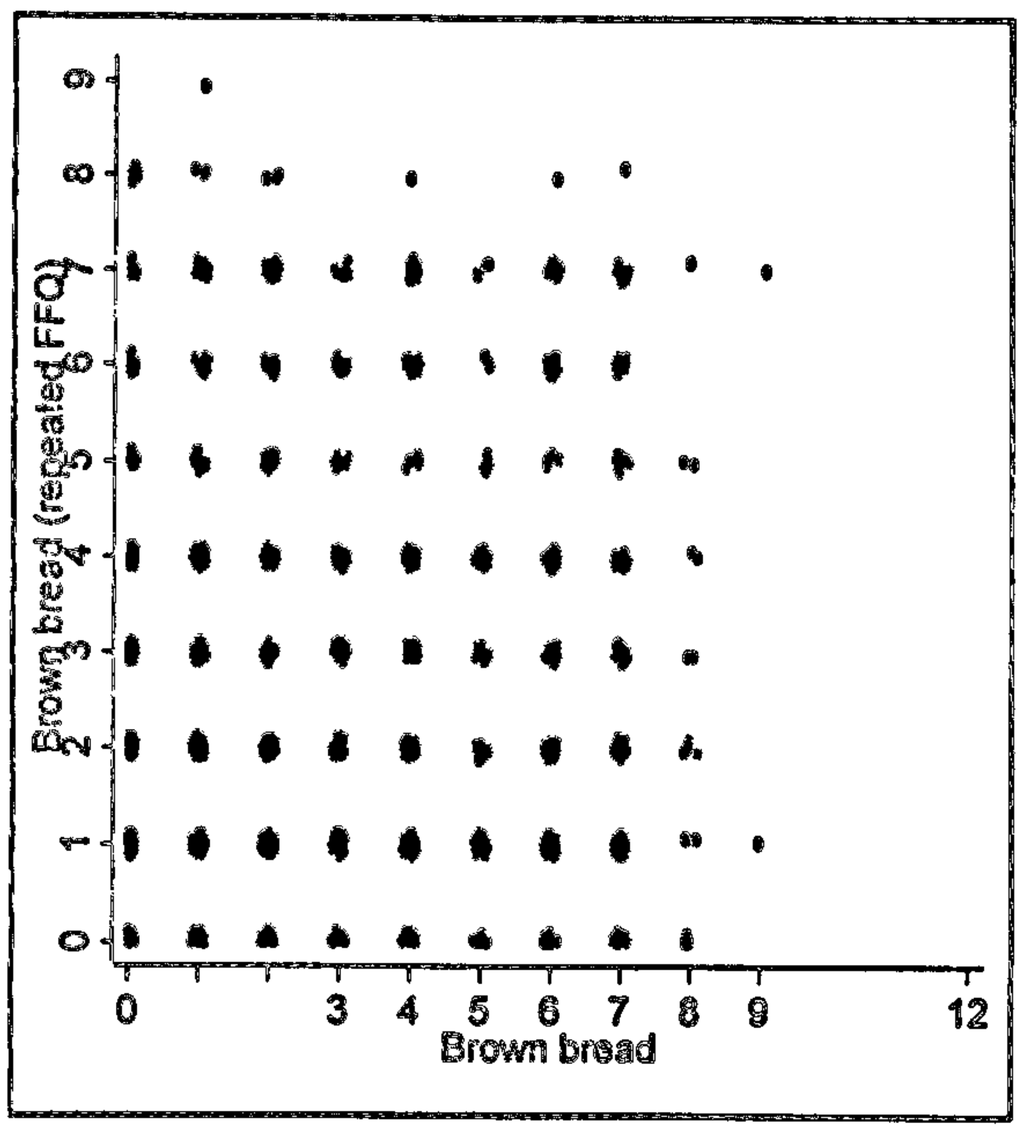


Figure 6.2: Scatter plot of Brown bread in original questionnaire vs repeated questionnaire

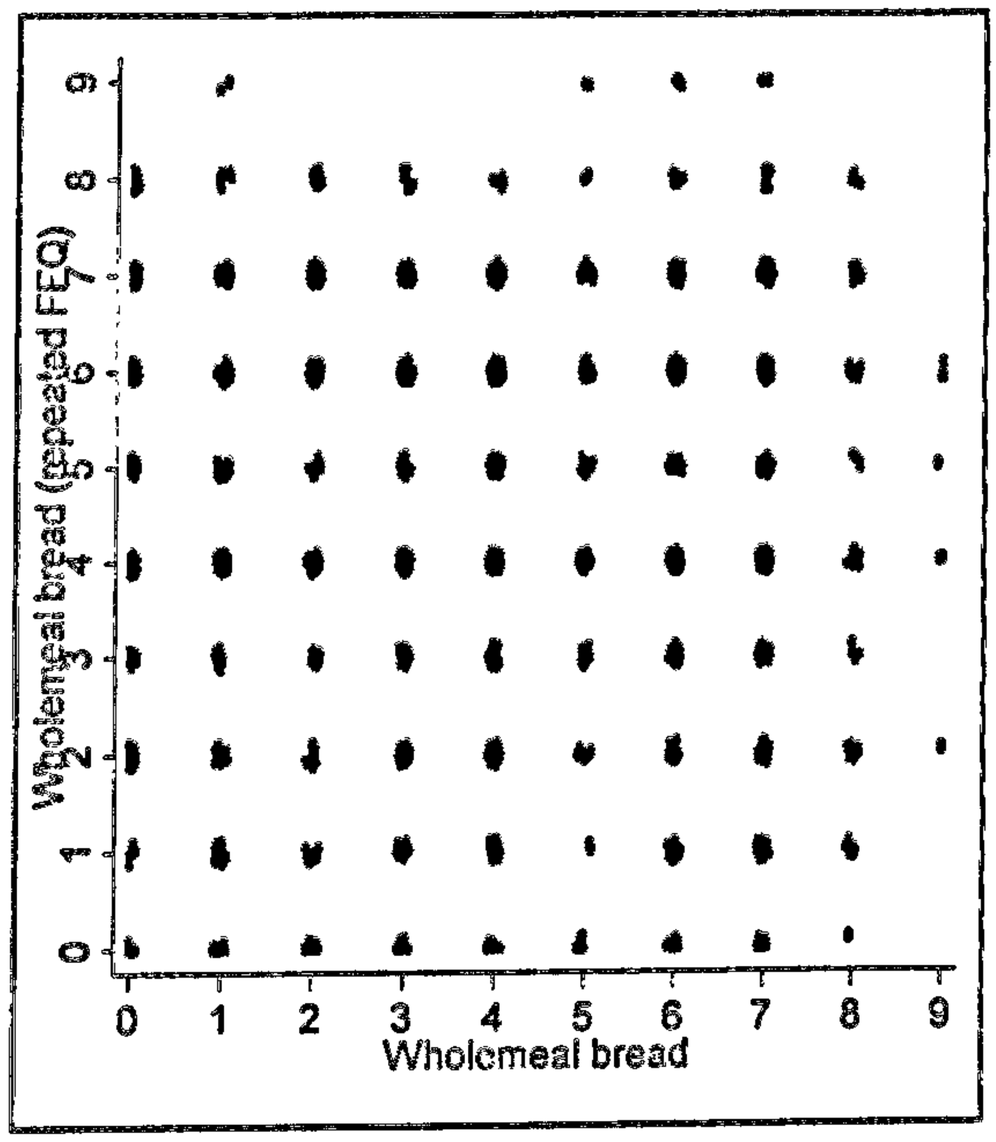


Figure 6.3: Scatter plot of Wholemeal bread in original questionnaire vs repeated questionnaire.

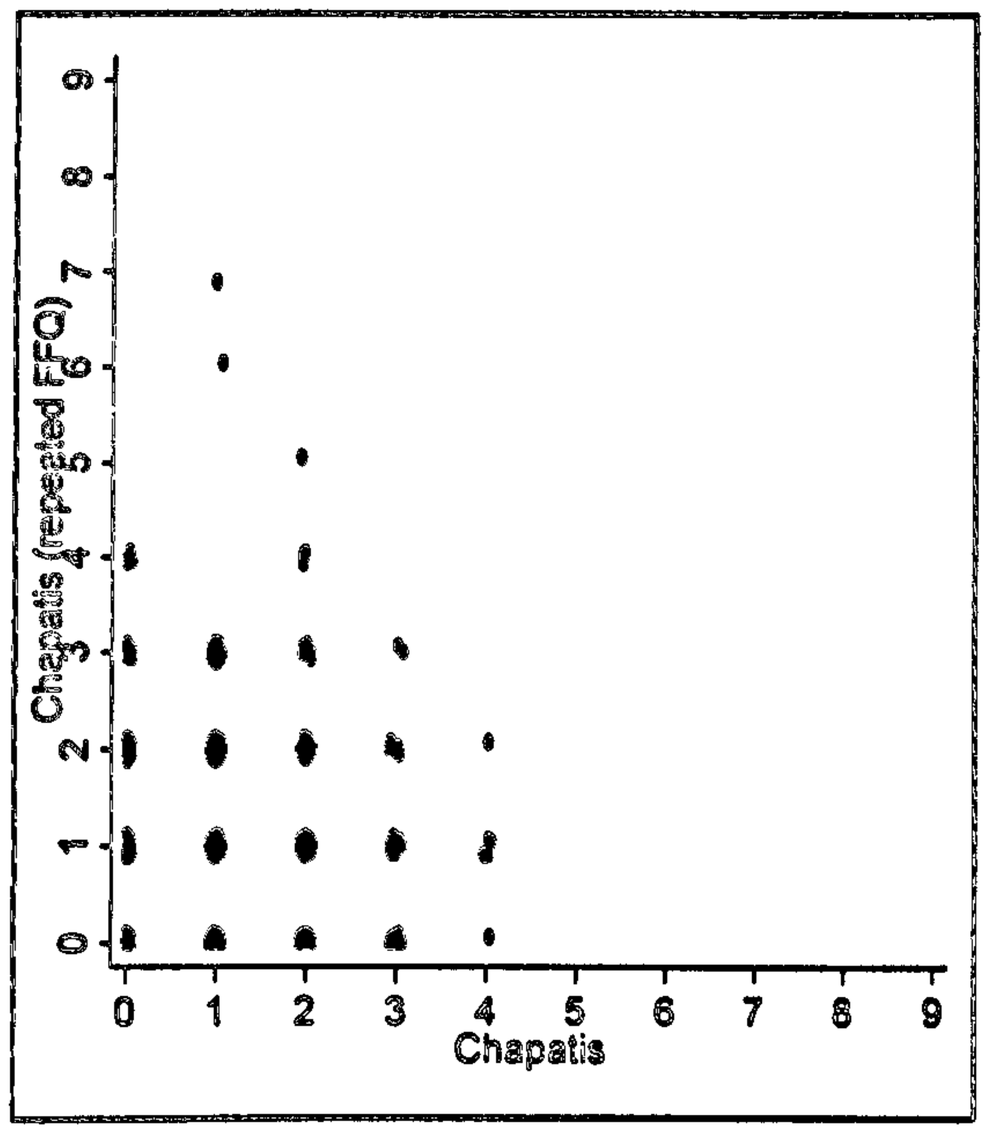


Figure 6.4: Scatter plot of Chapatis in original questionnaire vs repeated questionnaire.

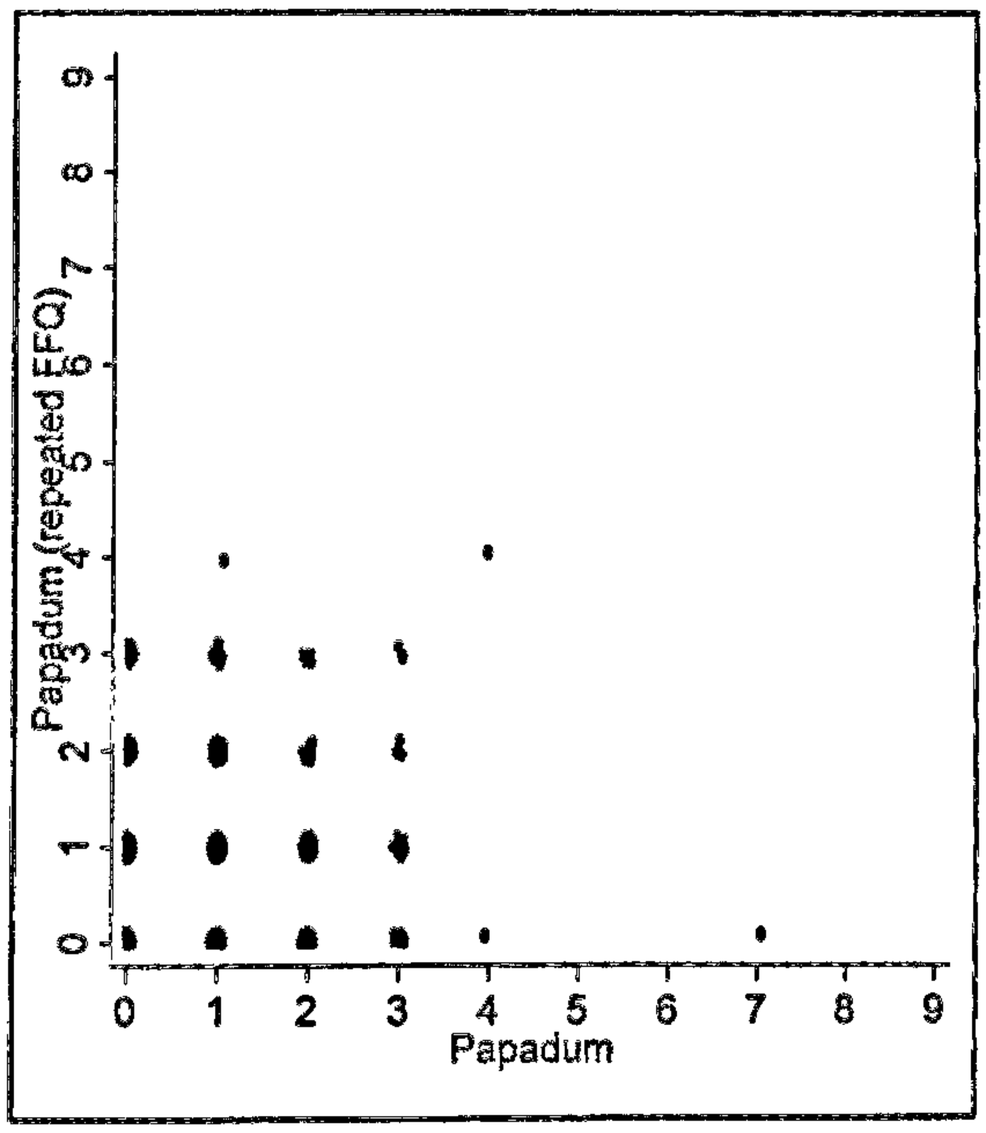


Figure 6.5: Scatter plot of Papadum in original questionnaire vs repeated questionnaire.

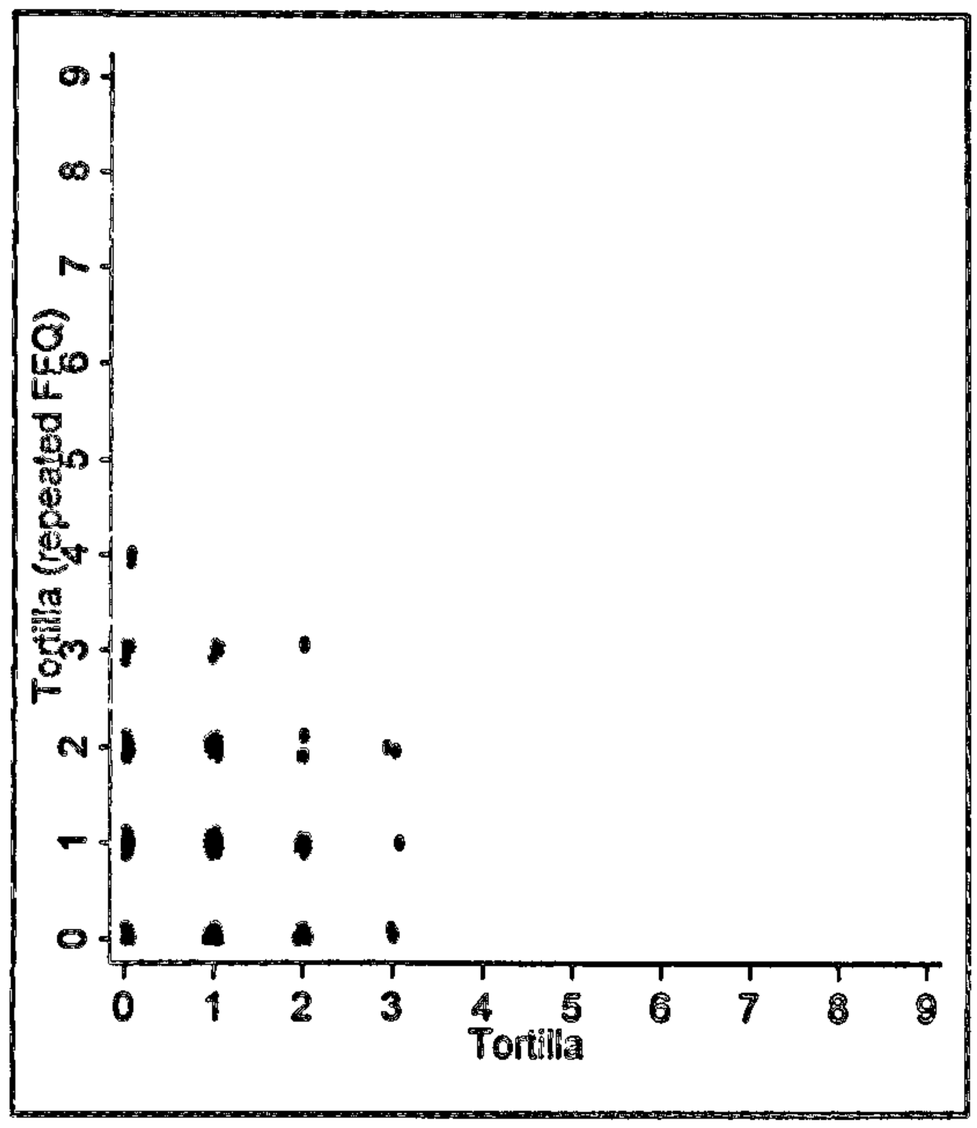


Figure 6.6: Scatter plot of Tortilla in original questionnaire vs repeated questionnaire.

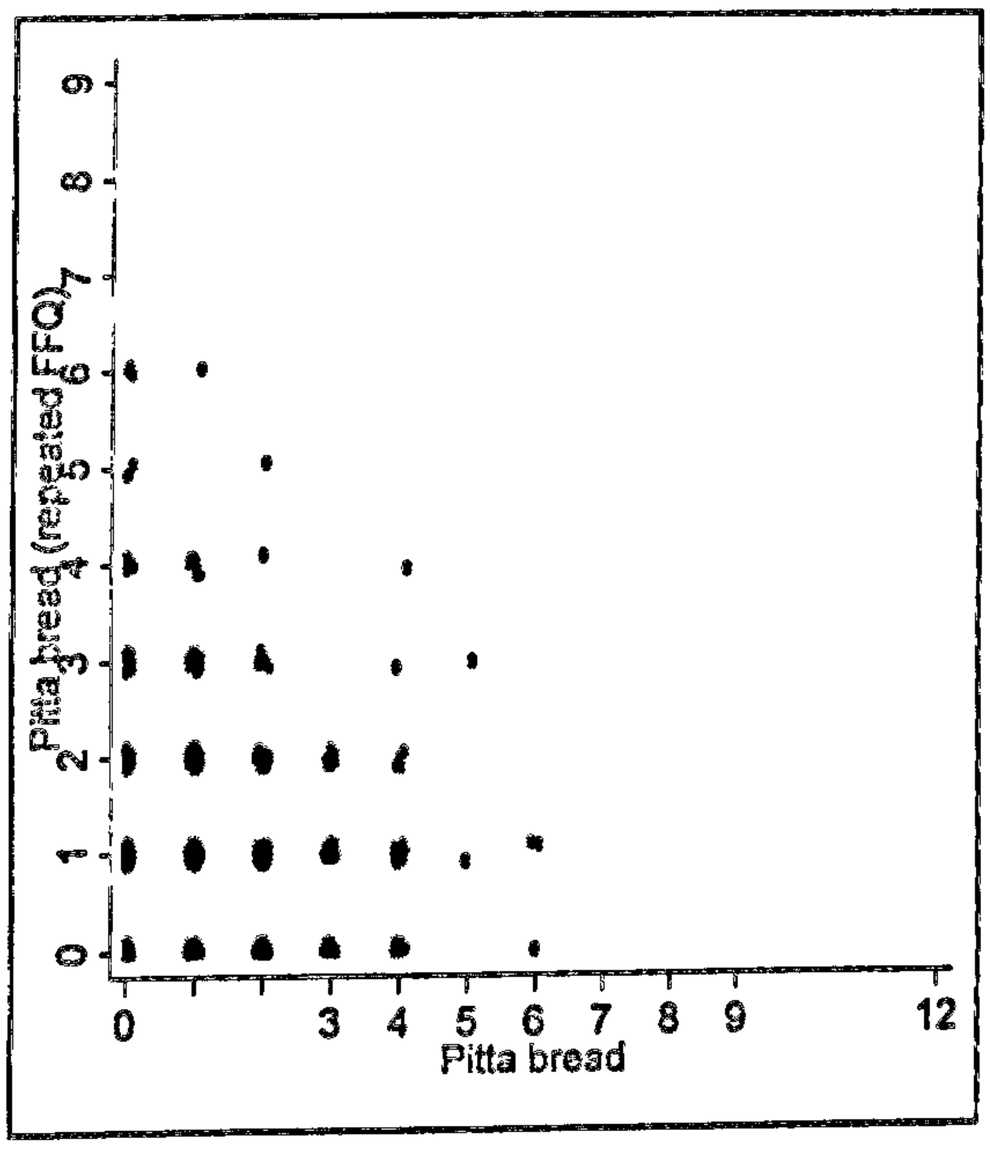


Figure 6.7: Scatter plot of Pitta bread in original questionnaire vs repeated questionnaire.

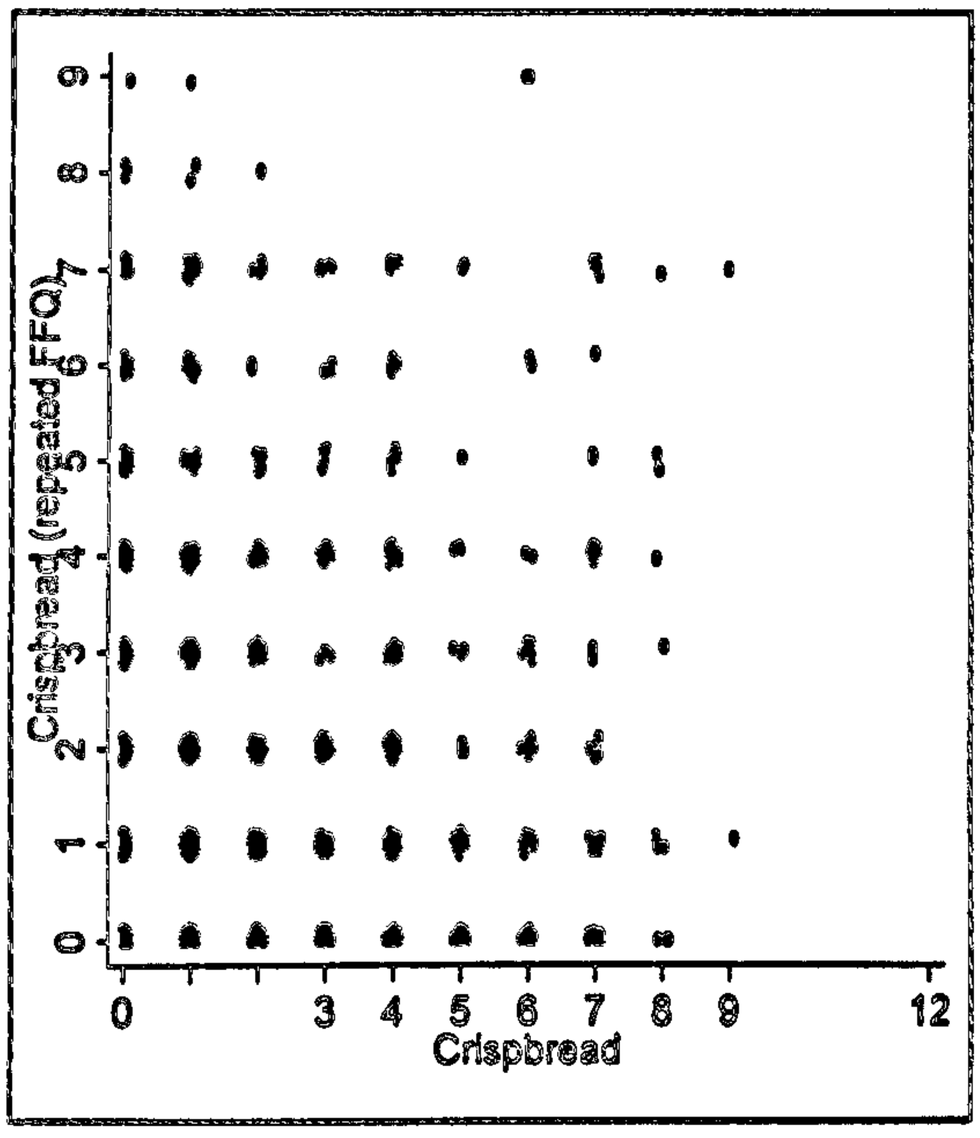


Figure 6.8: Scatter plot of Crispbread in original questionnaire vs repeated questionnaire.

6.4 Measuring agreement in bread block of FFQ versus the repeated FFQ.

In this section, agreement between responses was assessed. A scatter plot of 'white bread' in original and repeated questionnaire was plotted to visually measure the level of agreement between frequencies. If frequencies agree, one would expect all points to lie in the line of equality. Figure 6.9 shows that most of the frequencies do not lie on the line of equality implying rather poor agreement between the two questionnaires.

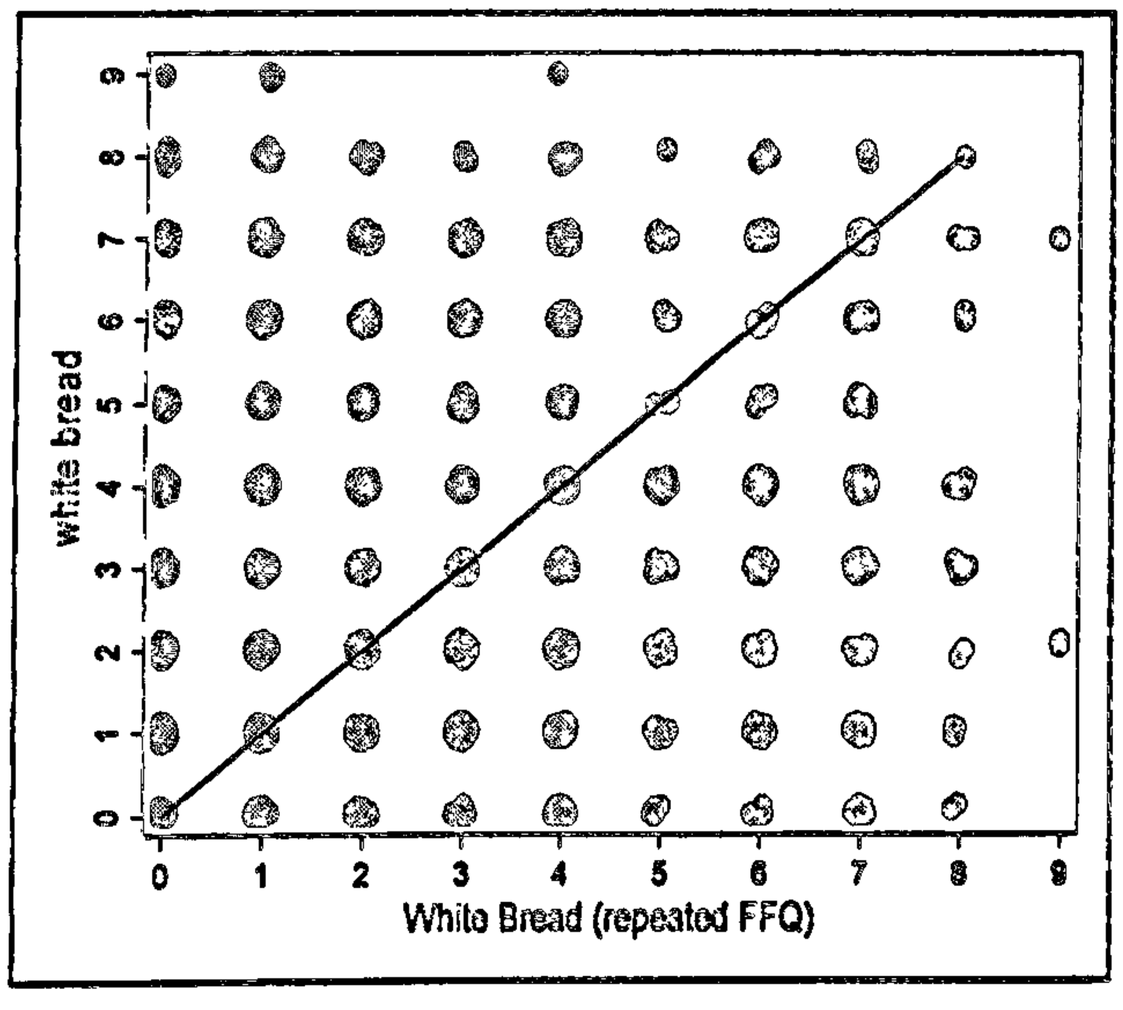


Figure 6.9: Scatter plot of 'white bread' in original questionnaire vs repeated questionnaire, and the line of equality.

The difference between frequencies for 'white bread' consumption in repeated and original was next calculated and plotted against the average frequency of the two questionnaires. The plot of the difference against mean helps to investigate

any possible relationship between the measurement error and the true value, see Figure 6.10. This type of plot was found to give clearer view of agreement between two methods (Bland, 1999).

One would expect most of the differences to lie between the mean difference (D) ± 1.96×S, where D is the mean difference and S is the standard deviation of the differences. These differences are expected to follow a normal distribution. The upper and lower limits of agreement for 'white bread' are calculated as:-

$D- (1.96 \times S) = 0.64 - (1.96 \times 3.22) = -5.67$

$D+ (1.96 \times S) = 0.64 + (1.96 \times 3.22) = 6.95$

This indicates that frequencies for 'white bread' consumption in the repeated questionnaire can be almost 7 points more than the original questionnaire or 6 points less than the original questionnaire. Calculations and graph show obvious lack of agreement between reported frequencies of 'white bread' in original and repeated questionnaire.

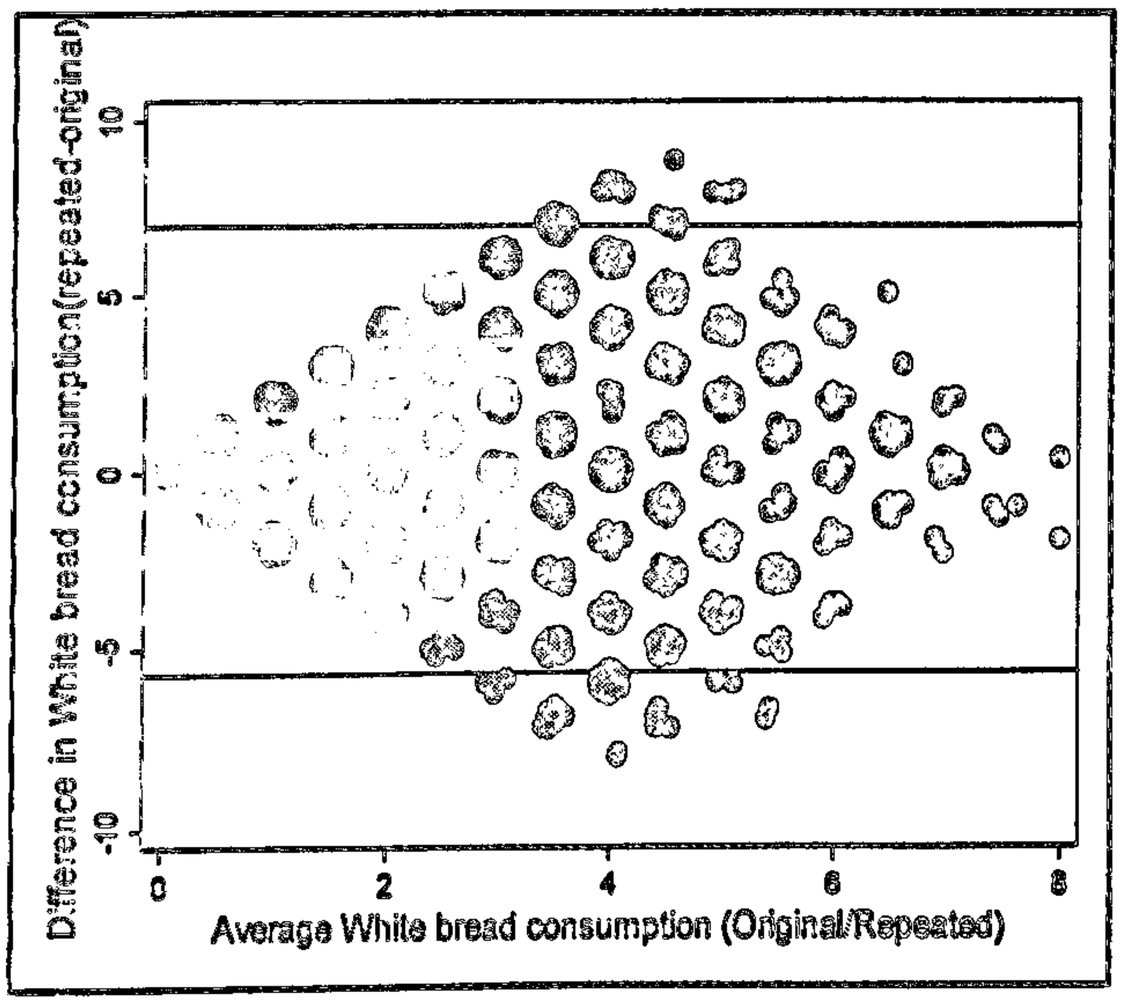


Figure 6.10: Difference against mean of frequencies of 'white bread' in original and repeated questionnaire.

The same calculations were next repeated for the other six types of bread. Table 6.6 shows that reported frequencies do not agree for 'white bread', 'brown bread' 'wholemeal bread' and 'crispbread'. The Table shows that the limits of agreement are narrower for 'chapatis', 'papudum', 'tortillas' and 'pitta bread'.

Variable	D	S	Upper limit	Lower limit
White bread	0.64	3.22	6.95	-5.67
Brown bread	0.38	3.13	6.51	-5.75
Wholemeal	-0.63	3.35	5.93	-7.19
Chapatis	-0.20	1.23	2.38	-2.61
Papudum	-0.05	1.05	2.00	-2.11
Tortillas	-0.11	0.85	1.55	-1.77
Pitta bread	-0.01	1.31	2.55	-2.57
Crispbread	-0.02	2.60	5.07	-5.11

Table 6.6: Upper and lower limits of agreement for each bread type, D= mean difference and S=standard deviation of differences, between original and repeated questionnaire

There was no evidence of agreement in all the above graphical explorations, between responses to 'white bread', 'brown bread' and 'wholemeal' blocks in original and repeated questionnaire. However, the range of upper and lower limits of agreement were much narrower for 'chapatis', 'tortillas', 'papudum' and 'pitta bread'.

6.5 Missing meant to be zero or 'never'

In this section the possibility of respondents skipping questions to say they don't consume that specific item was tested. One would expect respondents who left

item missing in the original questionnaire to fill in 'never' in the repeated questionnaire, if that were the case.

Responses to 'white bread' and 'brown bread' of the original questionnaire were cross-tabulated against responses in the repeated questionnaire in Tables 6.7 and 6.8. The FFQ variables were collapsed to three categories of no consumption, missing and positive (any reported frequency of consumption coded 1-9).

	'White bread' Original questionnaire		
'White bread' repeated questionnaire	Never	Positive	Missing
Never	17 (0.9%)	237 (12.0%)	0 (0.0%)
Positive	157 (8.2%)	1,487 (77.5%)	2 (0.1%)
Missing	1 (0.1%)	17 (0.9%)	0 (0.0%)

Table 6.7: Missing, zero and positive values of the responses to FFQ item about 'white bread' in original and repeated questionnaire.

Brown bread Repeated questionnaire	Brown bread Original questionnaire		
	Never	Positive	Missing
Never	63 (3.3%)	248 (12.9%)	1 (0.0%)
Positive	271 (14.1%)	1,146 (59.7%)	0 (0.0%)
Missing	38 (2.0%)	151 (7.9%)	0 (0.0%)

Table 6.8: Missing, zero and positive values of the responses to FFQ item about Brown bread at original and repeated questionnaire.

Table 6.7 shows that no subject had missing values on the two responses to 'white bread'; one filled in zero and was missing in the repeated questionnaire. Table 6.8 shows that more missing values were observed for 'brown bread' in the repeated questionnaire. The highest frequency in the two tables was the positive frequency,

which suggests that women who claimed eating white or brown bread with any frequency remained doing the same in the repeated questionnaire. The aim behind the sign tabulation in Tables 6.7 and 6.8 was to investigate if respondents left the question missing to indicate that they do not consume that type of bread, but no such evidence was clear from the Table and one can conclude that missing values were not intentional. The possibility that the question was skipped accidentally was much stronger. It was clear that items left missing were skipped or left missing at random, and omissions were not related to subject's diet.

6.6 Alcohol consumption in the original and repeated questionnaire

In Sections 6.2 and 6.3 agreements between responses to original vs. repeated questionnaires were investigated for 'white bread' and 'brown bread'. The aim was to understand the motives behind subjects leaving part of the questions missing, and whether these women were consistent in their responses. In this Section the same procedures were applied to 'alcohol' and 'smoking'. Alcohol consumption was collected in two parts of the questionnaire, see Section 3.1. The long questions of alcohol consumption and smoking were investigated. The recall of smoking habits and alcohol consumption can be easier than a food item eaten on everyday basis like bread, nevertheless people tend to hide their actual drinking habits if they were binge drinkers.

Table 6.9 shows the cross-tabulation of alcohol consumption in the repeated questionnaire with responses to the original questionnaire. Large numbers were observed on the diagonal, which was also supported by Figure 6.11 where the clouds of responses were denser on the diagonal. These results were further supported by the kappa statistics on Table 6.10, which shows good agreement between the original and repeated questionnaire for alcohol and very good agreement for smoking.

Out of the 48 subjects who left alcohol consumption in the original questionnaire missing, 13 responded as drinking more than once a week, and no one left it missing in the repeated questionnaire. Out of 12 who left the alcohol consumption of the repeated questionnaire missing, 3 claimed consuming alcohol more than once a week, 8 claimed consuming alcohol less than once a week. This suggests that most failure of responses to the alcohol consumption were not intentional.

One should bear in mind that the sample that filled in the repeated questionnaire might not be representative of the whole cohort. These results implies that there was considerable uncertainty about most missing values in responses to alcohol consumption. Although there was consistency in responses to alcohol consumption in the two occasions, however subjects who left alcohol consumption missing in the original questionnaire gave different responses in the repeated questionnaire.

Alcohol repeated questionnaire	Alcohol original questionnaire				
	More than once a week	Once a week	Less than once a week	Never drink alcohol	Missing
More than once a week	817	102	57	8	13
Once a week	63	95	58	5	8
Less than once a week	56	49	276	25	10
Never drink alcohol	8	7	49	183	17
Missing	3	0	8	1	0

Table 6.9: Distribution of responses to first question of alcohol consumption in original questionnaire against repeated questionnaire.

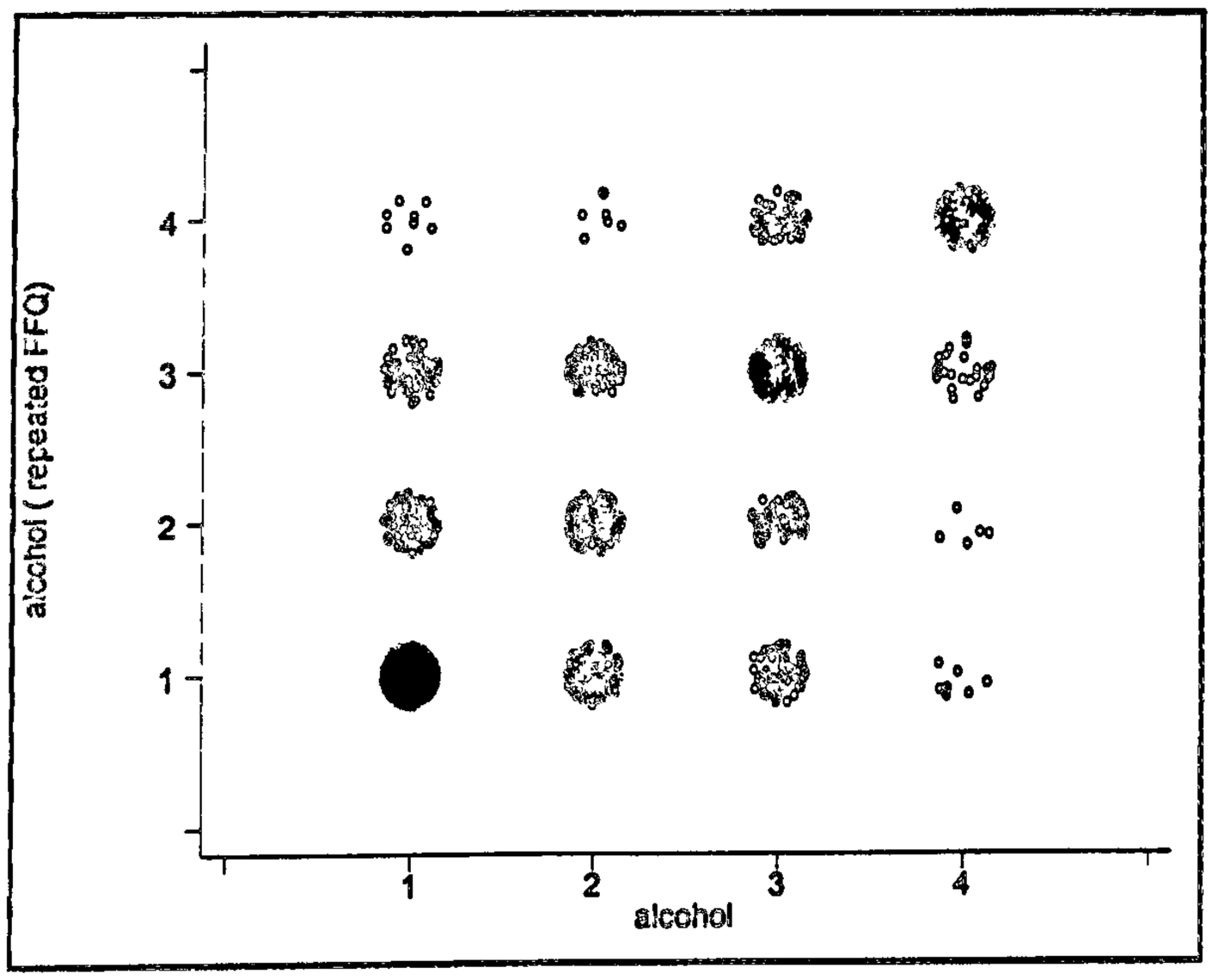


Figure 6.11: Scatter plot of alcohol consumption in original questionnaire against repeated questionnaire.

Note: 1=more than once a week, 2=Once a week, 3=Less than once a week, 4=Never drink alcohol.

Variable	Agreement	Expected agreement	Kappa	Std. Err.
Alcohol	88.45%	60.77%	0.706	0.018
Smoking	97.56%	81.68%	0.867	0.018

Table 6.10: Assessing the agreement of alcohol, smoking of Phase 1 and the repeated questionnaire using the Kappa statistics

The same exploration was done for smoking, in the original and repeated questionnaire, see Table 6.11. The Table shows greater consistency in responses to the two sets of the questionnaire, this consistency was even clearer in Figure 6.12, where clouds were denser and accumulated at the diagonal. Out of the 51 who left this question missing in the original questionnaire, 26 reported never smoked in the repeated questionnaire, and out of the 73 who left this question missing in the repeated questionnaire 54 reported never smoked in the original questionnaire, 8 left this question missing on the two sets of questionnaire. The percentage of missing values was 2.7% in the original questionnaire and 3.8% in the repeated questionnaire, out of the sample, which responded to the two questionnaires. This suggests that most of the small percentage that left this question missing were actually the category that never smoked.

The recall of smoking habit was easier than food items, and would not be considered as a stigma like alcohol consumption for women respondents, which was likely to be the case in the cohort.

Smoking repeated questionnaire	Smoking original questionnaire				
	Smoke everyday	Smoke occasionally	Used to smoke everyday	Never smoked	Missing
Smoke everyday	81	3	12	1	0
Smoke occasionally	3	18	16	1	1
Used to smoke everyday	27	16	507	23	16
Never smoked	0	3	24	1067	26
Missing	0	2	9	54	8

Table 6.11: Distribution of responses to smoking in original questionnaire against repeated questionnaire

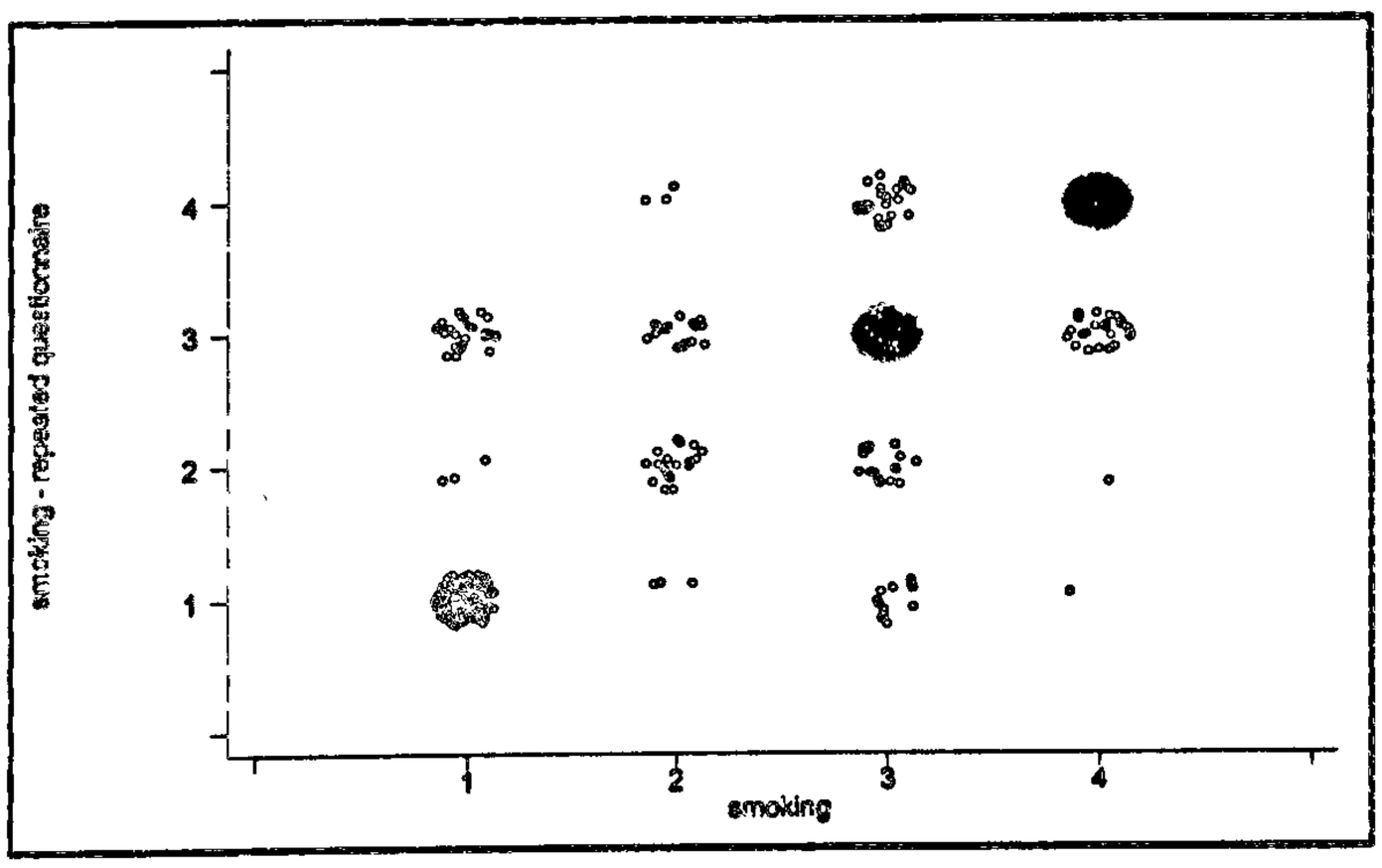


Figure 6.12: Scatter plot of smoking in original questionnaire against repeated questionnaire.

Note: 1= Smoke everyday, 2=Smoke occasionally, but not everyday, 3=Used to smoke, 4=Never smoked.

6.7 Discussion

The impact of the missing data on the results of the statistical analysis depends on the mechanism that caused the data to be missing. Data can be missing for many reasons; participants can sometimes skip a question accidentally, because they do not know the answer, because they do not want to disclose the required information, or even because they did not understand the question. These issues were discussed in Section 2.2.

Knowledge of the mechanism of missing data is the main element in determining a proper method for handling missing data, and largely determines the performance of this method. It is however, impossible to verify the MAR assumption and the causes of missingness unless one gets hold of the missing data by revisiting respondents, which can be difficult in most of the cases. One can otherwise investigate the missing data patterns and use the available information to make reasonable guesses about the mechanism.

The test of this assumption has not received much attention. When the missing values are in one variable X, for example, the mechanism of missingness can be tested by separating the fully observed variables in the dataset into respondents and non-respondents of X, by applying a t-test of difference in means. I tried a different method in Chapter 4 in which a dummy category was assigned to missing values and then fitting the analysis model to find out if they acted differently to observed values. That was also repeated in the survival analysis model. A binary variable was generated for each incomplete variable, by assigning missing values to one and observed values to zero. Logistic regression

models were next fitted for each incomplete variable, to test if there was significant difference in survival between missing and observed values.

An important assumption about the nature of the missing data, which is often made, is that it is missing at random. Rubin (1987, pp.155) states:

"An important feature of the assumption of ignorable non-response is that generally there would be direct evidence in the data to contradict it. [...] Since no X values are observed for no respondents, without external information there will be no way to judge whether the nonrespondents' missing values are systematically different from the respondents' observed values".

In other words the un-testable assumption of missing should be valid unless there is evidence from the data that contradicts it.

For a massive dataset, like the UK Women Cohort Study, the task of investigating the mechanism of missing data was not straightforward. First, the data consisted of categorical as well as continuous variables; second, the content of the variables covered diet as well as many life-style factors. With 35,000 subjects taking part, respondents could have had different motives behind skipping part of the questionnaire; these motives could never be guessed, and that made the task of testing the mechanism of missing data even harder.

In this chapter, I tried to check the mechanism of missing data within a number of questions by looking at the sample that completed the repeated questionnaire. As the only way of testing the exact mechanism of missing data was to have the missing data itself, I studied the consistency of the responses in the two sets of the questionnaire.

As mentioned earlier, the missing data mechanism has strong impact on the method to be used for handling missing data. The aim of this chapter was to explore the fact that the missing data mechanism was MNAR or in other words non-ignorable. As this was the only case of missing data mechanism when the application of handling missing data by multiple imputation would result in biased inferences.

Two variables in FFQ format were studied in detail, as well as two variables in the form of long questions.

Following the above investigations of missing items on selected variables, there was no evidence that the missing value mechanism was non-ignorable. As the assumption of MAR indicates that missing values depends on observed values rather than values which are missing. It is recommended that one should test for predictors of missingness for every variable to be imputed, see Section 5.4.2. Including, as many predictors in the imputation model as possible tend to make the MAR assumption even more plausible.

One could argue that the mechanism of missing data would be different for the rest of the variables, which were not studied. But even if that was the case, and the analysis was applied to complete cases only, the analyst would base inferences about variables on complete cases excluding any incomplete record due to any form of non-response. The complete case analysis discussed in Section 2.2.1, assumes that the mechanism of missing data is MCAR, i.e. the missing cases are a random sample of the complete cases, which is even a stronger assumption than MAR required by multiple imputation for the method to be valid.

Chapter 7
Summary

7.1 Discussion

Like most large-scale surveys, and postal surveys in particular, UKWCS suffers from the problem of missing data. Resolving it is not an easy task. Researchers in the last two decades started to understand the drawbacks of naïve methods for handling of missing data and the effect they have on conclusions. The old style default method of analysing only complete cases is another method of dealing with missing data, but in a way that introduces bias unless the strong MCAR assumption is met.

No method of handling missing data can replace the actual dataset. Hence, in planning surveys, researchers must think of research designs and data collection modes that minimize missing data. New portable computers, and interviews held by professional interviewers, help a lot to reduce the amount of missing data in large surveys. However, surveys of this type can sometimes be quite expensive or not feasible in particular fields. Subjects in some cases can have every intention of filling in the entire questionnaire, but they may not understand or get confused by part of the questionnaire. As a result, they skip some questions. Therefore, a short and easy questionnaire helps in getting better quality data, which can be almost complete. If all the efforts of avoiding missing data fail there should also be resources to make it possible to go back and contact subjects taking part in the

survey to be able to complete the information, which was not reported, or was not filled out properly.

In all multi-variable analyses, omitting all the records with missing data introduces bias and reduces the precision of estimation. For large datasets, like the UKWCS, a general solution for the problem of missing data has to be devised. This can help in resolving the problem of missing data once and for all research questions that will be analysed from the cohort in the future. The secondary analysts working on new studies will not need the skills or the knowledge of handling missing data, as they will be dealing with a complete dataset rather than a dataset full of gaps.

Great improvements to techniques of handling missing data to replace ad hoc methods traditionally used are becoming available to data analysts. As missing data are not avoidable most of the time, the dataset to be analysed must be examined prior to any analysis for the amount of missing information and the mechanism of missing data. This investigation can give a wider view on the best approach to handle missing data. Even if the researcher could not reach the best method to handle the problem efficiently, a compromise is to be made to improve the methods, had the missing data been ignored, or not dealt with at all.

The EM algorithm Dempster, Laird and Rubin (1977) is a powerful tool for handling missing data. However, the maximum likelihood methods are computationally complicated and require a special implementation for each statistical model. In large-scale surveys, where many analyses are expected to be conducted from the same dataset, and different statistical models have to be

implemented; the EM algorithm might not be feasible. The EM algorithm is difficult to implement for some complex statistical models, and that's when the sufficient statistics at the E-step cannot be calculated.

Multiple imputation is one of the attractive methods that should be considered. This method is especially powerful because of its generality and ease of use. Although multiple imputation is getting more and more popular in handling missing data in large-scale surveys, particularly in the USA (Rubin and Schenker, 1991), it has not been widely applied in health-care research and epidemiological surveys in the UK and Europe. The major drawback of multiple imputation is that it involves extensive programming, and requires large computer storage space, but once generated at the database construction stage, the multiple datasets can be analysed by standard statistical methods. Secondary data analysts can analyse the data at a later stage without requiring any specialized software or expertise for handling missing data.

To assess the impact of missing data and its implications on the results of statistical models applied to the UKWCS, several methods of handling missing data have been compared in three different analyses.

Firstly, I investigated the impact on alcohol intake. As alcohol questions were in two parts of the questionnaire, it was found that the women's responses to the FFQ were more complete. The long questions on alcohol consumption were originally intended as crosscheck items. Response to these crosscheck questions was very poor. The estimate of the overall alcohol nutrient intake varied greatly

depending on the method used for handling missing data. The complete case analysis, the default methods for such problems in large-scale surveys, underestimated the actual intake. This was because a lot of valuable information was lost as a result of discarding incomplete records. Some of these discarded records were not included because subjects reported consuming one or two alcohol types and left the others missing. It was found that these respondents with incomplete information on alcohol consumption, drank more so excluding them lead to biased results. Imputation of a default value underestimated the actual missing values, and standard errors were deceptively small, as the result of not taking account of uncertainty. Although this method of imputing a default value has been criticized a lot recently, it is still the established solution for the missing data problem in many surveys. In the field of nutritional epidemiological research, missing values are most of the time substituted by zeros. The basis of substituting such a value is the belief that the subjects skipped the question to say that they do not consume that particular item. Although this assumption can be realistic in part of the questions in FFQ format, it was not well founded in other cases. Imputing zeros did not have a strong effect on the FFQ version of the alcohol consumption question. That was mainly because missing data were very rare in that part. However, a single imputation with zero can be unsafe for some of the nutrients.

Two applications of multiple imputation were applied to the alcohol consumption variables, however each application had its own drawback.

- Multiple imputation by conditioning on one variable was capable of generating plausible values and thus make use of a lot of information that

would be an used in the complete case analysis. Uncertainty was taken into consideration by the addition of variability, however the applied method for the generation of the plausible values was further modified in later chapters to make the assumption of MAR more realistic.

- Multiple imputation by MCMC using the multi-level modelling software *MLwiN*, generated a lot of values that were not consistent with the dataset, the software was slow and the main drawback was its incapability of dealing with categorical and binary variables.

The impact of handling missing data was investigated on a published study from the cohort (Pollard *et al.*, 2001). The paper based all the results on complete cases, which were around one third of the dataset. The analysis was repeated following multiple imputation. The effect of the complete case analysis on the final results was very clear. Information on more than 13,000 extra records could be included in the analysis, after imputation. This gain had a great effect on the final results. The paper also reported that missing data had no effect on the results of the multi-variable analysis; this conclusion was based on the fact that when single variable analyses were carried on the 10,316 subjects there was no change in results. However, odds-ratios after multiple imputation were changed for most of the variables, combined with greater precision, lower standard errors.

Rubin (1987) and Schafer (1987) discussed that, to generate missing values using multi-variable regression a monotone pattern of missingness is required, see Section 2.4 for a description of monotone pattern. However, the problem in practice is that it is very rare that any real dataset has a perfectly monotone pattern of missingness. I have shown that multiple imputation by chained equation proved

to be very powerful in this problem of non-monotone missingness. Multiple imputation by the chained equations was capable of getting around the difficulty of having missing values in every variable by imputing starting values and then generating plausible values through a loop of ten iterations using the Gibbs sampling, which specifies a set of conditional distributions one for each incomplete variable. The method does not specify a form for the multivariate distribution, but do assume that a multivariate distribution exists. Specifications of imputation models were easier, and this depends on the type of variable to be imputed (logistic regression for binary variables, polytomous regression for categorical and linear regression for continuous variables). All imputations were developed under the MAR assumption and did not require modelling of the missing data mechanism. The algorithm was quick to converge. The program routines were written in STATA 8 code and can easily be modified to impute for any variable in the dataset.

Results of the complete case analysis and multiple imputation were also compared in a different type of multi-variable analysis, survival analysis. This model investigated the association between incidence of breast cancer and a number of prognostic factors. A similar model can be used in future when information on cancer is complete, to find the association between incidence of cancer and life-style factors, and so this is a good illustration of how the use of multiple imputation could enhance a future key analysis. In this analysis two different methods of handling missing data were compared, and a survival analysis model (Cox's Proportional Hazards) was fitted after each method. Hotdeck imputation was applied to impute the missing values. The hotdeck was improved by running

it several times in the form of a multiple imputation routine. Thus uncertainity due to missing values was taken into account. The negative aspect of this method is in the technique STATA 8 uses for hotdeck imputation, in which some observed values are sometimes changed, by substituting an incomplete record with a complete record rather than just filling in the missing values.

Comparison of hazard ratios and standard errors, after the complete case analysis, hotdeck imputation and multiple imputation, showed improvement in results after hotdeck imputation, and multiple imputation.

The principal assumptions of multiple imputation are that the missing data mechanism is correctly specified, the imputations are proper and that the planned analysis is efficient. The proper-ness refers to reflection of the uncertainty about the missing values; this usually entails uncertainty in the model fit (sampling variation) and the random variation implied by the model. Correctness of the model cannot be verified, but more complex models involving many or all the available variables, come closer to the ideal.

Multiple imputation involve assumptions that are either un-testable or cannot be tested with sufficient power. In the UKWCS subjects who failed to return the questionnaire were reminded by telephone calls. But nothing could be done for item non-response, where only parts of the questionnaires had no responses. The repeated questionnaire, although not originally sent out to assist with missing data, helped to understand the missing data mechanism. Responses to repeated questionnaire, of subjects who left the same questions missing in the original, suggested that non-response was not intentional, suggesting that the mechanism of MAR might hold, although there was no solid proof for it. Rubin (1987)

commented that even if MAR does not hold, multiple imputation based on MAR tends to be less biased than naïve methods, such as analysis carried out on complete cases only.

Multiple imputation by chained equations, assumes that a multivariate distribution exists and draws from it are generated using a Gibbs sampling. The method proved to be very efficient and was capable of generating plausible values consistent with the original observed values. It is recommended that the imputation model should use all variables in the analysis model as predictors of missingness as well as extra variables from the dataset if they proved to predict missing values. This technique of selecting as many predictors as possible for the imputation model tends to make the assumption of MAR more reasonable. The UKWCS consist of more than six hundred variables. Computational complexities as well as multicolinearity problems make it not practical to use all the variables in the dataset for imputation models to the generate plausible values. This is also not necessary, as it was shown in Chapter 5 that little was gained by the inclusion of an additional variable in the imputation model after the best set of variables were selected. Results in Chapter 4 suggest that the algorithm works well in this application and convergence was achieved in around six iterations. No burn-in iterations were discarded, and convergence was much quicker than the MCMC method applied in *MLwiN* in which convergence was achieved by thousands of iterations. The Gibbs sampler simulates draws from the multivariate posterior distribution by repeatedly drawing from a set of conditional distributions, provided the former exists, however it is possible that the conditional distributions are incompatible and therefore no joint distribution exists. Therefore, assessing

convergence is recommended, and this can by achieved by plotting the mean and standard deviations of imputations, similar to that presented in Chapter 4.

Specialized multiple imputation software packages are becoming available. One of the most popular packages is SOLAS, which implements an approximation for the Bayesian bootstrap (Rubin 1987). NORM and CAT, to handle continuous and categorical variables, respectively, are becoming more and more popular. Allison (2000) discussed a simulation study in which Schafer's multiple imputation packages simulated missing data with little bias. In case of the UKWCS, the dataset is massive with a large number of variables, so using NORM or CAT is not feasible. Hence, programming features of the more flexible statistical package STATA 8 (STATA Corporation, 2003) was used for developing all the routines used in multiple imputation. These routines can be used to impute values for all variables very easily, just by changing variable names.

Imputation always provides made-up values and the imputed values can never be an ideal substitute for the real observed values. However, the benefits and improvement to the methods are not by the values substituted for missing data but by the improved inferences and results after the analysis.

The comparison of complete case and multiple imputation analysis suggests that ignoring missing data or applying naïve methods of single imputation can alter the actual associations. Examples found are the total consumption of alcohol, the association between high and low consumption of fruit and vegetables and socio-economic factors, as well as the relation between getting the cancer and life style factors. An investment in improving the applied method of handling missing data

will lead to greater precision of the inferences, and with it better exploitation of the survey as a resource for research in nutrition and epidemiology.

This thesis recommends multiple imputation as a solution for handling missing data in the UKWCS. Although it needs more computing effort, and is more difficult to implement than most conventional methods, it greatly improves the results, compared to the default method of imputing zeros or to the complete case analyses.

7.2 Future work

It is recommended for future work that the STATA 8 routines developed for the generation of multiple imputation should be written in the form of a menu driven STATA ado file. This will make it more users friendly for a non-statistician to use. Steps for the selection of predictor variables developed in Chapter 5 will help in the selection of predictors in imputation models. For each analysis, results from the five datasets should be combined as outlined in Section 2.3, using the program developed in this thesis. This will help in obtaining more precise results in future analyses of the cohort data. A similar method named as multiple imputation by 'switching regression' (Royston, P., 2004) was implemented and presented at the STATA 10th users group meeting. This implemented method assumes the mechanism of missing data to be MAR. The method is a type of Gibbs Sampler in which the distribution of the missing values of a covariate is sampled conditional on the distribution of the remaining covariates in the

imputation model. Let us assume that *yvar* is the incomplete variable to be imputed, and *xvarlist* are the list of variables in the imputation model, which can also be incomplete. The method proceeds by generating plausible values in *yvar* by multiple regression of yvar on xvarlist, combined by random draws from the conditional distribution of the missing observations given the observed data and covariates. Let the variables in the main variable list be $x_1, x_2, x_3,\ldots,x_n$, (Royston, P., 2004) presented the procedure by the following the steps:-

1- Ignore observations for which every member of *yvar* and *xvarlist* has a missing value.

2- For each variable with missing values in *xvarlist*, initialise the iterative procedure by filling in missing values by random draws from the observed values.

3- For each of $x_1, x_2, x_3,\ldots,x_n$ in turn, impute missing values using multiple regression with the remaining variables as covariates.

4- Repeat 3 for *L* times, named as cycles or iterations. Van Buuren *et al.* (1999) recommended 20 iterations but went on to say that 10 or even 5 are usually sufficient. At each iteration replace previous imputations with the last updated ones.

5- This creates one complete dataset. To obtain m completed datasets the procedure is repeated m times independently.

This method differs from the one implemented in this thesis in the generation of starting values, which is set as mean for continuous variables and mode for categorical variables in the thesis, (Royston, P. 2004), also has the option of generating plausible values by the regression of *yvar* on *xvarlist* within a

bootstrap sample. or by prediction matching, which ensures that the values are imputed only within the range of *yvar*.

A comparison of results using this method to the STATA routines developed in this thesis can help in getting a wider view of the multiple imputation by chained equations method.

Multiple imputation by chained equations works by specifying a set of regression models for each incomplete variable. The method does not assume a specific form for the multivariate distribution. However, it assumes that these set of regression models converge to the joint multivariate distribution. The method converged well in this thesis and in simulation work by (Brand, 1999), but this is not always guaranteed. It is possible that two conditional distributions $P(Y_1|Y_2)$ and $P(Y_2|Y_1)$ are incompatible and their joint distribution $P(Y_1, Y_2)$ does not exist, i.e. this iterative models may never converge. This calls for more work to be done in testing convergence and cases in which convergence cannot be achieved.

The main goal of the UKWCS is to investigate plausible links between long-term diet and cancer. However, information captured by the FFQ was subject to different types of assumptions. First, that the subjects make no deliberate or accidental misjudgements of the pattern of their consumption; second, that every subject's frequency is in the middle of the range of frequencies in the FFQ category; third, that all portions are of equal size and no food is discarded. In addition, the FFQ covers the average intake over one year, but any relation between diet and chronic diseases is based on the life long intake of diet and cannot be based on the average intake of just one year. Diet of a specific person

can fluctuate from one day to the next, and many outside measures can have an influence on it, for example age, the person's mood or even time of the year. Shahar *et al.* (2001) found that although FFQs are designed to assess average yearly food intake, significant seasonal changes were identified in actual dietary intake.

The effect of diet on health is not instant, therefore long-term diet has to be considered and not just diet consumed in the previous month or year before the research. Therefore, more work is needed in the following unresolved issues:-

1. Approximation of portion sizes consumed.
2. Computing the amount of nutrients using conversion formulas and tables, which sometimes does not include every aspect of nutrient in a specific diet.
3. The effect of diet is not instant, therefore long-term diet has to be considered and not just diet consumed in the previous month or year before the research.

The same survival model developed in this thesis can also be applied once data on incidence of cancer is complete. This can be a start of interesting findings that can help to understand the actual impact of diet and a number of other factors, for example smoking and alcohol consumption on the possibility of getting a serious illness like cancer.

References

1. Ajzen, I. (1988). *Attitudes, Personality and Behaviour.* Milton Keynes, UK: Open University Press.
2. Ajzen, I. (1999). The theory of planned behaviour. *Organisational Behaviour and Human Decision Processes* **50**, 179-211.
3. Allison, P. (2000). Multiple imputation for missing data. A cautionary tale. *Sociological Methods and Research* **28**, 301-309.
4. Bailar, B.A., Bailey, L. and Corby, C. (1978). *A Comparison of Some Adjustment and Weighting Procedures for Survey Data. In Survey Sampling and Measurement.* New York: Academic Press. 175-198.
5. Barnard, J. and Meng, X.L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research* **8**, 17-36.
6. Beral, V., Doll R, Peto R and Reeves G on behalf of the Collaborative Group on Hormonal Factors in Breast Cancer (2002). Collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. *British Journal of Cancer* **87**, 1234-1245.
7. Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307-310.
8. Bland, J.M. and Altman, D.G. (1999). Statistical methods for assessing agreement between two methods of clinical measurement. Reproduced by permission of the Lancet. http://www.mbland.sghms.ac.uk/ba.htm.

9. Block, G., Patterson, B. and Subar, A. (1992). Fruit, vegetables and cancer prevention: a review of the epidemiological evidence. *Nutrition and Cancer* **18**, 1-29.

10. Boughton, B. (2001). Dietary folate protects against breast cancer in women who drink alcohol. *The Lancet Oncology* **2**, 528.

11. Box, G.E.P., and Cox, D.R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society Series B* **26,** 211-252.

12. Brand, J.P.L. (1999). *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets*. Ph.D. Thesis, Erasmus University Rotterdam.

13. Breslow, N.E., Cain, K.C. (1988). Logistic regression for two stage case control data. *Biometrika* **75**, 11-20.

14. Brown, M.L. (1990). *Present Knowledge in Nutrition*. 6th Edition. Washington D.C., 212-223.

15. Bueno de Mesquita, H.B., Smeets, F.W., Runia, S. and Hulshof, K.F. (1992). The reproducibility of a food frequency questionnaire among controls participating in a case-control study on cancer. *Nutrition and Cancer* **18,** 143-156.

16. Buzzard, I.M., Stanton, C.A., Figueiredo, M., Fries, E.A. et al (2001). Development and reproducibility of a brief food frequency questionnaire for assessing the fat, fiber, and fruit and vegetable intakes of rural adolescents. *Journal of the American Dietetic Association* **101,** 1438-1446.

17. Cade, J., Thompson, R., Burley, V. and Warm, D. (2002). Development, validation and utilisation of food-frequency questionnaires - a review. *Public Health Nutrition* **5**, 567- 587.

18. Calvert, C., Cade, J., Barrett, J.H. and Woodhouse A. (1997). Using crosscheck questions to address the problem of miss-reporting of specific food groups on Food Frequency Questionnaires. UKWCS Steering Group. United Kingdom Women's Cohort Study Steering Group. *European Journal of Clinical Nutrition* **51**, 708-712.

19. Cancer Research UK. http://www.cancerhelp.org.uk, Cancer Help UK.

20. Casella, G., and George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician* **46**, 167-174.

21. CGHFBC (2002). Collaborative group on hormonal factors in breast cancer. Alcohol, tobacco and breast cancer- collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 without the disease. *British Journal of Cancer* **87**, 1234-1245.

22. Chen, M.H. and Ibrahim, J.G. (2001) Maximum likelihood methods for cure rate models with missing covariates. *Biometrics* **57**, 43-52.

23. Chow, W.K. (1979). A look at various estimators in logistic models in the presence of missing values. *American Statistical Association*. In: Proceedings of the Business and Economics Section, 417-420.

24. Coder, J. (1978). Income data collection and processing for the March income supplement to the current population survey. *Proceedings of the Data Processing Workshops: Survey of Income and Programs*

Participation. Washington: US Department of Health, Education and Welfare.

25. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37-46.

26. Colledge, M.J., Johnson, H., Pare, R. and Sande, I.G. (1978). Large scale imputation of survey data. *American Statistical Association: Proceedings on the Section on Survey Research Methods*, 431-435.

27. Collins, L.M., Schafer, J.L. and Kam, C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* **6**, 330-351.

28. COMA (1994). Cardiovascular review group, Committee on medical aspects of food policy. *Nutritional aspects of cardiovascular disease: Department of Health report on health and social aspects* **46**. London: HMSO.

29. Congdon, P. (2001). *Bayesian Statistical Modelling*. Chichester: Wiley.

30. Conner, M., Kirk, S.F.L., Cade, J.E. and Barrett, J.H. (2001). Why do women use dietary supplements? The use of the theory of planned behaviour to explore beliefs about their use **52**, 621-633.

31. Cowin, I. and Emmett, P. (1999). The effect of missing data in the supplements to McCance and Widdowson's food tables on calculated nutrient intakes. *European Journal of Clinical Nutrition* **53**, 891-894.

32. Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

33. Crawley, H. (1994). *Food portion sizes*. London: HMSO.

34. Daling, J.R., Malone, K.E., Doody, D.R., Johnson, L.G., et al. (2001). Relation of body mass index to tumour markers and survival among young women with invasive ductal breast carcinoma. *Cancer* **92**, 720-729.

35. De Leeuw, E.D. (2001). Reducing missing data in surveys: an overview of methods. *Quality & Quantity* **35**, 147-160.

36. De Stavola, B.L. and Hardy, R. (2000). Birthweight, childhood growth and risk of breast cancer in a British cohort. *British Journal of Cancer* **83**, 964-968.

37. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* Ser. B **39**, 1-38.

38. Dobson, R. (2002). Risk of breast cancer increases with number of years smoking. *British Medical Journal* **325**, 298.

39. Dixon, W.J. (1983). BMDP statistical software. Berkeley University: University of California Press.

40. Ebbert, J.O., Yang, P., Vachon, R.A., Vierkant, R.A. et al. (2003). Lung cancer risk reduction after smoking cessation: Observations from a prospective cohort of women. *Journal of Clinical Oncology* **21**, 921-926.

41. Efron, B., (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association* **89**, 463-475.

42. Enger S.M., Ross R.K., Paganini-Hill A, Longnecker MP, et al. (1999). Alcohol consumption and breast cancer oestrogen and progesterone receptor status. *British Journal of Cancer* **79**, 1308-1314.

43. Engle, A., Lynn, L.L., Koury, K. and Boyar, A.P. (1990). Reproducibility and comparability of a computerized, self-administered food frequency questionnaire. *Nutrition and Cancer* **13**, 281-292.

44. Ewartz, M., Duffy, S., Adami, H. (1990). Age at first birth, parity and risk of breast cancer: a meta analysis of 8 studies from Nordic countries. *International Journal of Cancer* **46**, 597-603.

45. Ezzati-Rice, T.M., Khare, M. and Schafer, J.L. (1993). Multiple imputation of missing data in NHANES III, paper presented at the American Statistical Association Annual Meeting, San Francisco.

46. Ezzati-Rice, T.M., Johneson, W., Khare, M., Little, R.J.A. et al. (1995). A simulation study to evaluate the performance of model-based multiple imputations in CHHS Health Examination Surveys. In: *Proceedings of the Bureau of the Census Annual Research Conference*, 257-266.

47. Flanders, W.D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* **10**, 739-747.

48. Gandini, S., Merzenich, H., Robertson, C. and Boyle, P. (2000). Meta-analysis of studies on breast cancer risk and diet: the role of fruit and vegetable consumption and the intake of associated micronutrients. *European Journal of Cancer* **36**, 636-646.

49. Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

50. Gelfand, A.E., Hills, S.E., Racine Poon, A. and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85**, 972-985.

51. Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences with (discussion). *Statistical Science* **7**, 457-472.

52. Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741.

53. Geskus, R.B. (2001). Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored. *Statistics in Medicine* **20**, 795-812.

54. Geyer, C. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **7**, 473-483.

55. Ghadirian, P., Lynch, H.T., Krewski and D. (2003). Epidemiology of pancreatic cancer: an overview. *Cancer Detection and Prevention* **27**, 87-93.

56. Gilks W.R., Richardson, S. and Spiegelhalter, D.J. (1996). Introducing Markov Chain Monte Carlo. In: *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall, 1-19.

57. Gillum, R.F., Mussolino, M.E. and Ingram, D.D., (1996). Physical activity and stroke incidence in women and men: The NHANES I epidemiologic follow-up study. *American Journal of Epidemiology* **143**, 860-869.

58. Graham, J.W. and Donaldson, S.I. (1993). Evaluating interventions with differential attrition: The importance of non-response mechanisms and use of follow-up data. *Journal of Applied Psychology* 78, 119-128.

59. Graham, J.W., Hofer, S.M., and Piccinin, A.M. (1994). Analysis of missing data in drug prevention (MCLA). *Advances in data analysis for prevention intervention research* **142.** Rockville MD: National Institute on Drug Abuse.

60. Graham, J.W., Hofer, S.M. and MacKinnon, D.P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Science* **32**, 197-218.

61. Graham, J.W. and Schafer, J.L. (1999). On the performance of multiple imputation for multivariate data with small sample size. *Statistical Strategies for Sample Size Researc,* 1-29. Rick Hoyle. Thousand Oaks, CA: Sage.

62. Grambsch, P.M. and Therneau, T.M.(1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81, 515-526.**

63. Greenland, S. and Finkle, W.D. (1995). A critical look at methods for handling missing covariates in epidemiological regression analyses. *American Journal of Epidemiology* **142, 1255-1264.**

64. Greenwood, D.C., Cade, J.E., Draper, A., Barrett, J.H. et al. (2000). Seven unique food consumption patterns identified among the women in the UK Women's Cohort Study. *European Journal of Clinical Nutrition* **54,** 314-320.

65. Gronbaek, M., Becker, U., Johansen, D., Gottschau, A. et al. (2000). Type of alcohol consumed and mortality from all causes, coronary heart disease, and cancer. *Annals of Internal Medicine* **133**, 411-419.

66. Gunnell, D., Oliver, S.E., Peters, T.J., Donovan, J.L. et al. (2003). Are diet-prostate cancer associations mediated by the IGF axis? A cross-sectional analysis of diet, IGF-I and IGFBP-3 in healthy middle-aged men. *British Journal of Cancer* **88**, 1682-1686.

67. Haroon, A. (2003). WHO's diet report prompts food industry backlash. *The Lancet* **361**, 1442.

68. Hartley, H.O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics* **14**, 174-194.

69. Haverkos, H.W., Soon, G.X., Steckley, S.L. and Pickworth, W. (2003). Cigarette smoking and cervical cancer: Part I: a meta-analysis. *Biomedicine and Pharmacotherapy* **27**, 67-77.

70. Hediger, M.L., Overpeck, M.D., Mc Glynn, A., Kuczmarski, R.J., et al. (1999). Growth and fatness at theree to six years of age of children born small- or large-for-gestational age. *Paediatrics* **104**, e33.

71. Heitjan, D.F. (1997). Annotation: What can be done about missing data? Approaches to imputation. *American Journal of Public Health* **87**, 548-50.

72. Heitjan, D.F. and Little, R.J. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics* **40**, 13-29.

73. Horton, N.J. (2001). Maximum likelihood analysis of logistic regression model with incomplete covariate data and auxiliary information. *Biometrics* **57**, 34-42.

74. Horton, N.J., and Lipsitz, S.R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician* **55**, 244-254.

75. Hosmer, D.W. Jr., and S. Lemeshow (1999). *Applied Survival Analysis. Regression Modelling of Time to Event Data.* Wiley, New York (Wiley Series in Probability and Statistics).

76. Hsu, F.C., Starkebaum, G., Boyko, E.J. and Dominitz, J.A. (2003). Prevalence of rheumatoid arthritis and hepatitis C in those age 60 and older in a US population based study. *Journal of Rheumatology* **30**, 455-458.

77. Hu, F.B., Manson J.E., Stampfer, M.J., Colditz G. et al. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England Journal of Medicine* **345**, 790-797.

78. Huisman, M., Krol, B. and Sonderen, E. (1998). Handling missing data by re-approaching non-responders. *Quality and Quantity* **32**, 77-91.

79. Hunsberger, S., Murray, D., Davis, C.E. and Fabsitz, R.R. (2001). Imputation strategies for missing data in a school-based multi-centre study: the Pathways study. *Statistics in Medicine* **20**, 305-316.

80. International Agency for Research on Cancer (1990). *Cancer: Causes, Occurrence and Control*, IARC Scientific Publications, Lyon.

81. Ibrahim, J.G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85**, 765-769.

82. International Life Science Institute (1996). *Present Knowledge in Nutrition*. ILSI Press, Washington, DC.

83. Jones, M. P. (1994). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Technical report.* Department of Statistics, University of Iowa.

84. Josefson, D. (2001). Obesity and inactivity fuel global cancer epidemic. *British Medical Journal* **322**, 945.

85. Joshipura, K.J., Manson, J.E., Stampfer, M.J., Rimm, E.B. et al. (2001). The effect of fruit and vegetable intake on risk for coronary heart disease. *Annals of Internal Medicine* **134**, 1106-1114.

86. Key, T.J., Davey, G.K. and Apple, P.N. (1999). Health benefits of a vegetarian diet. *Proceedings of the Nutritional Society* **58**, 271-275.

87. Key, T.J., Verkasalo, P.K. Banks, E. (2001). Epidemiology of breast cancer. *The Lancet Oncology* **2**, 133-140.

88. Key, T.J., Allen, N.E., Spencer, E.A. and Travis, R.C. (2002). The effect of diet on risk of cancer. *Lancet* **360**, 861-868.

89. Kimati, V.P. (1985). The nutritional status of Tanzanian schoolchildren- A cross-sectional anthropometric community survey report. *East African Medical Journal* **62**, 105-117.

90. Kirk, S. F., Cade, J.E., Barrett, J.H. and Conner, M. (1999). Diet and lifestyle characteristics associated with dietary supplement use in women. *Public Health Nutrition* **2**, 69-73.

91. Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795-806.

92. Krol, B. (1996). *Waiting politely: Investigation of the Duration of Waiting Lists for Orthopaedic Patients*. Groningen: Groeneland verzekeringen, Northern Center for Healthcare Research, University of Groningen.

93. Laird, N.M. (1988). Missing data in longitudinal studies. *Statistics in Medicine* **7**, 305-315.

94. Landis, J.R., and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174.

95. Lenz, S.K., Goldberg, M.S., Labreche, F., Parent, M.E., et al. (2002). Association between alcohol consumption and postmenopausal breast cancer: results of a case-control study in Montreal, Quebec, Canada. *Cancer Causes and Control* **13**, 701-710.

96. Li, K.H. (1980). Imputation using Markov Chains. *Journal of Statistical Computation Simulation* **30**, 57-79.

97. Lipsitz, S.R., Ibrahim, J.G., Chen, M. and Petterson, H. (1999). Non ignorable missing covariates in generalized linear models. *Statistics in Medicine* **18**, 2435-2448.

98. Little, R.J. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* John Wiley and Sons, New York.

99. Little, R.J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* **83**, 1198-1202.

100. Little, R.J. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association* **87**, 1227-1237.

101. Liu, M., Taylor, J.M.G., and Belin, T.R. (2000). Multiple imputation and posterior simulation of multivariate missing data in longitudinal studies. *Biometrics* **56**, 1157-1163.

102. Longford, N.T. (2000). Multiple imputation in an international database of social science surveys. *ZA-Information* **46**, 72-95.

103. Longford, N.T., Ely, M., Hardy, R., and Wadsworth M.E.J. (2000). Handling missing data in diaries of alcohol consumption. *Journal of the Royal Statistical Society Series A* **163**, 381-402.

104. Lyde, P.M., Webster, L.A., Baughman, A.L. et al. (1989).The independent associations of parity, age at first full term pregnancy, and duration of breast feeding with the risk of breast cancer. Journal of Clinical Epidemilogy **42**, 963-973.

105. Lyles, R.H., Fan, D.J. and Chuachoowong, R. (2001). Correlation coefficient estimation involving a left censored laboratory assay variable. *Statistics in Medicine* **20**, 2921-2933.

106. MAFF (1999). Ministry of Agriculture, Fisheries and Food. *National Food Survey 1998*. London: HSMO.

107. Madigan, M.P., Troisi, R. Potischman, N., Brogan, D. et al. (2000). Characteristics of respondents and non-respondents from a case-control study of breast cancer in younger women. *International Journal of Epidemiology* **29**, 793-798.

108. Makimoto, K., Oda, H. and Higuchi, S. (2000). Is heavy alcohol consumption an attributable risk factor for cancer related deaths among Japanese men? *Alcohol, Clinical and Experimental Research* **24**, 382-385.

109. Manjer, J. and Andersson, I. (2000). Survival of women with breast cancer in relation to smoking. *European Journal of Surgery.* **166**, 852-858.

110. Margetts, B.M. and Nelson, M. (1997). *Design Concepts in Nutrition Epidemiology*. Oxford University Press, Oxford.

111. Martinussen, T. (1999). Cox regression with incomplete covariate measurements using the EM-algorithm. *Scandinavian Journal of Statistics* **26**. 479-491.

112. McKendrick, A.G. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society* **44**, 98-130.

113. Mellemkjaer, L., Olsen, M.L., Sorensen, H.T., Thulstrup, A.M. et al. (2003). Birth weight and risk of early-onset breast cancer (Denmark). *Cancer Causes and Controls* **14**, 61-64.

114. Meng, X.L. (1997). The EM algorithm and medical studies: a historical link. *Statistical Methods in Medical Research* **6**, 3-23.

115. Meng X.L. and Van Dyk, D. (1997). The EM algorithm – and old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B* **59**, 511-567.

116. Miller, Y.E. and Fain, P. (2003). Genetic susceptibility to lung cancer. *Seminars in Respiratory and Clinical Care Medicine* **24**, 197-204.

117. Musil, C.M., Warner, C.B., Yobas P.K., Jones, S.L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research* **24**, 815-829.

118. Ness, A.R. and Powels, J.W. (1997). Fruit and vegetables, and cardiovascular disease: a review. *International Journal of Epidemiology* **26**, 1-13.

119. Neyman, J. and Pearson, E.S. (1933). On the problem of most efficient tests of statistical hypotheses. *Philosophical transactions of the Royal Society of London, Series A* **231**, 289-337.

120. ONS (2001). Office of National Statistics. *Mortality Statistics: Causes England and Wales 2000* **27**. TSO.

121. Paton, A. (1994). *ABC of Alcohol*, 3rd edn. London: British Medical Journal.

122. Penny, K. I. and Jolliffe, I. E. (1999). Multivariate outlier detection applied to multiply imputed laboratory data. *Statistics in Medicine* **18**, 1979-1897.

123. Pietinen, P., Hartman, A.M., Happa, E., Rasanen, L., et al. (1988). Reproducibility and validity of dietary assessment instruments. II. A quantitative food frequency questionnaire. *American Journal of Epidemiology* **128**, 667-676.

124. Pollard, J., Greenwood, D., Kirk, S. and Cade, J. (2001). Lifestyle factors affecting fruit and vegetable consumption in the UK Women's Cohort Study. *Appetite* **37**, 71-79.

125. Ramon, J.M., Escriba, J.M., Casas, I., Benet, J. (1996). Age at first full-term pregnancy, lactation and parity and risk of breast cancer: a case-control study in Spain **12**, 449-53.

126. Rehm, J. and Bondy, S. (1998). Alcohol and all-cause mortality: an overview. In: Novartis Foundation Symposium, ed. *Alcohol and Cardiovascular Diseases Chichester*. New York: John Wiley & Sons.

127. Riboli, E. (1992). Nutrition and cancer: Background and rationale of the European prospective investigation into cancer and nutrition (EPIC). *Annals of Oncology* **3**, 783-791.

128. Robert C.P.and Casella G. (1999). *Monte Carlo statistical methods*. New York: Springer.

129. Rosenbaum, P.R. (1995). *Observational Studies*. Springer-Verlag, New York

130. Roth, P. L. (1994). Missing data: a conceptual review for applied psychologists. *Personnel Psychology* **47**, 537-560.

131. Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581-92.

132. Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

133. Rubin, D. B., and Schenker, N. (1991). Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine* **10**, 585-598.

134. Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473-489.

135. Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall.

136. Schafer, J.L. and Graham, W.J. (2002). Missing data: Our view of the state of the art. *Psychological Methods* **7**, 147-177.

137. Schill, W., JoChel, K. H., Drescher, K. and Timm, J. (1993). Logistic regression of case-control studies under validation sampling. *Biometrika* **80**, 339-352.

138. Schill, W. and Drescher, K. (1997). Logistic analysis of studies with two stage sampling: A comparison of four approaches. *Statistics in Medicine* **16**, 117-132.

139. Scoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239-241.

140. Sempos, C.T., Looker, A.C. Gillum, R.F. and Mcgee, D.L., (2000). Serum ferritin and death from all causes and cardiovascular disease: The NHANES II mortality study. *Annals of Epidemiology* **10**, 441-448.

141. Shahar, D.R. et al. (2001). Seasonal variations in dietary intake affect the consistency of dietary assessment. *European Journal of Epidemiology* **17**, 129-133.

142. Shen, S.M. and Lai, Y.L. (2000). Handling incomplete quality of life data. *Social Indicators Research* **55**, 121-166.

143. Skarin, A.T. and Herbst, R.S. (2001). Lung cancer in patients under age 40. *Lung Cancer* **32**, 255-264.

144. Simon, G. A. and Simonoff, J. S. (1986). Diagnostic plots for missing data in least squares regression. *Journal of the American Statistical Association* **81**, 501-509.

145. Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B* **55**, 3-23.

146. Smith-Warner, S.A. and Spiegelman, D., (1998). Alcohol and breast cancer in women: a pooled analysis of cohort studies. *Journal of the American Medical Association* **279**, 535-540.

147. Stampfer, M.J., Hu, F.B., Manson, J.E., Rimm, E.B., et al. (2000). Primary prevention of coronary heart disease in women through diet and lifestyle. *New England Journal of Medicine* **343**, 16-22.

148. Stata Corporation (2000). *Statistical Software: Release 7.0.* College Station, TX: Corporation.

149. Streiner, D.L. (2002). The case of the missing data: Methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry - Revue Canadienne de Psychiatrie.*

150. Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communication, Part B - Simulation Computation* **B5**, 55-64.

151. Sweeney, C., Blair C.K., Anderon, K.E., Lazovich, D. (2004). Risk factors of breast cancer in elderly women. *American Journal of Epidemiology* **160**, 868-875.

152. Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augumentation (with discussion). *Journal of the American Statistical Association* **82**, 528-550.

153. Tjonneland A., Thomsen B.L., Stripp C., Christensen J, Overvad K, et. al. (2003). Alcohol intake, drinking patterns and risk of postmenopausal breast cancer in Denmark: a prospective cohort study. *Cancer Causes and Controls* **14**, 277-284.

154. Touloumi, G., Babiker, A.G., Kenward, M.G. Pocock,S.J. et al. (2003). A comparison of two methods for the estimation of precision with incomplete longitudinal data, jointly modelled with a time-to-event outcome. *Statistics in Medicine* **22**, 3161-3175.

155. Turrell, G., Patterson, C., Oldenburg, B., Gould, T., et al. (2003). The socio-economic pattering of survey participation and non-response error in a multilevel study of food purchasing behaviour: Area- and individual level characteristics. *Public Health Nutrition* **6**, 181-189.

156. Van Buuren, S., Boshuizen, H.C., and Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 18, 681-694.

157. Van Buuren, S. and Oudshoorn, C.G.M. (1999). *Flexible Multivariate Imputation by MICE*. Leiden: TNO Preventie en Gezondheid.

158. Voss, S., Kroke. A., Klipstein-Grobusch, K. and Boeing, H. (1998). Is macronutrients composition of dietary intake data affected by underreporting? Results from the EPIC-Potsdam study. *European Journal of Clinical Nutrition* **52**, 119-26.

159. WHO/FAO (2003). *Diet, Nutrition and the Prevention of Chronic Diseases.* Geneva: WHO.

160. Wilson, A.C.; Forsyth, J.S.; Greene, S.A.; Irvine, L. et al. (1998). Relation of infant diet to childhood health: seven year follow up of cohort of children in Dundee infant feeding study. *British Medical Journal* **316**, 21-25.

161. Wolk, A., Ljung, H., Vessby, B., Hunter, D. et al., (1998). Effect of additional questions about fat on the validity of fat estimates from a food frequency questionnaire. Study Group of MRS SWEA. *European Journal of Clinical Nutrition* **52**, 186-192.

162. Wood, A.M., White, I.R., Thompson, S.G. (2004). Are missing outcome data adequately handled? A review of published randomised controlled trials in major medical journals. *Clinical trials* **1**, 368-376.

163. Woodbury, M.A. (1971). Discussion of paper by Hartley and Hocking. *Biometrics* **27**, 808-817.

164. Wu, H. and Wu, L. (2001). A multiple imputation method for missing covariates in non-linear mixed effects models with application to HIV dynamics. *Statistics in Medicine* **20**, 1755-1769.

165. Xie, F. and Paik and M.C. (1997). Multiple imputation methods for the missing covariates in generalized estimating equation. *Biometrics* **53**, 1538-1546.

166. Zhao, L.P., Lipsitz, S. and Lew, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics* **52**, 1165-1182.

APPENDIX A

Questionnaire: The UK Women's Nutrition and Lifestyle Survery

Please estimate how often you eat the following foods, and please answer every question.

PLEASE PUT A TICK(✓) ON EVERY LINE

FOODS AND AMOUNTS	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS?									
	NEVER	Less than once a month	1-3 per month	once a week	2-4 per week	5-6 per week	once per day	2-3 per day	4-5 per day	6+ per day
BREAD/SAVOURY BISCUITS										
White bread & rolls	0	1	2	3	4	5	6	7	8	9
Brown bread & rolls	0	1	2	3	4	5	6	7	8	9
Wholemeal bread & rolls	0	1	2	3	4	5	6	7	8	9
Chapatis, Nan, Paratha	0	1	2	3	4	5	6	7	8	9
Papadums	0	1	2	3	4	5	6	7	8	9
Tortillas	0	1	2	3	4	5	6	7	8	9
Pitta Bread	0	1	2	3	4	5	6	7	8	9
Crispbread e.g. Ryvita	0	1	2	3	4	5	6	7	8	9
Cream crackers, cheese biscuits	0	1	2	3	4	5	6	7	8	9
BREAKFAST CEREALS										
Porridge, Readybrek	0	1	2	3	4	5	6	7	8	9
Sugar coated cereals e.g. Sugar Puffs	0	1	2	3	4	5	6	7	8	9
Non-sugar coated cereals e.g. Cornflakes, Rice Krispies	0	1	2	3	4	5	6	7	8	9
Muesli	0	1	2	3	4	5	6	7	8	9
All Bran, Bran Flakes	0	1	2	3	4	5	6	7	8	9
Weetabix, Shredded Wheat	0	1	2	3	4	5	6	7	8	9
POTATOES, RICE & PASTA										
Potatoes e.g. boiled, mashed	0	1	2	3	4	5	6	7	8	9
Chips	0	1	2	3	4	5	6	7	8	9
Jacket Potato	0	1	2	3	4	5	6	7	8	9
Roast Potatoes	0	1	2	3	4	5	6	7	8	9
Potato Salad	0	1	2	3	4	5	6	7	8	9
White Pasta e.g. Spaghetti, Green Pasta, Red Pasta, Noodles	0	1	2	3	4	5	6	7	8	9
Wholemeal Pasta, Brown Spaghetti	0	1	2	3	4	5	6	7	8	9
White Rice	0	1	2	3	4	5	6	7	8	9
Brown Rice	0	1	2	3	4	5	6	7	8	9
Wild Rice	0	1	2	3	4	5	6	7	8	9
Macaroni Cheese	0	1	2	3	4	5	6	7	8	9

Please estimate how often you eat the following foods, and please answer every question

PLEASE PUT A TICK(✓) ON EVERY LINE

FOODS AND AMOUNTS	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS?									
	NEVER	Less than once a month	1-3 per month	once a week	2-4 per week	5-6 per week	once per day	2-3 per day	4-5 per day	6+ per day
DAIRY & NON-DAIRY PRODUCTS										
Thick & Creamy Yoghurt (125g carton)	0	1	2	3	4	5	6	7	8	9
Low fat Yoghurt (125g carton)	0	1	2	3	4	5	6	7	8	9
Diet Yoghurt (125g carton)	0	1	2	3	4	5	6	7	8	9
Greek Yoghurt (125g carton)	0	1	2	3	4	5	6	7	8	9
Fromage Frais/Creme Fraiche (125g carton)	0	1	2	3	4	5	6	7	8	9
Dairy Desserts (125g carton)	0	1	2	3	4	5	6	7	8	9
Single/Sour Cream (tablespoon)	0	1	2	3	4	5	6	7	8	9
Double/Clotted Cream (tablespoon)	0	1	2	3	4	5	6	7	8	9
Icecream	0	1	2	3	4	5	6	7	8	9
Milk Puddings	0	1	2	3	4	5	6	7	8	9
Low-fat Cheese	0	1	2	3	4	5	6	7	8	9
Cheese e.g. Cheddar, Brie, Edam	0	1	2	3	4	5	6	7	8	9
Cottage Cheese	0	1	2	3	4	5	6	7	8	9
Cheese and Onion Pastie	0	1	2	3	4	5	6	7	8	9
Soya Cheese	0	1	2	3	4	5	6	7	8	9
Soya Yoghurt	0	1	2	3	4	5	6	7	8	9
MARGARINES/BUTTERS & SPREADS										
Butter (enough for 1 slice of bread)	0	1	2	3	4	5	6	7	8	9
Block Margarine e.g. Stork, Krona, NOT in tub (enough for 1 slice of bread)	0	1	2	3	4	5	6	7	8	9
Polyunsaturated Margarine e.g. Flora, Sunflower, Granose, in tub (enough for 1 slice of bread)	0	1	2	3	4	5	6	7	8	9
Other soft Margarine, Dairy spreads e.g. Blue Band, Clover, in tub (enough for 1 slice of bread)	0	1	2	3	4	5	6	7	8	9
Low fat spread e.g. Outline, Gold, Flora Lite, in tub (enough for 1 slice of bread)	0	1	2	3	4	5	6	7	8	9
Very low fat spread, in tub e.g. St Ivel Lowest Fat Spread (enough for 1 slice of bread)	0	1	2	3	4	5	6	7	8	9
Monounsaturated Margarine eg Mono, Olivio (enough for 1 slice of bread)	0	1	2	3	4	5	6	7	8	9

Please estimate how often you eat the following foods, and please answer every question.

PLEASE PUT A TICK(✓) ON EVERY LINE

FOODS AND AMOUNTS	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS?									
	NEVER	Less than once a month	1-3 per month	once a week	2-4 per week	5-6 per week	once per day	2-3 per day	4-5 per day	6+ per day
SPREADS										
Marmite/Bovril/Vegemite	0	1	2	3	4	5	6	7	8	9
Peanut Butter	0	1	2	3	4	5	6	7	8	9
Chocolate/Chocolate & Nut Spread	0	1	2	3	4	5	6	7	8	9
Jam/Marmalade	0	1	2	3	4	5	6	7	8	9
Honey	0	1	2	3	4	5	6	7	8	9
Vegetable pâté	0	1	2	3	4	5	6	7	8	9
Nut Pâté	0	1	2	3	4	5	6	7	8	9
SAUCES & SOUPS										
Low Calorie Salad Cream (tablespoon)	0	1	2	3	4	5	6	7	8	9
Mayonnaise, Salad Cream Type Dressing (tablespoon)	0	1	2	3	4	5	6	7	8	9
French Type Dressing (tablespoon)	0	1	2	3	4	5	6	7	8	9
Sauces e.g. white/cheese/'Cook In'/curry	0	1	2	3	4	5	6	7	8	9
Tomato Ketchup (tablespoon)	0	1	2	3	4	5	6	7	8	9
Pickles/Chutney/Pesto sauce	0	1	2	3	4	5	6	7	8	9
Packet Soups - Meat & Veg (Bowl)	0	1	2	3	4	5	6	7	8	9
Other - Vegetable Soups (Bowl)	0	1	2	3	4	5	6	7	8	9
Other - Meat Soups (Bowl)	0	1	2	3	4	5	6	7	8	9
Low Calorie Soups (Bowl)	0	1	2	3	4	5	6	7	8	9
GRAINS (Medium serving)										
Barley	0	1	2	3	4	5	6	7	8	9
Oats	0	1	2	3	4	5	6	7	8	9
Bulgar Wheat	0	1	2	3	4	5	6	7	8	9
Wheat Germ (tablespoon)	0	1	2	3	4	5	6	7	8	9
Cous-cous	0	1	2	3	4	5	6	7	8	9
White Rice	0	1	2	3	4	5	6	7	8	9
Brown Rice	0	1	2	3	4	5	6	7	8	9
NUTS & SEEDS										
Peanuts/Pistachio Nuts	0	1	2	3	4	5	6	7	8	9
Cashew Nuts & Almonds	0	1	2	3	4	5	6	7	8	9
Pecan Nuts/Walnuts	0	1	2	3	4	5	6	7	8	9
Sunflower Seeds/ Sesame Seeds	0	1	2	3	4	5	6	7	8	9

Please estimate how often you eat the following foods, and please answer every question.

PLEASE PUT A TICK(✓) ON EVERY LINE

FOODS AND AMOUNTS	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS?									
	NEVER	Less than once a month	1-3 per month	once a week	2-4 per week	5-6 per week	once per day	2-3 per day	4-5 per day	6+ per day
PULSES (Include when used in recipes)										
Lentils, dals	0	1	2	3	4	5	6	7	8	9
Chick Peas, Chanas	0	1	2	3	4	5	6	7	8	9
Hummus	0	1	2	3	4	5	6	7	8	9
Baked beans	0	1	2	3	4	5	6	7	8	9
Mung Beans & Red Kidney Beans	0	1	2	3	4	5	6	7	8	9
Bean Sprouts	0	1	2	3	4	5	6	7	8	9
Black Eyed Beans	0	1	2	3	4	5	6	7	8	9
Butter Beans/Broad Beans	0	1	2	3	4	5	6	7	8	9
EGGS/EGG DISHES										
Boiled/ Poached egg	0	1	2	3	4	5	6	7	8	9
Omelette, Scrambled egg	0	1	2	3	4	5	6	7	8	9
Fried egg	0	1	2	3	4	5	6	7	8	9
Quiche	0	1	2	3	4	5	6	7	8	9
VEGETABLE DISHES										
Quorn	0	1	2	3	4	5	6	7	8	9
Textured vegetable protein/ Sosmix/burger mix/soya sausages	0	1	2	3	4	5	6	7	8	9
Vegetarian Chilli/Vegetable Curry	0	1	2	3	4	5	6	7	8	9
Mixed Bean Casserole/Ratatouille	0	1	2	3	4	5	6	7	8	9
Stir-fry vegetables	0	1	2	3	4	5	6		8	9
Vegetable - Lasagne/Moussaka/Ravioli/ filled pasta with sauce	0	1	2	3	4	5	6	7	8	9
Vegetable Pizza	0	1	2	3	4	5	6	7	8	9
MEAT										
Beef e.g. roast, steak	0	1	2	3	4	5	6	7	8	9
Beef Stew/Casserole/Mince/Curry	0	1	2	3	4	5	6	7	8	9
Beefburger/Hamburger	0	1	2	3	4	5	6	7	8	9
Pork e.g. Roast, Chops, Slices	0	1	2	3	4	5	6	7	8	9
Pork Stew/Casserole	0	1	2	3	4	5	6	7	8	9
Lamb e.g. Roast, Chops	0	1	2	3	4	5	6	7	8	9
Lamb Stew/Casserole	0	1	2	3	4	5	6	7	8	9

Please estimate how often you eat the following foods, and please answer every question.

PLEASE PUT A TICK(✓) ON EVERY LINE

FOODS AND AMOUNTS	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS?									
	NEVER	Less than once a month	1-3 per month	once a week	2-4 per week	5-6 per week	once per day	2-3 per day	4-5 per day	6+ per day
OTHER MEATS										
Chicken/Turkey roast, slices	0	1	2	3	4	5	6	7	8	9
Breadcrumbed e.g.chicken nuggets/kievs	0	1	2	3	4	5	6	7	8	9
Chicken/Turkey in creamy sauce, curry	0	1	2	3	4	5	6	7	8	9
Bacon	0	1	2	3	4	5	6	7	8	9
Ham	0	1	2	3	4	5	6	7	8	9
Corned Beef, Spam, Luncheon Meats	0	1	2	3	4	5	6	7	8	9
Sausages e.g. Beef Pork	0	1	2	3	4	5	6	7	8	9
Pies/Pasties/Sausage Rolls	0	1	2	3	4	5	6	7	8	9
Offal e.g. Liver, Kidney	0	1	2	3	4	5	6	7	8	9
Liver Pâté/Sausage, Salami	0	1	2	3	4	5	6	7	8	9
Meat - Lasagne/Moussaka/Ravioli/ filled pasta with sauce	0	1	2	3	4	5	6	7	8	9
Meat Pizza	0	1	2	3	4	5	6	7	8	9
FISH										
Fish fingers/cakes	0	1	2	3	4	5	6	7	8	9
Fried fish in batter (as in fish and chips)	0	1	2	3	4	5	6	7	8	9
White fish e.g. Cod, Haddock, Plaice, Sole, Halibut (fresh or frozen)	0	1	2	3	4	5	6	7	8	9
Oily fish e.g. Mackerel, Kippers, Tuna, Salmon, Sardines, Herring	0	1	2	3	4	5	6	7	8	9
Shellfish e.g. Crab, Prawns, Mussels	0	1	2	3	4	5	6	7	8	9
Fish Roe, Taramasalata	0	1	2	3	4	5	6	7	8	9
Fish Pie/Fish Lasagne	0	1	2	3	4	5	6	7	8	9
VEGETABLES										
Beetroot	0	1	2	3	4	5	6	7	8	9
Broccoli, Spring Greens, Kale	0	1	2	3	4	5	6	7	8	9
Brussel Sprouts	0	1	2	3	4	5	6	7	8	9
Cabbage	0	1	2	3	4	5	6	7	8	9
Carrots	0	1	2	3	4	5	6	7	8	9
Cauliflower	0	1	2	3	4	5	6	7	8	9
Celery	0	1	2	3	4	5	6	7	8	9

Please estimate how often you eat the following foods, and please answer every question

PLEASE PUT A TICK(✓) ON EVERY LINE

FOODS AND AMOUNTS	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS?									
	NEVER	Less than once a month	1-3 per month	once a week	2-4 per week	5-6 per week	once per day	2-3 per day	4-5 per day	6+ per day
VEGETABLES (continued)										
Coleslaw	0	1	2	3	4	5	6	7	8	9
Low-calorie Coleslaw	0	1	2	3	4	5	6	7	8	9
Courgettes, Marrow, Squash	0	1	2	3	4	5	6	7	8	9
Cucumber	0	1	2	3	4	5	6	7	8	9
Garlic	0	1	2	3	4	5	6	7	8	9
Green Beans, Runner Beans	0	1	2	3	4	5	6	7	8	9
Leeks	0	1	2	3	4	5	6	7	8	9
Lettuce	0	1	2	3	4	5	6	7	8	9
Mushrooms	0	1	2	3	4	5	6	7	8	9
Aubergine, Okra/Ladies Finger	0	1	2	3	4	5	6	7	8	9
Olives	0	1	2	3	4	5	6	7	8	9
Parsnips	0	1	2	3	4	5	6	7	8	9
Peas, Mushy peas, Mange-tout	0	1	2	3	4	5	6	7	8	9
Peppers - Red, Green, Yellow, Black etc	0	1	2	3	4	5	6	7	8	9
Swedes	0	1	2	3	4	5	6	7	8	9
Sweetcorn	0	1	2	3	4	5	6	7	8	9
Tomatoes - raw/canned/sauce	0	1	2	3	4	5	6	7	8	9
Turnip	0	1	2	3	4	5	6	7	8	9
Watercress, Mustard & Cress	0	1	2	3	4	5	6	7	8	9
FRUIT										
Apples	0	1	2	3	4	5	6	7	8	9
Avocado	0	1	2	3	4	5	6	7	8	9
Bananas	0	1	2	3	4	5	6	7	8	9
Grapes	0	1	2	3	4	5	6	7	8	9
Kiwi Fruit	0	1	2	3	4	5	6	7	8	9
Mangoes	0	1	2	3	4	5	6	7	8	9
Oranges, Satsumas, Grapefruit, etc	0	1	2	3	4	5	6	7	8	9
Papaya	0	1	2	3	4	5	6	7	8	9
Pears	0	1	2	3	4	5	6	7	8	9
Pineapple	0	1	2	3	4	5	6	7	8	9

lease estimate how often you eat the following foods, and please answer every question.

LEASE PUT A TICK(✓) ON EVERY LINE

FOODS AND AMOUNTS	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS?									
	NEVER	Less than once a month	1-3 per month	once a week	2-4 per week	5-6 per week	once per day	2-3 per day	4-5 per day	6+ per day
SEASONAL FRUIT **How often have you eaten these fruits, when they are in season?**										
Apricots	0	1	2	3	4	5	6	7	8	9
Melon	0	1	2	3	4	5	6	7	8	9
Nectarines	0	1	2	3	4	5	6	7	8	9
Peaches	0	1	2	3	4	5	6	7	8	9
Plums	0	1	2	3	4	5	6	7	8	9
Raspberries	0	1	2	3	4	5	6	7	8	9
Red currants/Black currants	0	1	2	3	4	5	6	7	8	9
Rhubarb	0	1	2	3	4	5	6	7	8	9
Strawberries	0	1	2	3	4	5	6	7	8	9
DRIED FRUIT										
Dates	0	1	2	3	4	5	6	7	8	9
Figs	0	1	2	3	4	5	6	7	8	9
Prunes	0	1	2	3	4	5	6	7	8	9
Mixed Dried Fruit e.g. Apricots, Apples, Pears, Mangoes	0	1	2	3	4	5	6	7	8	9
Currants, Raisins, Sultanas	0	1	2	3	4	5	6	7	8	9
SWEET SNACKS										
Cereal Bars/Flapjacks (one)	0	1	2	3	4	5	6	7	8	9
Fruit bars (one) eg Apricot, Date	0	1	2	3	4	5	6	7	8	9
Chocolate Snack Bars e.g. Mars, Crunchie (1 bar)	0	1	2	3	4	5	6	7	8	9
Mini chocolate snack bars, Chocolates - singles or squares (1)	0	1	2	3	4	5	6	7	8	9
Boiled Sweets, Toffees, Mints	0	1	2	3	4	5	6	7	8	9
SAVOURY SNACKS										
Crisps (1 bag)	0	1	2	3	4	5	6	7	8	9
Other fried snacks e.g. Wotsits (1 bag)	0	1	2	3	4	5	6	7	8	9
Low fat or baked snacks e.g. Low-fat Crisps (1 bag)	0	1	2	3	4	5	6	7	8	9
Bombay Mix (small handful)	0	1	2	3	4	5	6	7	8	9
Peanuts/Pistachio Nuts (small handful)	0	1	2	3	4	5	6	7	8	9
Mixed Nuts and Raisins (small handful)	0	1	2	3	4	5	6	7	8	9

Please estimate how often you eat the following foods, and please answer every question.

PLEASE PUT A TICK(✓) ON EVERY LINE

FOODS AND AMOUNTS	HOW OFTEN HAVE YOU EATEN THESE FOODS IN THE LAST 12 MONTHS?									
	NEVER	Less than once a month	1-3 per month	once a week	2-4 per week	5-6 per week	once per day	2-3 per day	4-5 per day	6+ per day
BEVERAGES										
Tea (cup)	0	1	2	3	4	5	6	7	8	9
Herbal Tea (cup)	0	1	2	3	4	5	6	7	8	9
Coffee - instant/ground (cup)	0	1	2	3	4	5	6	7	8	9
Coffee - decaffeinated (cup)	0	1	2	3	4	5	6	7	8	9
Coffee substitute e.g Caro/Bambu (cup)	0	1	2	3	4	5	6	7	8	9
Coffee whitener (teaspoon)	0	1	2	3	4	5	6	7	8	9
Cocoa, Hot Chocolate (cup)	0	1	2	3	4	5	6	7	8	9
Horlicks, Ovaltine (cup)	0	1	2	3	4	5	6	7	8	9
Low Calorie/Low-fat Horlicks, Ovaltine, Hot Chocolate, (cup)	0	1	2	3	4	5	6	7	8	9
Orange Juice (Pure Fruit) (glass)	0	1	2	3	4	5	6	7	8	9
Other -(100%) Pure fruit juices (glass)	0	1	2	3	4	5	6	7	8	9
Fruit Squash/Cordial - diluted (glass)	0	1	2	3	4	5	6	7	8	9
Fizzy soft drinks e.g. Coke, Lemonade (glass/can)	0	1	2	3	4	5	6	7	8	9
Low Calorie/Diet Soft Drinks (glass/can)	0	1	2	3	4	5	6	7	8	9
ALCOHOLIC BEVERAGES										
Wines (wineglassful)	0	1	2	3	4	5	6	7	8	9
Beer, Lager (half pint)	0	1	2	3	4	5	6	7	8	9
Cider (half pint)	0	1	2	3	4	5	6	7	8	9
Port, Sherry, Liqueurs (glass)	0	1	2	3	4	5	6	7	8	9
Spirits e.g. Whisky, Gin, Vodka, Brandy (single/1 measure)	0	1	2	3	4	5	6	7	8	9
BISCUITS, SWEETS & PUDDINGS										
Plain Biscuits e.g. Marie, Nice, Digestive (one)	0	1	2	3	4	5	6	7	8	9
Chocolate Biscuits (one)	0	1	2	3	4	5	6	7	8	9
Sandwich/Cream Biscuits (one)	0	1	2	3	4	5	6	7	8	9
Fruitcake (1 slice)	0	1	2	3	4	5	6	7	8	9
Sponge cakes (1 slice)	0	1	2	3	4	5	6	7	8	9
Buns/Pastries e.g. Croissants, Doughnuts, Tray Bakes, (one)	0	1	2	3	4	5	6	7	8	9
Scones/Pancakes/Muffins/Crumpets (1)	0	1	2	3	4	5	6	7	8	9
Fruit Pies, Tarts, Crumbles, (1 slice)	0	1	2	3	4	5	6	7	8	9
Sponge Puddings (1 serving)	0	1	2	3	4	5	6	7	8	9

1: Other Foods

Are there any other foods which you eat more than once a week? Yes ☐ 1 No ☐ 2

If yes, please list below

Food	Usual serving size	Number of times eaten each week

2: Would you describe yourself as a vegetarian? Yes ☐ 1 No ☐ 2

If yes, how long have you been vegetarian? ☐☐ years.

Would you describe yourself as a vegan? Yes ☐ 1 No ☐ 2

If yes, how long have you been vegan? ☐☐ years.

3: Do you use herbs and spices at least once per week when cooking food? Yes ☐ 1 No ☐ 2

Which fresh herbs and spices would you use at least once a week? Please list here

Which dried herbs and spices would you use at least once a week? Please list here

PORTION SIZE:

4: Compared to other people would you describe your typical average portion size of foods as?

Small? ☐ 1 Medium? ☐ 2 Large? ☐ 3

PULSES:

5: Do you eat pulses e.g. beans, peas, lentils etc. Yes ☐ 1 No ☐ 2

If no, please go to question 7.

6: Can you please indicate how much of the Pulses you eat are Fresh, Frozen, Canned or Dried. Please tick the appropriate boxes, e.g. 1/4 Dried, 3/4 Frozen.

	Proportion				
	Never	1/4	1/2	3/4	All
Fresh	1	2	3	4	5
Frozen	1	2	3	4	5
Dried	1	2	3	4	5
Canned	1	2	3	4	5

How do you usually cook pulses? Tick all applicable.

Steaming/Boiling/Pressure Cooking ☐ 1

Microwaving ☐ 2

Roasting ☐ 3

Stewing/Casseroling/Baking ☐ 4

Stir Frying/Frying ☐ 5

Raw/soaked/Raw-sprouted ☐ 6

VEGETABLES:

7: How many servings of vegetables or vegetable containing dishes, (excluding potatoes) do you usually eat each week? ☐☐

8: Can you please indicate how much of the vegetables you eat are Fresh, Frozen, Canned or Dried. Please tick the appropriate boxes, e.g. 1/4 Dried, 3/4 Frozen.

	Proportion				
	Never	1/4	1/2	3/4	All
Fresh	1	2	3	4	5
Frozen	1	2	3	4	5
Dried	1	2	3	4	5
Canned	1	2	3	4	5

9: Do you ever eat raw vegetables apart from salad vegetables? Yes ☐ 1 No ☐ 2

10: How do you usually cook your vegetables.? (Excluding potatoes). Tick more than one box if necessary.

Boiling ☐
Steaming ☐
Grilling/Barbecuing/Baking/Roasting (Cooked dry or using a small amount of oil) ☐
Stir Frying/Frying/Sauté ☐
Microwaving ☐
Deep frying - including in batter ☐
Casseroling//Baking in sauce ☐
Other ☐

Please describe

FRUIT:

11: How many servings of fruit or fruit containing dishes do you usually eat each week? ☐☐

Can you please indicate how much of the fruit you eat is Fresh, Canned, Dried or Stewed.

Please tick the appropriate boxes e.g. 1/4 Fresh, 3/4 Canned

	Proportion				
	Never	1/4	1/2	3/4	All
Fresh	1	2	3	4	5
Stewed	1	2	3	4	5
Dried	1	2	3	4	5
Canned	1	2	3	4	5

12: Do you ever cook the fruit you eat? Yes ☐ 1 No ☐ 2

13: If so, how do you usually cook your fruit?

Stewing ☐ 1
Poaching/Steaming ☐ 2
Baking ☐ 3
Microwaving ☐ 4
Other ☐ 5

Please describe

MEAT: (If you **never** eat meat please go to question 16)

14: How many servings of meat or meat containing dishes do you usually eat each week? ☐☐

What do you do with the visible fat on your meat?

Eat all/most of the fat ☐ 1

Eat some of the fat ☐ 2

Eat as little as possible/none ☐ 3

15: How do you usually cook meat? Tick more than one box if necessary.

Grilling/Barbecuing/Baking/Roasting (Cooked dry or using a small amount of oil) ☐ 1

Stir Frying/Frying ☐ 2

Microwaving ☐ 3

Deep frying - including in batter ☐ 4

Casseroling/Baking in sauce ☐ 5

Other ☐ 6

Please describe

16: How many servings of fish or fish containing dishes do you usually eat each week? ☐☐

How do you usually cook fish. Tick more than one box if necessary.

Boiling ☐ 1

Steaming ☐ 2

Grilling/Barbecuing/Baking/Roasting (Cooked dry or using a small amount of oil) ☐ 3

Stir Frying/Frying ☐ 4

Microwaving ☐ 5

Deep frying - including in batter ☐ 6

Casseroling//Baking in sauce ☐ 7

Other ☐ 8

Please describe

MILK:

17. What type of milk do you use most often? Select one only

Full cream (Silver Top) ☐ 1

Semi-skimmed (Red/White Top) ☐ 2

Skimmed/fat free ☐ 3

Channel Islands (Gold Top) ☐ 4

Dried Milk ☐ 5

Soya ☐ 6

Sterilised ☐ 7

None ☐ 8

Other ☐ 9 Specify

If you used soya milk, please describe brand and type

18: How much milk do you drink each day, including milk with tea, coffee, milky drinks, cereals etc?

None ☐ 1

¼ Pint ☐ 2

½ Pint ☐ 3

¾ Pint ☐ 4

1 Pint ☐ 5

More than 1 Pint ☐ 6

BREAKFAST:

Q19: Are there any breakfast cereals that you normally eat that were not mentioned earlier? Yes ☐ 1 No ☐ 2

If yes, which brand and type of breakfast cereal, do you usually eat?

List the types most often used

Brand	Type

Q20: Do you usually take sugar on your breakfast cereal? Yes ☐ 1 No ☐ 2

If yes, how many teaspoons? ☐☐ teaspoons

Q21: Do you usually take sugar/honey in tea, herbal tea, coffee or coffee substitute? Yes ☐ 1 No ☐ 2

If Yes, please write the number of teaspoons per cup.

Sugar/honey in tea ☐☐ teaspoons

Sugar/honey in herbal tea ☐☐ teaspoons

Sugar/honey in coffee ☐☐ teaspoons

Sugar/honey in coffee substitute ☐☐ teaspoons

Do you use sweeteners instead of sugar or honey, Yes ☐ 1 No ☐ 2

Which brand of sweetener do you use, please specify

If yes how many tablets per day, or how many teaspoons of powder per day? ☐☐

Q22: On days when you eat bread, how many slices of bread or rolls do you eat ? ☐☐ slices/rolls per day

USE OF FATS:

Q23: Do you usually spread butter/margarine on your bread?

Yes ☐ 1 No ☐ 2 Sometimes ☐ 3

How many slices of bread/rolls/crackers do you have with spread each day? ☐☐

How much spread do you use?

Just a scrape/thinly spread ☐ 1

medium ☐ 2

Thickly spread ☐ 3

Q24: What kind of fat do you most often use for frying, roasting, grilling etc?

Tick more than one if applicable

Butter ☐ 1

Lard/Dripping ☐ 2

Vegetable Oil ☐ 3

Solid White Vegetable Fat ☐ 4

Margarine ☐ 5

None ☐ 6

If you used vegetable oil, or margarine, please give type e.g. corn, sunflower

25: What kind of fat do you most often use for baking cakes etc.? Tick more than one if applicable

Butter ☐ 1 Solid White Vegetable Fat ☐ 2

Lard/Dripping ☐ 3 Margarine ☐ 4

Vegetable Oil ☐ 5 None ☐ 6

If you use margarine, please give Brand e.g. Flora, Stork

USE OF SALT:

26: How often do you add salt to food while cooking?

Always ☐ 1 Usually ☐ 2

Sometimes ☐ 3 Rarely ☐ 4

Never ☐ 5

27: How often do you add salt to any food at the table?

Always ☐ 1 Usually ☐ 2

Sometimes ☐ 3 Rarely ☐ 4

Never ☐ 5

28: Do you regularly use a salt substitute (e.g. Losalt)? Yes ☐ 1 No ☐ 2

If yes, which brand?

USE OF SUPPLEMENTS:

29: Do you take any vitamins, minerals, fish oils, fibre or other food supplements? Yes ☐ 1 No ☐ 2

If yes, please fill in details below.

Name and Brand of Supplements	How much do you take at a time	How often do you take these? Daily	Weekly	Monthly	Less often
		1	2	3	4
		1	2	3	4
		1	2	3	4

SPECIAL DIETS:

30: I) Have you changed your diet over the last 12 months? Yes ☐ 1 No ☐ 2

If yes, please indicate if the change was for any of the reasons listed below?

Tick more than one box if applicable

High Blood Pressure ☐ 0 Stomach problems (e.g. ulcer or gastritis) ☐ 1

Bowel problems (e.g. irritable bowel or diverticulitis) ☐ 2 Concern over eating a healthy Diet ☐ 3

Concern over a family history of illness ☐ 4 High Blood Cholesterol/Lipids ☐ 5

Overweight/Obesity ☐ 6 Diabetes ☐ 7

Allergies (e.g. skin rash) ☐ 8 Other ☐ 9

Specify

II) Describe below how your diet has changed

Do you currently follow any of these diets? Tick more than one box if neccessary.

Low fat ☐ 1 Low salt ☐ 2 Diabetic ☐ 3

Slimming ☐ 4 Gluten free ☐ 5 High fibre ☐ 6

Other ☐ 7 Please give details

CONSUMPTION OF ALCOHOL:

31: How often, if ever do you drink alcohol?

More than once a week ☐ 1 Once a week ☐ 2

Less than once a week ☐ 3 Never drink alcohol ☐ 4

32: In a typical week, how much do you drink?

Beer or cider ☐☐ pints each week

Wine ☐☐ glasses each week

Sherry/Fortified Wines ☐☐ glasses each week

Spirits ☐☐ glasses (singles) each week

33: Five years ago, how many alcoholic drinks did you have each week?

Beer or cider ☐☐ pints each week

Wine ☐☐ glasses each week

Sherry/Fortified Wines ☐☐ glasses each week

Spirits ☐☐ glasses (singles) each week

SMOKING:

34: Which one of the following best describes you?

I smoke every day ☐ 1 I smoke occasionally, but not every day ☐ 2

I used to smoke every day, but do not smoke at all now ☐ 3 I have never smoked ☐ 4

IF you have never smoked, please go to question 37.

35: Do/did you smoke?

Cigarettes ☐ 1

Cigars ☐ 2

A combination of the above ☐ 3

If you currently smoke or used to smoke cigarettes how many do/did you smoke each day?

☐☐ cigarettes

If you currently smoke or used to smoke cigarettes which brand of cigarettes do/did you usually smoke?

36: If you have stopped smoking for what period of time have you been a non-smoker?

1 year or less ☐ 1 2-5 years ☐ 2

6-10 years ☐ 3 Over 10 years ☐ 4

SIZE:

37: Approximately how much did you weigh when you were born?

☐☐ lbs or ☐☐ Kg or Don't Know ☐

38: Approximately how much did you weigh when you were 20 years old?

☐☐ stones ☐☐ pounds or ☐☐☐ Kg or Don't Know ☐

39: Approximately how much do you weigh at present?

☐☐ stones ☐☐ pounds or ☐☐☐ Kg or Don't Know ☐

40: Have you lost more than half a stone in the last year? Yes ☐ 1 No ☐ 2

Have you gained more than half a stone in the last year?
(Please ignore weight gained during pregnancy.) Yes ☐ 1 No ☐ 2

41: What is your present waist size? ☐☐ inches or ☐☐☐ centimetres or Don't Know ☐ 1

42: What is your present hip size? ☐☐ inches or ☐☐☐ centimetres or Don't Know ☐ 1

43: What is your present height? ☐ ft ☐☐ inches or ☐☐☐ centimetres or Don't Know ☐ 1

44: What size of blouse do you wear? Size ☐☐

45: What size of skirt do you wear? Size ☐☐

PHYSICAL ACTIVITY

46: In a typical week during the last 12 months, how many hours did you spend on each of the following activities? Put "0" if none

Housework, such as cleaning, washing, cooking, child care		☐☐ hours ☐☐	minutes per week
Do-It-Yourself		☐☐ hours ☐☐	minutes per week
Gardening	In Summer	☐☐ hours ☐☐	minutes per week
	In Winter	☐☐ hours ☐☐	minutes per week
Walking, including to work, shopping & leisure	In Summer	☐☐ hours ☐☐	minutes per week
	In Winter	☐☐ hours ☐☐	minutes per week
Cycling, including to work & leisure	In Summer	☐☐ hours ☐☐	minutes per week
	In Winter	☐☐ hours ☐☐	minutes per week
Other physical exercise, such as keep-fit, aerobics, jogging, tennis, swimming	In Summer	☐☐ hours ☐☐	minutes per week
	In Winter	☐☐ hours ☐☐	minutes per week

47: In a normal week, do you do any of these activities vigorously enough to cause sweating or a faster heartbeat? Yes ☐ 1 No ☐ 2

If yes, for how long each week do you do such vigorous physical activity? ☐☐ hours ☐☐ minutes per week

ILLNESS:

48: Have you ever been told by a doctor that you have, or had, any of the following conditions? Please tick all which apply and give the age at which each condition was first diagnosed.

Heart attack, coronary thrombosis, myocardial infarction Yes ☐ 1 No ☐ 2 at age ☐☐ yrs old

Angina Yes ☐ 1 No ☐ 2 at age ☐☐ yrs old

Stroke Yes ☐ 1 No ☐ 2 at age ☐☐ yrs old

High Blood Pressure (Hypertension) Yes ☐ 1 No ☐ 2 at age ☐☐ yrs old

High Blood Cholesterol, Hyperlipidaemia Yes ☐ 1 No ☐ 2 at age ☐☐ yrs old

Diabetes Yes ☐ 1 No ☐ 2 at age ☐☐ yrs old

Gallstones Yes ☐ 1 No ☐ 2 at age ☐☐ yrs old

Polyps in the large intestine Yes ☐ 1 No ☐ 2 at age ☐☐ yrs old

Cancer Yes ☐ 1 No ☐ 2 at age ☐☐ yrs old

If yes, what type of cancer?

Any other illnesses or operations?

Do not include hysterectomy or breast surgery. These are covered later in the questionnaire.

Condition/ operation / disease	Age first diagnosed
	☐☐ yrs old
	☐☐ yrs old
	☐☐ yrs old
	☐☐ yrs old

49: Are you currently receiving long-term treatment for any illness or condition? Yes ☐ 1 No ☐ 2

If yes, please give details of treatment. If no please go to question 50:

Illness or condition	Treatment	Dose	Frequency

50: Have your mother and/or father ever suffered from cancer or heart attack/heart disease?

Yes ☐ 1 No ☐ 2 Don't Know ☐ 3

If yes, please give details

51: If you have brothers and/or sisters,have they ever suffered from cancer or heart attack/heart disease?

Yes ☐ 1 No ☐ 2 Don't Know ☐ 3

If yes, please describe details

EDUCATION:

52: How old were you when you finished your full time education? ☐☐ yrs old

53: Do you have any of the following qualifications? Tick all applicable

CSE	☐ 1	"A" Level, Highers	☐ 4
GCE "O" Level	☐ 2	Teaching diploma, HNC	☐ 5
City & Guilds	☐ 3	Degree	☐ 6
Other ☐ 7 describe		None of these	☐ 8

EMPLOYMENT:

54: Have you ever had a paid job? Yes ☐ 1 No ☐ 2 ☐

If yes, please answer for your current or most recent job

What is/was your job title?

What do/did you do in your job?

What does/did the organisation you work for make or do?

Are/were you a Manager? ☐ 1 Foreman/woman? ☐ 2

Supervisor? ☐ 3 None of these? ☐ 4

Are/were you self-employed? Yes ☐ 1 No ☐ 2

Do you have a paid job at present? Yes ☐ 1 No ☐ 2

If no, how would you describe yourself?

Housewife ☐ 1 Unemployed ☐ 2

Retired ☐ 3 Student ☐ 4

Other ☐ 5 describe _____

When did you last have paid employment 19☐☐ (year) or Never ☐ 1

S5: What is your marital status?

Married or living as married ☐ 1 Divorced ☐ 2

Widowed ☐ 3 Single ☐ 4

Separated ☐ 5

If you are not married or living as married, please go to question 57.

PARTNER'S EMPLOYMENT:

S6: If married or living as married, has your partner ever had a paid job? Yes ☐ 1 No ☐ 2 ☐

If yes, please answer for your partner's current or most recent job.

What is/was your partner's job title? _____

What does/did the organisation your partner works for make or do? _____

Is/was your partner a Manager? ☐ 1 Foreman/woman? ☐ 2

Supervisor? ☐ 3 None of these? ☐ 4

Is/was your partner self-employed? Yes ☐ 1 No ☐ 2

Does your partner have a paid job at present? Yes ☐ 1 No ☐ 2

If no, how would you describe your partner?

House-husband ☐ 1 Student ☐ 2

Unemployed ☐ 3 Retired ☐ 4

Other ☐ 5 describe _____

S7: Which of these groups would you consider you belong to?

White ☐ 1 Bangladeshi ☐ 2

Indian ☐ 3 Chinese ☐ 4

Pakistani ☐ 5 Black - Caribbean ☐ 6

Black - other ☐ 7 _____

Other ☐ 8 _____

MENSTRUAL & OBSTETRIC HISTORY:

S8: How old were you when you had your first menstrual period ☐☐ years old

S9: What is the usual length of your menstrual cycle?

(i.e. from the first day of one period to the first day of the next period e.g. 26 days)? ☐☐ days.

60: Have you ever been pregnant? Yes ☐ 1 No ☐ 2

Are you pregnant at the moment? Yes ☐ 1 No ☐ 2

How many times have you been pregnant? ☐☐

Have you ever had a miscarriage/still birth? Yes ☐ 1 No ☐ 2

If you have had children, please go to question 61. If not please go to question 63.

61: Have you had any children? Yes ☐ 1 No ☐ 2

If yes, how old were you when your first child was born ☐☐ years

If yes, how many children have you had? ☐☐ children

If none please go to question 63.

Please can you write in each child's sex and approximate birthweight.

Child	Sex of Child	Approximate Birthweight	Child's D.O.B
CHILD1:			19☐☐
CHILD 2:			19☐☐
CHILD 3:			19☐☐
CHILD 4:			19☐☐
CHILD 5:			19☐☐

62: Did you ever breast feed any of your children? Yes ☐ 1 No ☐ 2

If yes, for those children you breast-fed, please describe how long you continued breast feeding after each birth, (even only occasional breast feeding). Tick the appropriate box.

	1–6 days	1–4 weeks	1–3 months	4–6 months	6+ months	12+ months
CHILD 1:	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6
CHILD 2:	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6
CHILD 3:	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6
CHILD 4:	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6
CHILD 5:	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5	☐ 6

63: Have you ever seen a doctor because of fertility problems? Yes ☐ 1 No ☐ 2

If yes, has a doctor ever told you that you were infertile? Yes ☐ 1 No ☐ 2

64: Have you ever used oral contraceptives (the pill)? Yes ☐ 1 No ☐ 2

If yes, how old were you when you first started to use the pill? ☐☐ years old

For how long altogether did you use the pill? ☐☐ years

Are you currently using the pill? Yes ☐ 1 No ☐ 2

If no, how old were you when you last used it? ☐☐ years old

65: Have you ever used a coil or intra-uterine device (IUD)? Yes ☐ 1 No ☐ 2

If yes, do you have a coil or IUD at present? Yes ☐ 1 No ☐ 2

66: How many "natural" menstrual periods have you had in the last 12 months?
Do not count bleeding while using the pill or HRT (Hormone Replacement Therapy)

None ☐ 1
1 to 3 ☐ 2
4 to 5 ☐ 3
6 to 9 ☐ 4
10 or more ☐ 5
Not applicable because using the Pill or HRT or currently pregnant ☐ 6

67: When did you last have a "natural" menstrual period? Do not count bleeding while using the pill or HRT (Hormone Replacement Therapy). Record as fully as possible

Date: _____ or age ☐☐ years old Don't know ☐

68: Have you ever used HRT (Hormone Replacement Therapy) for menopause? Yes ☐ 1 No ☐ 2

If yes, how old were you when you first used HRT? ☐☐ years old

For how long altogether have you used HRT? years ☐☐ and months ☐☐

Are you currently using HRT? Yes ☐ 1 No ☐ 2

If no, how old were you when you last used HRT? ☐☐ years old

69: Have you had a hysterectomy ? If no please go to question 71.

Yes ☐ 1 Age at time of operation ☐☐ years old No ☐ 2 Don't know ☐ 3

70: Have you had an operation to remove one or both your ovaries?

Yes ☐ 1 No ☐ 2 Don't know ☐ 3

If yes, how old were you? ☐☐ years old

Were one or both ovaries removed? One ☐ 1 Both ☐ 2 Don't know ☐ 3

71: Have you ever had a breast biopsy (minor surgery to remove tissue from your breast for diagnostic purposes)? Yes ☐ 1 No ☐ 2 Don't know ☐ 3

If yes, how old were you (first occurrence)? ☐☐ years old

APPENDIX B

The Data Dictionary

UK Women's Cohort Study

Data Dictionary for Baseline Data.

N of Cases: 35374

Total No of Variables:599

Variable Information:

We have developed a scoring system for each of the variables . This indicates how clean the variable is .

For example:

***** 5 stars = Sweaky clean, we are confident that all variables are correct, & in the desirable frequency and that the question has been fully understood . No inconsistent values were observed.

**** 4 stars = Almost sweaky clean, although a there are a few extremes (possible extremes). <1% inconsistent values were observed.

*** 3 stars = There are noted extremes, some of which are feasible and so have been left, others have been recoded as –1 , or impossible . We have cross checked these

impossible values against other available variables prior to re-coding. 1-2 % inconsistent values were observed.

** 2 stars = There are quite a mixture of responses, quite a number of responses have been recoded as impossible. There is obviously a misunderstanding of the question, with other variables. 2-5 % inconsistent values were observed.

*1 star = Very poor. We are not confident that this question has been fully understood. There has been considerable recoding due to impossible variables. > 5% of inconsistent values were observed.

Data Dictionary

Section 1: Food Frequency Questionnaire (FFQ)

The following chapter is based on the ffq section of the questionnaire (p1-8). All ffq variables are denoted as "b-variables" and are split according to the food group as presented in the body of the questionnaire.

ID - identification number

BATCH

This is the batch number of the records received back from the data entry company. This variable will probably not be of any use.

61 = missing batch number.

GROUP

Group: Historical variable used for sampling

This should indicate what group the subject was first categorised into according to a questionnaire that they filled out for the World Cancer Research Fund, back in 1994. This is now superseded by other variables.

1= vegetarian

2= meat eater

3= fish eater

4+= Other

B- Variables

The following 211 variables (b1-b211), relate to the Food frequency section of the questionnaire. All of these have the same labels:

Frequency of consumption

0	never
1.00	less than once per month
2.00	1-3 per month
3.00	once a week
4.00	2-4 per week
5.00	5-6 per week
6.00	once per day
7.00	2-3 per day
8.00	4-5 per day
9.00	6 + per day

BREAD/SAVOURY BISCUITS 5 stars

B1 white bread

N Valid	34970
Missing	404
Mode	1.00
Minimum	00
Maximum	9.00

Valid % for mode 11.0

B2 brown bread

N Valid	33572
Missing	1802
Mode	.00
Minimum	.00
Maximum	9.0

Valid % for mode 18.8

B3 wholmeal bread

N Valid	34119
Missing	1255
Mode	7.00
Minimum	.00
Maximum	9.00

Valid % for mode 19.3

B4 chapati

N valid	34342
Missing	1032
Mode	.00
Minimum	.00
Maximum	9.00

Valid % for mode 48.3

B5 Papadum

N Valid	34381
Missing	993
Mode	.00
Minimum	.00
Maximum	9.00

Valid % for mode 48.5

B6 Tortilla

N Valid	34150
Missing	1224
Mode	.00
Minimum	.00
Maximum	6.00

Valid % for mode 73.0

BEVERAGES 5 stars

B184 tea (cup)

N Valid	35247
Missing	127
Mode	7
Minimum	0
Maximum	9

Valid % for mode 32.5

B185 herb tea (cup)

N Valid	34769
Missing	605
Mode	0
Minimum	0
Maximum	9

Valid % for mode 46.3

B186 coffee-instant,ground cup

N Valid	34666
Missing	708
Mode	7
Minimum	0
Maximum	9

Valid % for mode 25.6

B187 coffee decaffinated (cup)

N Valid	34773
Missing	601
Mode	0
Minimum	0
Maximum	9

Valid % for mode 51.1

B188 coffee substitute-Caro/bambu

N Valid	34957
Missing	417
Mode	0
Minimum	0
Maximum	9

Valid % for mode 93.1

B189 coffee whitener

N Valid	35081
Missing	293
Mode	0
Minimum	0
Maximum	9

Valid % for mode 80.9

B190 cocoa,hot chocolate/cup

N Valid	34985
Missing	389
Mode	0
Minimum	0
Maximum	9

Valid % for mode 49.6

B191 horlicks,ovaltine-cup

N Valid	34847
Missing	527
Mode	0
Minimum	0
Maximum	9

Valid % for mode 74.6

B192 low fat/cal horlicks/ovaltine

N Valid	35069
Missing	305
Mode	0
Minimum	0
Maximum	9

Valid % for mode 69.6

B193 Fresh orange juice

N Valid	35112
Missing	262
Mode	6
Minimum	0
Maximum	9

Valid % for mode 21.4

B194 other fruit juice

N Valid	34801
Missing	573
Mode	0
Minimum	0
Maximum	9

Valid % for mode 27.3

B195 squash/cordial diluted

N Valid	34983
Missing	391
Mode	0
Minimum	0
Maximum	9

Valid % for mode 47.7

B196 coke,lemonade

N Valid	34823
Missing	551
Mode	0
Minimum	0
Maximum	9

Valid % for mode 45.0

B197 low calorie soft drink

N Valid	34891
Missing	483
Mode	0
Minimum	0
Maximum	9

Valid % for mode 51.2

ALCOHOLIC BEVERAGES 5 stars

B198 wine glassful

N Valid	35203
Missing	171
Mode	4
Minimum	0
Maximum	9

Valid % for mode 20.3

B199 beer,lager half pint

N Valid	35085
Missing	289
Mode	0
Minimum	0
Maximum	9

Valid % for mode 55.5

B200 cider half pint

N Valid 3	5004
Missing	370
Mode	0
Minimum	0
Maximum	9

Valid % for mode 71.7

B201 port,sherry.liquers

N Valid	35080
Missing	294
Mode	0
Minimum	0
Maximum	9

Valid % for mode 39.5

B202 spirits single measure

N Valid	35154
Missing	220
Mode	0
Minimum	0
Maximum	9

Valid % for mode 41.0

BISCUITS, SWEETS & PUDDINGS

5 stars

B203 plain biscuits

N Valid	35115
Missing	259
Mode	1
Minimum	0
Maximum	9

Valid % for mode 20.2

B204 chocolate biscuits

N Valid	35140
Missing	234
Mode	1
Minimum	0
Maximum	9

Valid % for mode 23.0

B205 sandich, cream biscuits

N Valid	34986
Missing	388
Mode	0
Minimum	0
Maximum	9

Valid % for mode 43.8

B206 fruitcake

N Valid	35109
Missing	265
Mode	1
Minimum	0
Maximum	9

Valid % for mode 37.5

B207 sponge cakes

N Valid	35067
Missing	307
Mode	1
Minimum	0
Maximum	8

Valid % for mode36.5

VEGAN 4 stars

describe yourself as vegan?

Value Label
1 yes
2 no

describe yourself as vegan?

		Frequency	Percent
Valid	yes	401	1.2
	no	32535	92.0
	Total	32968	93.2
Missing	Inconsistent value	8	.0
	System	2398	6.8
	Total	2406	6.8
Total		35374	100.0

This variable was cross checked. If the response was **no** to being vegan, but then the subject continued to enter the number of years being vegan (yrsvegan), this was re-coded as value –1, which represents an inconsistent value. The variable was again cross checked if the response was yes to being vegan and then claimed to eat meat , it was re-coded again as –1 inconsistent value.

YRSVEGAN 4 stars

How many years have you been a vegan?

If years been a vegan was more than the age of the subject, response is re-coded as –1 inconsistent value.

		Frequency	***Percent***
Valid		470	1.3
Missing	Inconsistent value	3	.0
	System	***34901***	98.7
	Total	34904	98.7
Total		35374	100.0

HERBS 5 stars

Do you consume herbs/spices more than once per week?

Value Label
1 yes
2 no

		Frequency	***Percent***
Valid	yes	28018	79.2
	no	6623	18.7
	Total	34641	97.9
Missing	System	733	2.1
	Total	***35374***	***100.0***

APPENDIX C

Handling missing data in nutrition

Search History	Results	
1	missing data.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	2308
2	non response.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	1190
3	MCAR.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	62
4	missing completely at random.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	63
5	NMAR.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	2
6	missing at random.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	152
7	MNAR.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	8
8	(missing adj2 random).mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	209
9	(missing adj2 random).mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	209
10	complete case analyses.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	4
11	complete case analysis.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	65
12	last value carried forward.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	13
13	listwise deletion.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	33
14	mean substitution.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	38
15	em algorithm.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	723
16	multiple imputation.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	210
17	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16	4394
18	diet.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	181968

19	nutrition.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	79001
20	nutrients.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	21612
21	food frequency questionnaire.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	3056
22	18 or 19 or 20 or 21	260797
23	17 and 22	84
24	remove duplicates from 23	62
25	limit 24 to english language	57
26	from 25 keep 1-57	

Applications of the EM algorithm

#	Search History	Results
1	em algorithm.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	721
2	missing data.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	2264
3	non response.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	1129
4	MCAR.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	55
5	missing completely at random.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	63
6	missing at random.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	151
7	MAR.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	6523
8	MNAR.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	7
9	NMAR.mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	2
10	(missing not at random).mp. [mp=ti, ab, sh, tn, ot, dm, mf, rw, hw, ty, id]	25200
11	2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10	33175
12	1 and 11	79
13	remove duplicates from 12	53
14	limit 13 to english language	53

APPENDIX D

Multiple imputation by chained equation

```
local k=1
while `k'<6 {

use "C:\My Documents\Ula\Jennie\Jennie Newtheis\Newthesis"
gen age_0=age
summ age
gen agemean= r(mean)
replace age_0=agemean if age==.
drop agemean

gen sweatim_0=sweatim
summ sweatim
gen swemean=r(mean)
replace sweatim_0= swemean if sweatim==.
drop swemean

gen vitamin_0=vitamin
egen alm=mode(vitamin)
replace vitamin_0= alm if vitamin ==.
drop alm

gen alcholp1_0=alcholp1
egen alm=mode(alcholp1)
replace alcholp1_0 =alm if alcholp1==.
drop alm


gen smoking_0=smoking
egen alm=mode(smoking)
replace smoking_0=alm if smoking==.
drop alm

gen class_0=class
egen alm=mode(class)
replace class_0= alm if class==.
drop alm

gen imarried_0=imarried
egen alm=mode(imarried)
replace imarried_0=alm if imarried==.
drop alm

gen highedu_0=highedu
egen alm=mode(highedu)
replace highedu_0= alm if highedu==.
drop alm
```

```
gen vegen_0 =vegen
egen alm=mode(vegen)
replace vegen_0=alm if vegen==.
drop alm

**********************************************************************************************
local i=0
local j=1
while `i'<10 {

qui logit  vitamin_`i' alcholp1_`i' smoking_`i' sweatim_`i' class_`i' imarried_`i' age_`i'
highedu_`i' vegen_`i'  job region child16 illness hilo_f_v
predict p
gen q= 1-p
gen u=uniform()
gen vitamin_`j'= vitamin_`i'
replace vitamin_`j'= 1 if (u<p & vitamin==. & p~=.)
replace vitamin_`j'=0 if (u>p & vitamin==. &p~=.)
replace vitamin_`j'=. if p==. & q==. & vitamin==.
replace vitamin_`j'=vitamin_`i' if vitamin_`j'==.
drop u p q

qui mlogit  alcholp1_`i'  vitamin_`j' smoking_`i' sweatim_`i' class_`i' imarried_`i' age_`i'
highedu_`i' vegen_`i' job region child16 illness hilo_f_v
gen u=uniform()
predict p1 p2 p3 p4
gen p12=p1 + p2
gen p123= p1 + p2 + p3
gen alcholp1_`j'= alcholp1_`i'
replace alcholp1_`j'= 1 if alcholp1==.
replace alcholp1_`j'=2 if u>p1 & alcholp1==.
replace alcholp1_`j'=3 if u >p12 & alcholp1==.
replace alcholp1_`j'=4 if u>p123 & alcholp1==.
replace alcholp1_`j'=. if p1==. & p2==. & p3==. & p4==. & alcholp1==.
replace alcholp1_`j'= alcholp1_`i' if alcholp1_`j'==.
drop u p1 p2 p3 p4 p12 p123


qui regress sweatim_`i' vitamin_`j' alcholp1_`j' smoking_`i' class_`i' imarried_`i' age_`i'
highedu_`i' vegen_`i' job region child16 illness hilo_f_v
 predict pred_1, xb
 predict resid_1, r
 summ resid_1
 gen sweatim_`j' = (invnorm(uniform())*sqrt(r(Var))) + pred_1 if sweatim==.
 replace sweatim_`j'= sweatim if sweatim~=.
replace sweatim_`j'=sweatim_`i' if sweatim_`j'==.
replace sweatim_`j'=sweatim_`i' if sweatim_`j'<0
drop pred_1 resid_1


qui mlogit smoking_`i' vitamin_`j' alcholp1_`j' sweatim_`j' class_`i' imarried_`i' age_`i'
highedu_`i' vegen_`i' job region child16 illness hilo_f_v
gen u=uniform()
predict p1 p2 p3 p4
gen p12=p1 + p2
gen p123= p1 + p2 + p3
gen smoking_`j'= smoking
replace smoking_`j' = 1 if smoking==.
```

```
replace smoking_`j'=2 if u>p1 & smoking==.
replace smoking_`j'=3 if u>p12 & smoking==.
replace smoking_`j'=4 if u>p123 & smoking==.
replace smoking_`j'=. if p1==. & p2==. & p3==. &p4==. & smoking==.
replace smoking_`j'=smoking_`i' if smoking_`j'==.
drop u p1 p2 p3 p4 p12 p123

qui mlogit  class_`i' vitamin_`j' alcholp1_`j' sweatim_`j' smoking_`j' imarried_`i' age_`i'
highedu_`i'  vegen_`i' job region child16 illness hilo_f_v
gen u=uniform()
predict p1 p2 p3 p4 p5 p6 p7 p8 p9 p10
gen p12=p1 + p2
gen p123= p1 + p2 + p3
gen p1234= p1 + p2 + p3+p4
gen p12345=  p1 + p2 + p3 + p4 + p5
gen p123456=  p1 + p2 + p3 + p4 + p5 +p6
gen p1234567=  p1 + p2 + p3 + p4 + p5+p6 +p7
gen p12345678=  p1 + p2 + p3 + p4 + p5+p6 +p7 + p8
gen p123456789=  p1 + p2 + p3 + p4 + p5+p6 +p7 + p8 + p9
gen class_`j'= class
replace class_`j'=0 if class==.
replace class_`j'=1 if u>p1& class==.
replace class_`j'=2 if u>p12 & class==.
replace class_`j'=3 if u>p123 & class==.
replace class_`j'=4 if u>p1234 & class==.
replace class_`j'=5 if u>p12345 & class==.
replace class_`j'=6 if u>p123456 & class==.
replace class_`j'=7 if u>p1234567 & class==.
replace class_`j'=8 if u>p12345678 & class==.
replace class_`j'=9 if u>p123456789 & class==.
replace class_`j'=. if (p1 ==. & p2==. & p3==. & p4==. & p5==. & p6==. & p7==. & p8==. &
p9==. & p10==. & class==.)
replace class_`j'=class_`i' if class_`j'==.
drop   u p1 p2 p3 p4 p5 p6 p7 p8 p9 p10 p12 p123 p1234 p12345 p123456 p1234567
p12345678 p123456789


qui mlogit  imarried_`i' vitamin_`j' alcholp1_`j' sweatim_`j' smoking_`j' class_`j' age_`i'
highedu_`i' vegen_`i' job region child16 illness hilo_f_v
gen u=uniform()
predict p1 p2 p3 p4 p5
gen p12=p1 + p2
gen p123= p1 + p2 + p3
gen p1234= p1 + p2 + p3+p4
gen imarried_`j'=imarried
replace imarried_`j'=1 if imarried==.
replace imarried_`j'=2 if u>p1 & imarried==.
replace imarried_`j'=3 if u>p12 & imarried==.
replace imarried_`j'=4 if u>p123 & imarried==.
replace imarried_`j'=5 if u>p1234 & imarried==.
replace imarried_`j'=. if p1==. & p2==. & p3==. & p4==. & p5==. & imarried==.
replace imarried_`j'=imarried_`i' if imarried_`j'==.
drop  u p1 p2 p3 p4 p5 p12 p123 p1234


qui regress age_`i' vitamin_`j' alcholp1_`j' sweatim_`j' smoking_`j' class_`j' imarried_`j'
highedu_`i' vegen_`i' job region child16 illness hilo_f_v
 predict pred_1, xb
 predict resid_1, r
 summ resid_1
```

```
 gen age_`j' = (invnorm(uniform())*sqrt(r(Var))) + pred_1 if age==.
 replace age_`j'= age if age~=.
replace age_`j'=age_`i' if age_`j'==.
drop pred_1 resid_1


qui mlogit  highedu_`i' vitamin_`j' alcholp1_`j' sweatim_`j' smoking_`j' class_`j' imarried_`j'
age_`j' vegen_`i' job region child16 illness hilo_f_v
gen u=uniform()
predict p1 p2 p3 p4
gen p12=p1 + p2
gen p123= p1 + p2 + p3
gen p1234= p1 + p2 + p3+ p4
gen highedu_`j'= highedu_`i'
replace highedu_`j'=0 if u<p1 & highedu==.
replace highedu_`j'=1 if u>p1 & highedu==.
replace highedu_`j'=2 if u>p12 & highedu==.
replace highedu_`j'=3 if u>p123 & highedu==.
replace highedu_`j'=. if p1==. & p2==. & p3==. & p4==. & highedu==.
replace highedu_`j'= highedu_`i' if highedu_`j'==.
drop u p1 p2 p3 p4 p12 p123 p1234


qui logit  vegen_`i' vitamin_`j' alcholp1_`j' smoking_`j' sweatim_`j' class_`j' imarried_`j'
age_`j' highedu_`j' job region child16 illness hilo_f_v
predict p
gen q= 1-p
gen u=uniform()
gen vegen_`j'= vegen_`i'
replace vegen_`j'= 1 if (u<p & vegen==.)
replace vegen_`j'=0 if (u>p & vegen==.)
replace vegen_`j'=. if p==. & q==. & vegen==.
replace vegen_`j'=vegen_`i' if vegen_`j'==.
drop u p q

local i=`i'+1
local j=`j'+1
}

save "C:\My Documents\Ula\Jennie\Jennie Newtheis\dataset`k'", replace


local k=`k' +1
}

*********************************************************************************
 xi:logistic  hilo_f_v age_10 sweatim_10 i.vegen_10 i.vitamin_10 i.illness i.alcholp1_10
i.smoking_10 i.imarried_10 i.highedu_10 i.job i.class_10 i.region i.child16
logit
```

```
**********Survival Analysis**********

generate endcanre=mdy (12,31,2002)
format endcanre%d

gen surtime=breastcdate-formdate if breastcdate~=. & breastc==1
replace surtime=dod-formdate if dod<=endcanre
replace surtime=endcanre -formdate if breastc~=1
gen event =0
replace event = 1 if breastc==1
stset surtime , failure(event==1)

**********************************************************

**********to change survival time into years**********
replace surtime= surtime/365.25

*********BMI into groups*********
gen bmigroup=.
label define bmigroup 1"under weight" 2"Normal weight" 3"Over weight" 4"Obese"
replace bmigroup=1 if bmi<18.5
replace bmigroup=2 if bmi>=18.5 & bmi <=24.9
replace bmigroup=3 if bmi>24.9 & bmi<=29.9
replace bmigroup=4 if bmi>29.9 &bmi~=.
label values bmigroup bmigroup

**********Single variable  cox regression**********
char alcholp1 [omit]4
xi:stcox i.alcholp1

char bmigroup[omit]2
xi:stcox i.bmigroup

char smoking2 [omit]1
xi:stcox i.smoking2


xi: stcox i.agegroup


xi:stcox i.children

xi:stcox i.menopausal

**********************************************************************
char alcholp1_10 [omit]4
char bmigroup_10[omit]2
char smoking2_10 [omit] 3
char children_10[omit] 1
xi:stcox i.alcholp1_10 i.bmigroup_10 i.smoking2_10 i.children_10 i.menopausal_10
i.agegroup_10, schoenfeld(sch*) scaledsch(sca*)

stphtest, detail

stphtest, plot( _Ialcholp1_2) ytitle("Scaled Schoenfeld-Drink alcohol once a week")
```

**********assessing missing at random **********
gen
gen misalcholp1=alcholp1
replace misalcholp1=1 if alcholp1~=.
replace misalcholp1=0 if alcholp1==.
logistic misalcholp1 smoking2
logistic misalcholp1 highedu
logistic misalcholp1 class
logistic misalcholp1 sweatim
logistic misalcholp1 imarried
logistic misalcholp1 bmigroup
logistic misalcholp1 agegroup
logistic misalcholp1 children
logistic misalcholp1 menopausal

gen misbmi =bmigroup
replace misbmi=1 if bmigroup~=.
replace misbmi=0 if bmigroup==.
logistic misbmi smoking2
logistic misbmi highedu
logistic misbmi class
logistic misbmi sweatim
logistic misbmi imarried
logistic misbmi alcholp1
logistic misbmi agegroup
logistic misbmi children
logistic misbmi menopausal

gen mismoking= smoking2
replace mismoking=1 if smoking2~=.
replace mismoking=0 if smoking2==.
logistic mismoking bmigroup
logistic mismoking highedu
logistic mismoking class
logistic mismoking sweatim
logistic mismoking imarried
logistic mismoking alcholp1
logistic mismoking agegroup
logistic mismoking children
logistic mismoking menopausal

gen mischildren=children
replace mischildren=1 if children~=.
replace mischildren=0 if children==.
logistic mischildren bmigroup
logistic mischildren highedu
logistic mischildren class
logistic mischildren sweatim
logistic mischildren imarried
logistic mischildren alcholp1
logistic mischildren agegroup
logistic mischildren smoking2
logistic mischildren menopausal

```
gen misage = agegroup
replace misage=1 if agegroup~=.
replace misage=0 if agegroup==.
logistic  misage bmigroup
logistic  misage highedu
logistic  misage class
logistic  misage sweatim
logistic  misage imarried
logistic  misage alcholp1
logistic  misage children
logistic  misage smoking2
logistic  misage menopausal
```

```
**********logrank test to test difference in distribution of survival between missing and
nonmissng**********
sts test  misalcholp1, logrank
sts test   misbmi, logrank
sts test  mismoking, logrank
sts test  mismenopausal, logrank
sts test  mischildren, logrank
sts test   misage, logrank
```

```
**********Kaplain – Meier**********

gen grbmi=bmigroup
replace grbmi=5 if bmigroup==.
label define bmigroup 5"missing", modify
label values grbmi bmigroup
sts test grbmi, logrank

sts graph if _t>=2600, by(grbmi)title("Kaplan-Meier survival estimates, by BMI")
tmin(2000) noorigin ytitle("proportion surviving")


gen gralcholp1=alcholp1
replace gralcholp1=5 if alcholp1==.
label define gralcholp1 1"More than once a week" 2"Once a week" 3"Less than once a
week" 4"Never drink alcohol" 5"Missing"
label values gralcholp1 gralcholp1
sts test gralcholp1, logrank
sts graph if _t>=1000, by(alcholp1)title("Kaplan-Meier survival estimates, by alcohol")
tmin(2000) noorigin ytitle("proportion surviving")


gen grchild=children
replace grchild=3 if children==.
label define grchild 1"no children" 2"with children" 3 "missing"
label values grchild grchild

sts graph if _t>=2590, by(grchild)title("Kaplan-Meier survival estimates, by Do you have
children") tmin(2000) noorigin ylabel(0.5(.5)1)
sts test grchild, logrank

gen grsmoking=smoking2
replace grsmoking=4 if smoking2==.
label define SMOKING2 4"missing", modify
label values grsmoking SMOKING2
```

```
tab grsmoking
sts graph if _t>=2400 , by(grsmoking)title("Kaplan-Meier survival estimates, by smoking")
tmin(2000) noorigin ylabel(0.5(.5)1)

sts test grsmoking



gen grage=agegroup
replace grage=5 if agegroup==.
label define agegroup 5"Missing", modify
label values grage agegroup
sts graph if _t>=2400 , by(grage)title("Kaplan-Meier survival estimates, by age")
tmin(2000) noorigin ylabel(0.5(.5)1) legend (off)



gen agegroup=age
replace agegroup= 1 if age >=30 & age <40
replace agegroup=2 if age>=40 & age <50
replace agegroup=3 if age>=50 & age <60
replace agegroup=4 if age>=60 & age<75

label define agegroup 1"30-40" 2"41-50" 3"51-60" 4"61-75", modify
label values agegroup agegroup


recode children 2=0
label define newchildren 1"yes" 0"no"
label values children newchildren


*****************************************************************
local k=1
while `k'<6 {
use  "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\newchapter5.dta"
gen alcholp1_0=alcholp1
egen alm=mode(alcholp1)
replace alcholp1_0 =alm if alcholp1==.
drop alm

gen bmigroup_0 =bmigroup
egen alm=mode(bmigroup)
replace bmigroup_0=alm if bmigroup==.
drop alm

gen smoking2_0=smoking2
egen alm=mode(smoking2)
replace smoking2_0=alm if smoking2==.
drop alm

gen children_0 =children
egen alm=mode(children)
replace children_0=alm if children==.
drop alm

gen menopausal_0 =menopausal
```

```
egen alm=mode(menopausal)
replace menopausal_0=alm if menopausal==.
drop alm

gen age_0=age
summ age
gen agemean= r(mean)
replace age_0=agemean if age==.
drop agemean

gen imarried_0=imarried
egen alm=mode(imarried)
replace imarried_0=alm if imarried==.
drop alm

gen class_0=class
egen alm=mode(class)
replace class_0= alm if class==.
drop alm

gen highedu_0=highedu
egen alm=mode(highedu)
replace highedu_0= alm if highedu==.
drop alm


gen sweatim_0=sweatim
summ sweatim
gen swemean=r(mean)
replace sweatim_0= swemean if sweatim==.
drop swemean

gen surtime_0=surtime
summ surtime
gen sur=r(mean)
replace surtime_0= sur if surtime==.
drop sur
```

*****************multiple imputation by chained equations************

```
local i=0
local j=1
while `i'<10 {


qui mlogit  alcholp1_`i'  bmigroup_`i' smoking2_`i' children_`i'  menopausal_`i'  age_`i'
imarried_`i'  class_`i' highedu_`i' sweatim_`i'
gen u=uniform()
predict p1 p2 p3 p4
gen p12=p1 + p2
gen p123= p1 + p2 + p3
gen alcholp1_`j'= alcholp1_`i'
replace alcholp1_`j'= 1 if alcholp1==.
replace alcholp1_`j'=2 if u>p1 & alcholp1==.
replace alcholp1_`j'=3 if u >p12 & alcholp1==.
```

```
replace alcholp1_`j'=4 if u>p123 & alcholp1==.
replace alcholp1_`j'=. if p1==. & p2==. & p3==. & p4==. & alcholp1==.
replace alcholp1_`j'= alcholp1_`i' if alcholp1_`j'==.
drop u p1 p2 p3 p4 p12 p123

qui mlogit   bmigroup_`i'  alcholp1_`j' smoking2_`i' children_`i' menopausal_`i'  age_`i'
imarried_`i' class_`i' highedu_`i'
gen u=uniform()
predict p1 p2 p3 p4
gen p12=p1 + p2
gen p123= p1 + p2 + p3
gen bmigroup_`j'= bmigroup_`i'
replace bmigroup_`j'= 1 if bmigroup==.
replace bmigroup_`j'=2  if u>p1  & bmigroup==.
replace bmigroup_`j'=3 if u >p12 & bmigroup==.
replace bmigroup_`j'=4 if u>p123 & bmigroup==.
replace bmigroup_`j'=. if p1==. & p2==. & p3==. & p4==. & bmigroup==.
drop u p1 p2 p3 p4 p12 p123



qui mlogit smoking2_`i' alcholp1_`j' bmigroup_`j' children_`i' menopausal_`i'  age_`i'
class_`i'  highedu_`i'
gen u=uniform()
predict p1 p2 p3
gen p12=p1 + p2
gen smoking2_`j'= smoking2
replace smoking2_`j' = 1 if smoking2==.
replace smoking2_`j'=2 if u>p1 & smoking2==.
replace smoking2_`j'=3 if u>p12 & smoking2==.
replace smoking2_`j'=. if p1==. & p2==. & p3==. &  smoking2==.
replace smoking2_`j'=smoking2_`i' if smoking2_`j'==.
drop u p1 p2 p3 p12

qui logit  children_`i' alcholp1_`j' bmigroup_`j'  smoking2_`j' menopausal_`i' age_`i'
imarried_`i' class_`i'  highedu_`i' sweatim_`i'  surtime_`i'
predict p
gen q= 1-p
gen u=uniform()
gen  children_`j'=  children_`i'
replace  children_`j'= 1 if (u<p &  children==.)
replace  children_`j'=2 if (u>p &  children==.)
replace  children_`j'=. if p==. & q==. &  children==.
replace  children_`j'= children_`i' if  children_`j'==.
drop u p q

qui logit  menopausal_`i' alcholp1_`j' bmigroup_`j'  smoking2_`j' children_`j' age_`i'
imarried_`i' class_`i'  highedu_`i' sweatim_`i'
predict p
gen q= 1-p
gen u=uniform()
gen  menopausal_`j'=  menopausal_`i'
replace  menopausal_`j'= 1 if (u<p &  menopausal==.)
replace menopausal_`j'=0 if (u>p &  menopausal==.)
replace  menopausal_`j'=. if p==. & q==. &  menopausal==.
replace  menopausal_`j'= menopausal_`i' if  menopausal_`j'==.
drop u p q
```

```
qui regress age_`i'  alcholp1_`j' bmigroup_`j' smoking2_`j'  children_`j'  menopausal_`j'
class_`i' highedu_`i'
 predict pred_1, xb
 predict resid_1, r
 summ resid_1
 gen age_`j' = (invnorm(uniform())*sqrt(r(Var))) + pred_1 if age==.
 replace age_`j'= age if age~=.
replace age_`j'=age_`i' if age_`j'==.
drop pred_1 resid_1



local i=`i'+1
local j=`j'+1
}

save "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\dataset`k'", replace


local k=`k' +1
}
```

**********Test of the proportional hazard assumption**********

```
char alcholp1_10 [omit]4
char bmigroup_10[omit]2
char smoking2_10[omit] 3
xi:stcox i.alcholp1_10 i.bmigroup_10  i.smoking2_10  i.children_10  i.menopausal_10
i.agegroup_10, schoenfeld(sch*) scaledsch(sca*)

stphtest, detail

stphtest, plot( _Ialcholp1_2) ytitle("Scaled Schoenfeld-Drink alcohol once a week")
********imputing bw by bmi using hotdeck**************

use "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\newchapter5.dta", clear

hotdeck alcholp1, by(menopausal) store imp(5) keep(id)
use imp1
sort id
save  "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\alc1.dta", replace
use imp2
sort id
save  "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\alc2.dta", replace
use imp3
sort id
save "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\alc3.dta", replace
use imp4
sort id
save  "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\alc4.dta", replace
use imp5
sort id
save "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\alc5.dta", replace
```

Hotdeck imputation

```
**********imputing bmi by other variables  using hotdeck*********************

use "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\newchapter5.dta", clear
hotdeck bmigroup, by( agegroup) store imp(5) keep(id)
use imp1
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\bmi1.dta", replace
use imp2
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\bmi2.dta", replace
use imp3
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\bmi3.dta", replace
use imp4
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\bmi4.dta", replace
use imp5
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\bmi5.dta", replace


**********imputing smoking by other variables  using hotdeck*********************
use "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\newchapter5.dta", clear
hotdeck smoking2, by(menopausal) store imp(5) keep(id)
use imp1
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\smok1.dta", replace
use imp2
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\smok2.dta", replace
use imp3
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\smok3.dta", replace
use imp4
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\smok4.dta", replace
use imp5
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\smok5.dta", replace


**********imputing children by other variables  using hotdeck*********************
use "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\newchapter5.dta", clear
hotdeck children, by(menopausal) store imp(5) keep(id)
use imp1
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\ch1.dta", replace
use imp2
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\ch2.dta", replace
use imp3
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\ch3.dta", replace
use imp4
sort id
```

```
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\ch4.dta", replace
use imp5
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\ch5.dta", replace


**********imputing menopausal by other variables  using hotdeck**********
use "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\newchapter5.dta", clear
hotdeck   menopausal, by(menopausal) store imp(5) keep(id)
use imp1
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\men1.dta", replace
use imp2
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\men2.dta", replace
use imp3
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\men3.dta", replace
use imp4
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\men4.dta", replace
use imp5
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\men5.dta", replace



**********imputing agegrpoup by other variables  using hotdeck**********
use "C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\newchapter5.dta", clear
hotdeck   agegroup, by( menopausal) store imp(5) keep(id)
use imp1
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\age1.dta", replace
use imp2
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\age2.dta", replace
use imp3
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\age3.dta", replace
use imp4
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\age4.dta", replace
use imp5
sort id
save"C:\My data\Ula\Thesies\Newthesis\Chapter 5 -cancer\age5.dta", replace
```

***********Cox proportional hazard using the hotdeck imputed datasets****************

```
use "d:\ula\survival analysis\hialimp1", clear
gen surtime=formdate- dob if diagdat1==.
replace surtime= diagdat1 - dob if diagdat1 ~=.
replace surtime=. if surtime <0
replace surtime= surtime/365.25
```