The Effects of Missing Data on Multiple Linear & Logistic Regression

MATH5871M: Dissertation in Statistics

[redacted]

[redacted]

Supervisors: [redacted]

Submitted in accordance with the requirements for the degree of

Master of Science in Statistics

The University of Leeds, School of Mathematics

September 2011

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

Abstract

Background:

Missing data occurs frequently in data collection, especially in medical research. Often the impact of overlooking this missing data is underestimated. Methods for coping with missing values in a data set have recently become available to medical researchers. This study aimed to investigate such methods and carry out an analysis on a data set with missing values, concerning abdominal aortic aneurysm (AAA) surgery. The required outcome was a logistic regression model for whether or not subjects survived the AAA surgery.

Research & Study:

Research into each of the methods was carried out, using mainly journal articles, but also books. The books provided a useful introduction to new topics, whereas the journal articles provided more detailed information required for a deeper understanding. Definitions were given for the missingness mechanisms and explanations were given for the methods available for data with missing values. Small examples were given with the theory and then these ideas were applied to a larger data set to compare the methods under the different missingness mechanisms. Finally, the more complicated AAA surgery data set was analysed with an appropriate method and the logistic regression model obtained.

Results:

The data set contained 14010 subjects, with the missing values assumed to be missing at random, MAR. Multiple imputation was selected as the method for the data set, with different imputation methods selected for the different types of data; numerical, binary and categorical. The imputed values were concluded to be satisfactory and the logistic regression model produced, showing the impact of the predictors on the outcome of mortality.

Conclusion:

There are different missingness mechanisms and methods for coping with missing data. These methods have certain circumstances in which they would be most useful. Care should be taken when dealing with missing values to try to avoid bias in the results or loss of information. For the AAA surgery data set, multiple imputation proved to be the most suitable method.

Contents

List of Figures

List of Tables

Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or institution of learning.

In the attached submission I have not presented anyone else's work as my own. Where I have taken advantage of the work of others, I have given full acknowledgment. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the School's policy on mitigation and procedures for the submission of statements and evidence of mitigation.

Signed: ____________________

Chapter 1

Introduction

The study will investigate missing data and the different methods which are available to cope when data are missing. Missing data are unavoidable in clinical research and epidemiological studies and can often lead to invalid results, (Sterne *et al*, 2009). Missing data can occur for a number of different reasons. A value may be missing if the subject does not wish to disclose the information, such as household income, or it may be missing if the person is no longer available, for example if they have left the country. The reasons for the data being missing lead to different types of missing data, which will be investigated. Whatever the reasons for the missing data, it cannot be ignored from the analysis stage, since this can lead to bias in the results, (Sterne *et al*, 2009). Statistical methods have only recently become available to medical researchers to help tackle problems associated with missing data, (Sterne *et al*, 2009). There are now many different method available and some of these will be considered during this study.

Chapter 2 will introduce the main data set used during this study and provide basic summaries concerning any missing values in this data. Chapter 3 will outline some theory relating to missing data. It will introduce a small data set to help explain the theory, which includes different types of missing data and a variety of different methods which can be used to cope with data sets with missing values. Next, chapter 3 will introduce a package which is available in the statistical software R, which has been designed to implement some of the missing data methods. Chapter 4 contains a large data set and investigates how well the different methods cope with the different types of missing data. As the theory has now been introduced and smaller or simpler examples have been demonstrated, the main data set from chapter 2 returns in chapter 5 to be investigated and the missing data dealt with. Chapter 5 includes graphical summaries of the missing data, an appropriate method for handling the missing data and finally the results of the analysis. Lastly, conclusions are given and the appendices holds any additional information such as the R statistical software commands used.

Chapter 2

Abdominal Aortic Aneurysm (AAA) Surgery Data

The main data set being used during this project is concerning subjects who have undergone Abdominal Aortic Aneurysm (AAA) surgery. The outcome of interest is whether or not the subject survives the procedure. There are 14010 subjects in total and 23 variables recorded. Of these 14010 subjects, only 1964 are complete cases, that is, the subject has a recorded value for each of the 23 variables. The remaining subjects have a value missing from at least one of the 23 variables. These missing values will be the focus of this study. Smaller and more simple data sets will be used to help explain new methods as they are introduced. However, the aim of the study is to consider the missing values within this data set.

Various different methods for dealing with missing data will be considered. One will be chosen as the most appropriate method and used to create a regression model for this data set, with the outcome being whether or not the subject survives the surgery. This model can then be used to assess the risk involved for future patients considering AAA surgery, and highlight which variables are important in increasing or decreasing the chances of survival during the surgery. The outcome variable is called the 'dischargeStatus' and takes a value of 0 for 'survived' and 1 for 'died'. These steps can be found in chapter 5.

2.1 Explanation of the Data and Basic Summaries

The 23 variables are shown in table 2.1, along with a brief explanation. The first 15 variables are categorical, with the majority having simply yes/no answers. Table 2.2 shows the number of subjects in each of the categories for all of the categorical variables. It can be seen that there are a number of NA values, these represent the missing values for each of the variables.

Table 2.3 shows the odds ratios for the binary categorical variables against the outcome of death. It was not taken into account where the subject died, simply whether they did or did not survive. *If* the missing values had been included in these calculations, they may have affected these odds ratios. In some cases, the missing values *could* cause the odds ratios to change from above/below one to below/above one. Recall that an odds ratio of one suggests that those with, say, diabetes have the same odds of death as those without diabetes. The missing values, if added, could therefore affect whether this variable increases or decreases the subjects chances of survival after surgery.

The final 8 variables are continuous, but with maximum and minimum plausible values. Table 2.4 shows the upper and lower clinically plausible limits for each of the continuous variables. The units used are those given in table 2.1. Therefore, subjects with values which lie outside these boundaries, will be edited to NA for the relevant variables. This is so they can still be included in the analysis. Reasons for values outside of these boundaries may include;

- Data entry errors,
- Incorrect codes, such as entering '0' when 'NA' should be entered,
- Incorrect units used, for example, centimetres instead of millimetres.

Table 2.5 shows the upper and lower data values for each of the continuous variables. It can be seen that many of the variables have values outside the clinically plausible limits. The final column in table 2.5 shows the number of subjects with a value for that variable which lies outside the clinically plausible limits in table 2.4. It can be seen that these restrictions affect between 157 (1.1%) and 324 (2.3%) subjects for each of the continuous variables. Therefore, there are many subjects which need to be changed to NA for these variables. Note that for the age variable, the lowest value is -53. There were many recordings for age with a negative value, which may be a data entry error, whereby the number of years is correct, but the sign is incorrect. However, as this is only a suggestion, these negative values must be set to NA.

Finally, table 2.6 shows each of the variables along with how many missing values (NA) they have and the corresponding percentage of missing values. Recall that the data set contains 14010 subjects. It can be seen that the percentage of missing values for any one variable with any missing, ranges from $7.1 \times 10^{-3}\%$ for gender to 67.3% for stroke. Note there are no missing values for hospitalID. It is likely that the figure is so high for variables such as stroke and haemorrhage, as some hospitals or physicians may not record a value when the answer is 'no'. Therefore, some of the missing values may be those without a haemorrhage or without a stroke, but nothing was recorded for the subject, rather than a value of '0'.

2.1.1 Data Input Adjustments

There were some amendments which needed to be made to the raw data before it could be used. Some of these are listed below.

- The subjects were provided on every other line of a spreadsheet, which resulted in twice as many values being read into R. The blank spaces were interpreted as additional missing values. Therefore, these blank rows needed to be deleted before the data could be read into R. As deleted rows in spreadsheets are never truly removed, a new spreadsheet needed to be produced by copying and pasting the original data, once every other row had been deleted.
- The gender variable had values recorded in seven different ways. These were: f, Female, F, and F(space) for women and M, Male and m for men. Therefore, R was showing seven different categories for gender. Therefore, f, Female and F(space) were changed to F and m and Male were changed to M. Therefore, two categories resulted, namely F and M.
- There was one subject with a blank cell for gender, so this was coded to NA.
- The admission mode variable had only blank spaces for the missing values, so these were changed to NA for continuity in the data set.

- As previously mentioned, the status variable was simplified to 0=alive and 1=died. Therefore, where the subject died was not taken into consideration.

- There were also six categories amongst the categorical variables which were not specified in the description of the data set. For example, there were some subjects with an admission mode in a category 4, which is not listed. These were therefore followed up to see what the categories were, so the data could be interpreted correctly. However, no information was available for these missing categories, hence they had to be edited to NA also. It is likely that these additional categories are codings within say, a certain hospital or region. It can be seen from figure 2.1 that, for example, the admission mode categories 0, 3,4 and 4, which are not listed, only affect three of the 132 hospitals. It is therefore likely that these codes were used only within these hospitals. Previously, hospital staff could record data, such as admission mode, more freely, hence 'new' categories could be formed. However, now the categories are recorded on computer systems, using a drop-down box, and so only the listed categories can be chosen. It is therefore likely that the admission modes 0, 3,4 and 4 were not recorded recently. Note, the last option in the key, without a label, are those left blank, which will be coded as missing. Note also that those categories listed are shown in blue whilst those which are not listed are shown in orange.

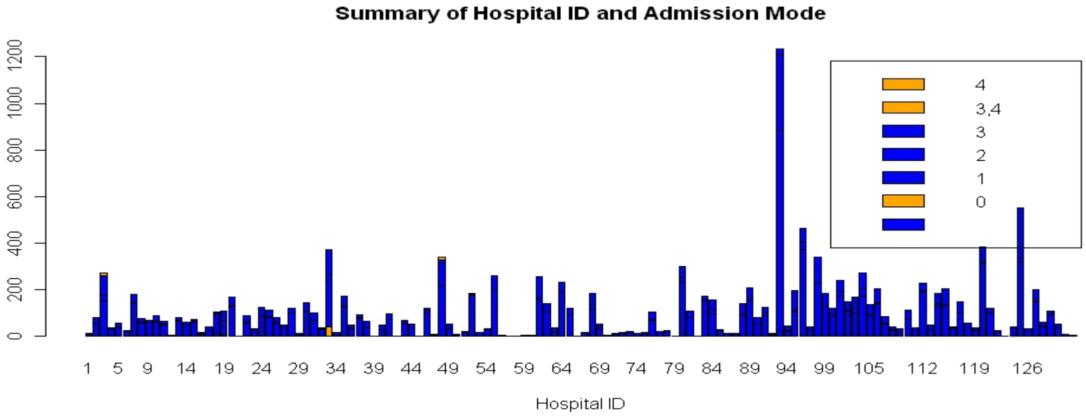


Figure 2.1: Additional admission mode categories

Variable	Explanation	Measure
Status	Did the patient survive	Alive=0, Died in theatre=1 Died in Recovery=2, Died in Ward=3

Gender	Sex	Male=M,
		Female=F

hospitalID	Each hospital given a number	Integer 1-132

adMode	Mode of admission	Elective=1
		Unplanned=2
		Emergency=3

Diabetes	Does the patient have	No=0, Yes=1
Smoker	Up to within 2 months	No=0, Yes=1
Dialysis	Has the patient had renal	No=0, Yes=1
Transplant	Has the patient had renal	No=0, Yes=1
Previous	Previous surgery/stent	No=0, Yes=1

AAASurgery	Type of surgery	Tube=1,
		Groin=2,
		Intraabdominal=3,
		Other=9

Haemorrhage	Has the patient had	None=0, Major=1

Stroke	Has the patient had	Non-Disabling=0,
		Disabling=1,
		None=2

Myo	Has the patient had myocardial infarction	No=0, Yes=1
Cardiac	Patient had heart failure	No=0, Yes=1
Hypo	Has the patient had hypotension	No=0, Yes=1
Age	Age of patient	Number of full years
Haemoglobin	(Hb) [g/dl]	Number
WhiteCell	WCC [109/l]	Number
Urea	[mmol/l]	Number
Sodium	[mmol/l]	Integer
Potassium	[mmol/l]	Number
LowBP	BP of patient	Integer
HighPulse	Pulse of patient	Integer

Table 2.1: An explanation of the variables

Variable	0	1	2	3	9	NA
Gender	F=2019	M=11990	-	-	-	1
adMode	43	9911	895	-	-	3161
Diabetes	11698	1490	-	-	-	822
Smoker	8725	2274	-	-	-	3011
Dialysis	11943	150	-	-	-	1917
Transplant	11741	101	-	-	-	2168
Previous	12060	354	-	-	-	1596
AAASurgery	-	5596	1096	2706	931	3681
Haemorrhage	4699	409	-	-	-	8902
Stroke	1337	43	3208	-	-	9422
Myo	4671	569	-	-	-	8770
Cardiac	4240	602	-	-	-	9168
Hypo	4154	697	-	-	-	9159
Status	12183	279	850	255	-	443

Table 2.2: Quantities of the categorical variables

Variable	Odds Ratio (2dp)	p-value (2dp)	95% Confidence Interval (2dp)
Gender	0.66	1.69e-08	(0.57, 0.76)
Diabetes	1.12	0.21	(0.93, 1.34)
Smoker	1.13	0.14	(0.96, 1.33)
Dialysis	1.16	0.18	(0.82, 2.31)
Transplant	0.84	0.86	(0.35, 1.74)
Previous	1.28	0.14	(0.90, 1.79)
Haemorrhage	5.37	$< 2.2e-16$	(4.33, 6.66)
Myo	6.09	$< 2.2e-16$	(5.04, 7.35)
Cardiac	9.94	$< 2.2e-16$	(8.22, 12.04)
Hypo	12.77	$< 2.2e-16$	(10.62, 15.38)

Table 2.3: The odds ratios

Variable	Lower Limit	Upper Limit
Age	18	100
Haemoglobin	2	20
WhiteCell	2	50
Urea	0.1	800
Sodium	105	165
Potassium	2	40
LowBP	20	250
HighPulse	20	200

Table 2.4: Clinically plausible limits of the continuous variables

Variable	Data Lower Limit	Data Upper Limit	Number of Values set to NA
Age	-53	108	169
Haemoglobin	0	1531	276
WhiteCell	0	298	324
Urea	0	808	201
Sodium	0	220	218
Potassium	0	140	215
LowBP	0	200	247
HighPulse	0	800	157

Table 2.5: Continuous variables: implausible values

Variable	Number of Missing Values	Percentage of Missing Values (1dp)
Gender	1	0.0
adMode	3161	22.6
Diabetes	822	5.9
Smoker	3011	21.5
Dialysis	1971	13.7
Transplant	2168	15.5
Previous	1596	11.4
AAASurgery	3681	26.3
Haemorrhage	8902	63.5
Stroke	9422	67.3
Myo	8770	62.6
Cardiac	9168	65.4
Hypo	9159	65.4
Status	443	3.2
Age	169	1.2
Haemoglobin	1460	10.4
WhiteCell	1719	12.3
Urea	1978	14.1
Sodium	1522	10.9
Potassium	1575	11.2
LowBP	2348	16.8
HighPulse	2427	17.3
hospitalID	0	0.0

Table 2.6: A summary of the missing values by variable

Chapter 3

Missing Data: Theory

A missing data value is where there is no data for a particular variable under a certain subject. Note this is different from a recorded value of zero. Missing data can commonly occur, especially in medical data, since some information may not be recorded. As this results in less information than was initially intended, the missing data can affect the conclusions drawn.

Throughout this chapter, an example complete data set will be used to demonstrate the theory being explained. The details of this body fat data set can be found below.

3.1 The Body Fat Data Set

This data set was obtained from Neter (1996) and concerns the body fat of females. It contains 20 healthy females, each with their body fat, triceps skinfold thickness, thigh circumference and midarm circumference measured. The body fat reading was obtained using an expensive method whereby the person was immersed in water. The aim is therefore to find a regression model which can provide the estimated body fat, using the predictors provided. No units were provided with the data set.

There follows some summary statistics and a regression model for the complete example data set, so these can be used for comparisons when the data set is used to explain the theory. The data set contains only continuous variables, which seem logically linked. The variable of interest is body fat, with triceps skinfold thickness, thigh circumference and midarm circumference as predictors. There are correlations between the predictors as shown below, with triceps skinfold thickness and thigh circumference being the most highly correlated at 0.92 (2dp). Intuitively, one may assume that body fat is related to triceps skinfold thickness and thigh circumference, since these are areas which appear to be more likely to store fat. Whereas the midarm circumference seems, to the untrained, to be less affected by body fat. The correlations show that triceps skinfold thickness and thigh circumference are both highly correlated with body fat, since they are both over 0.8 (1dp), whereas the midarm circumference and body fat have only a correlation of 0.14 (2dp). This may suggest that triceps skinfold thickness and thigh circumference are more important at predicting body fat than midarm circumference. Note how the predictors triceps and thigh are very highly correlated. If a variable selection method were to be applied, it may result in one of these variables being removed from the final model. However, for this example, they will both remain in the model. It will be noted that this could cause problems associated with collinearity, such as the estimated coefficients changing signs.

```
    Fat Triceps Thigh Midarm #the original data set
1  11.9    19.5  43.1   29.1
2  22.8    24.7  49.8   28.2
3  18.7    30.7  51.9   37.0
4  20.1    29.8  54.3   31.1
5  12.9    19.1  42.2   30.9
6  21.7    25.6  53.9   23.7
7  27.1    31.4  58.5   27.6
8  25.4    27.9  52.1   30.6
9  21.3    22.1  49.9   23.2
10 19.3    25.5  53.5   24.8
11 25.4    31.1  56.6   30.0
12 27.2    30.4  56.7   28.3
13 11.7    18.7  46.5   23.0
14 17.8    19.7  44.2   28.6
15 12.8    14.6  42.7   21.3
16 23.9    29.5  54.4   30.1
17 22.6    27.7  55.3   25.7
18 25.4    30.2  58.6   24.6
19 14.8    22.7  48.2   27.1
20 21.1    25.2  51.0   27.5
```

```
> cor(body) #correlation matrix     #significant cors:
         Fat  Triceps Thigh Midarm        fat/thigh
Fat      1.00    0.84  0.88   0.14        fat/triceps
Triceps  0.84    1.00  0.92   0.46        triceps/thigh
Thigh    0.88    0.92  1.00   0.09        triceps/midarm
Midarm   0.14    0.46  0.09   1.00
```

Figure 3.1 shows a pairs plot of the body fat data and suggests that the relationships between the variables are linear, therefore multiple linear regression seems sensible. A linear regression model can now be formed for the complete data set. This can then be used to compare with the models which will be produced with the data set once values have been removed, that is, a data set with ‘missing values’. The complete data set model is shown below. It can be seen that the model produced is:

$$\text{bodyfat} = 117.1 + 4.3(\text{triceps}) - 2.9(\text{thigh}) - 2.2(\text{midarm}). \tag{3.1}$$

Note the unusual output produced. None of the variables are considered to be significant, yet the overall model is very significant. This may be explained by the high correlation between the triceps and thigh variables.

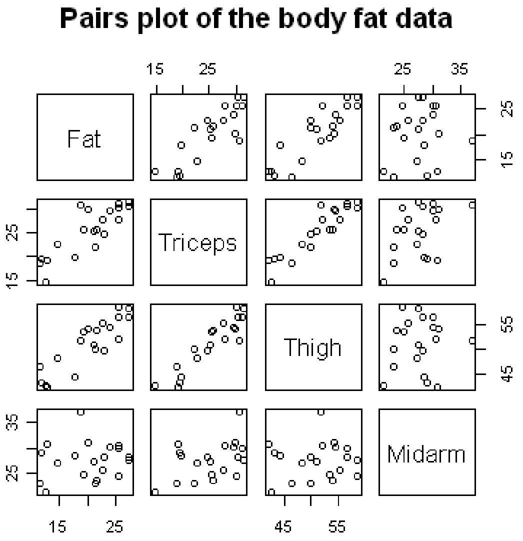


Figure 3.1: Pairs plot of the body fat data

```
Call: #output showing the regression model
lm(formula = fat ~ triceps + thigh + midarm)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
triceps        4.334      3.016   1.437    0.170
thigh         -2.857      2.582  -1.106    0.285
midarm        -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014,     Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

Models will be formed using the methods which follow, to demonstrate how the methods are used. These models can then be compared to this complete data set model to see whether there is a difference, that is, how much the missing values affect the final model. Note however, that this data set is very small, there are only 20 subjects. Therefore, for a more accurate comparison of the different methods, these ideas will then be carried out using a larger data set with 1000 subjects, see chapter 4.

3.2 Types of Missing Data

Missing data can occur for a number of different reasons, some of which are listed below.

- The subject does not wish to answer the question. This can occur frequently during some data collection methods, such as questionnaires, since it is easy for the subject to miss out questions and not feel under pressure to complete all the questions. There are certain types of questions which may be more likely to be ignored than others, such as questions concerning income or medical history, which may be considered to be personal. Note that if there is an entire section regarding income or medical history, say, then this may result in some subjects missing out entire sections of the questionnaire.

- When a study is over a period of time, that is a longitudinal study, there may be missing data if the subject is removed from the study. For example, if the study is over a two year period, and during month 10 the subject moves to another country and so can no longer participate, then there will be missing data for this subject from month 10 to month 24. Other reasons for subjects being removed from the study may include death, or no longer satisfying the conditions required to participate, for example, having a certain illness.

- In medicine, there are many reasons why data may be missing. For example, there may be a shortage of machinery and so it may be the case that some variables may not be recorded, if not viewed to be essential, for example, height or weight. There may also be a shortage of staff, which could result in the subjects visit being rushed and hence any unnecessary steps ignored. Alternatively the member of staff, if busy, may not record a ‘satisfactory’ value, for example a healthy blood pressure reading, if they are particularly rushed. Or, if this value is recorded, it may be recorded quickly and it is possible that the value could be input incorrectly onto a computer, or the numbers unclear if on a handwritten report. These steps could result in an implausible value being recorded. Missing data could also arise from a patient who has passed away part way through their treatment, and hence the most recent readings will be missing values.

There are also different types of missing data, and the type can determine the severity by which the conclusions are affected by the missing data. Commonly used types are mentioned below.

3.2.1 Missing Completely at Random (MCAR)

There are no systematic differences between the missing values and those which have been recorded (Sterne *et al*, 2009). This type of missing data is called missing completely at random, MCAR. It may arise from, say, a broken piece of machinery, which results in some subjects not having their pulse measured. If the data are missing completely at random, the recorded data is considered to be a random sample of the complete data set. It therefore has some desirable qualities which other types of missingness do not possess.

This explanation is taken from Little & Rubin (2002). Let the complete data be $Y = (y_{ij})$, which denotes an $(n \times K)$ rectangular data set which contains no missing values. The ith row is $y_i = (y_{i1}, ..., y_{iK})$, where y_{ij} is the value for variable Y_j for subject i. Where there are missing data, let the missing data indicator matrix be $A = (A_{ij})$, where $a_{ij} = 1$ if y_{ij} is a missing value and $a_{ij} = 0$ if y_{ij} is not missing. Note that A is then a matrix which shows the pattern of the missing data. The missing data mechanism is characterised by the conditional distribution of A

given Y, say $f(A \mid Y, \phi)$, where ϕ denotes the unknown parameters. If the missingness does not depend on the values of the data Y, whether missing or recorded, then the data are called missing completely at random. Therefore,

$$f(A \mid Y, \phi) = f(A \mid \phi),\ \forall\, Y, \phi. \tag{3.2}$$

Note that this definition does not mean that the pattern of the missing values in the data is random, but instead that the missingness does not depend on the data values in any of the variables.

In the example data set regarding body fat, this means that the missing values will not differ from those which have been recorded. For example, it may be that a subject does not have their triceps skinfold thickness measured as the skinfold calipers used for measuring may have been misplaced or may be in use elsewhere. Below is the data set with ten values missing completely at random from the predictors only, that is around 17% MCAR overall, $\left(\frac{10}{60}\right)$, or $15 - 20\%$ for each predictor, $\left(\frac{3}{20}\right)$ or $\left(\frac{4}{20}\right)$.

```
    Fat  Triceps Thigh Midarm
1  11.9    19.5  43.1   29.1
2  22.8     -    49.8   28.2
3  18.7    30.7  51.9   37.0
4  20.1    29.8  54.3    -
5  12.9    19.1  42.2   30.9
6  21.7    25.6   -     23.7
7  27.1    31.4  58.5   27.6
8  25.4    27.9  52.1   30.6
9  21.3    22.1   -      -
10 19.3    25.5  53.5   24.8
11 25.4    31.1  56.6   30.0
12 27.2    30.4  56.7   28.3
13 11.7    18.7  46.5    -
14 17.8    19.7  44.2   28.6
15 12.8    14.6  42.7   21.3
16 23.9     -    54.4   30.1
17 22.6     -    55.3   25.7
18 25.4    30.2   -     24.6
19 14.8    22.7  48.2   27.1
20 21.1    25.2  51.0    -
```

3.2.2 Missing at Random (MAR)

Any systematic difference between the missing values and the recorded values can be explained by differences in the other variables (Sterne *et al*, 2009). This is called missing at random, MAR. For example, it may be the case that the missing blood pressure measurements are generally lower than the recorded blood pressure measurements. However, this may be explained by the fact that younger people, with generally lower blood pressures, are more likely to have missing blood pressure measurements (Sterne *et al*, 2009). Note that missing at random can be viewed as a generalisation of missing completely at random (Cattle *et al*, 2011).

This explanation is taken from Little & Rubin (2002). As above, let the complete data be $Y = (y_{ij})$ and let the missing data indicator matrix be $A = A_{ij}$. Let Y_{obs} denote the observed entries of Y and let Y_{mis} denote the missing components. The assumption which is less restrictive than missing completely at random is that the missingness depends on the components of Y which are observed Y_{obs}, but not on those which are missing Y_{mis}. Therefore,

$$f(A \mid Y, \phi) = f(A \mid Y_{obs}, \phi), \; \forall \, Y_{mis}, \phi. \tag{3.3}$$

3.2.3 Missing Not at Random (MNAR)

Once the recorded data have been taken into account, there may still be systematic differences between the missing values and the recorded values (Sterne *et al*, 2009). This is termed as missing not at random, MNAR. For example, people who have high blood pressure may be more likely to miss clinic appointments because they have headaches resulting from their high blood pressure (Sterne *et al*, 2009).

Following on from the notation given for missing completely at random and missing at random taken from Little & Rubin (2002), if the mechanism is missing not at random, this means that the distribution of A depends on the missing values Y_{mis}, in Y.

3.2.4 Missing by Design

The data are missing due to the design of the data collection, for example, the style of the questionnaire. This is a less commonly used term. Note that missing by design is often equivalent to missing completely at random (Cattle *et al*, 2011). One example may be the duration of diabetes. If the person is not diabetic, the answer to this question will be missing.

3.3 Deletion Methods

There are methods which have been suggested to try to reduce the amount of missing data or cope with missing data when it is present. The amount of missing data can be reduced by carefully planning the method of data collection. For example, if a questionnaire is the most suitable method to use, rather than selecting a paper questionnaire, an electronic questionnaire could be used, which does not allow the subject to miss out any questions. It is easy to incorporate a function whereby the subject cannot progress to the next question, without first completing the current question. Generally, if missing data are still present, which is likely, then try to use a data analysis method which is relatively robust to missing data. Deletion methods are one group of methods which can be used

3.3.1 Complete Case Analysis/Listwise Deletion/Casewise Deletion

A commonly used approach for dealing with missing data is to only analyse the complete cases. A complete case is a subject who has a value recorded for each of the variables, that is, they have no missing data. Therefore, any subjects who have one or more value missing from the variables are excluded from the analysis. The main advantage of complete case analysis is the simplicity of the method along with the fact that standard complete data analysis can be used (Little & Rubin, 2002). However, complete case analysis can lead to bias in the results. Bias usually occurs when the data are not missing completely at random and the complete cases are not a random sample of the original cases (Little & Rubin, 2002). Also, in data sets with a large number of variables, this can cause many of the subjects to be excluded from the analysis, which results in a large loss of both power and precision (Sterne *et al*, 2009). Note that medical data often contains many variables, both about the subject generally and their disease of interest.

There are situations where complete case analysis will not lead to bias. Examples, taken from Sterne *et al* (2009), are listed below.

- When the missing data occurs only in the outcome variable which is measured only once for each subject. This is only so if all variables associated with the outcome being missing, can be included as covariates. This requires an assumption of missing at random.

- If the missing values are in the predictor variables and the reasons for the missing data are unrelated to the outcome.

Complete case analysis should only be considered when the bias is minimal and there is not much loss of precision, thus making the simple method attractive (Little & Rubin, 2002). Note this will often be when there are only a small number of incomplete cases in relation to the number of complete cases.

The example data set regarding body fat can be used here. From the original data set, let ten of the original values be missing completely at random as shown in figure 3.2; the missing values are represented with black circles. This revised data set will be used in the examples which follow in this section.

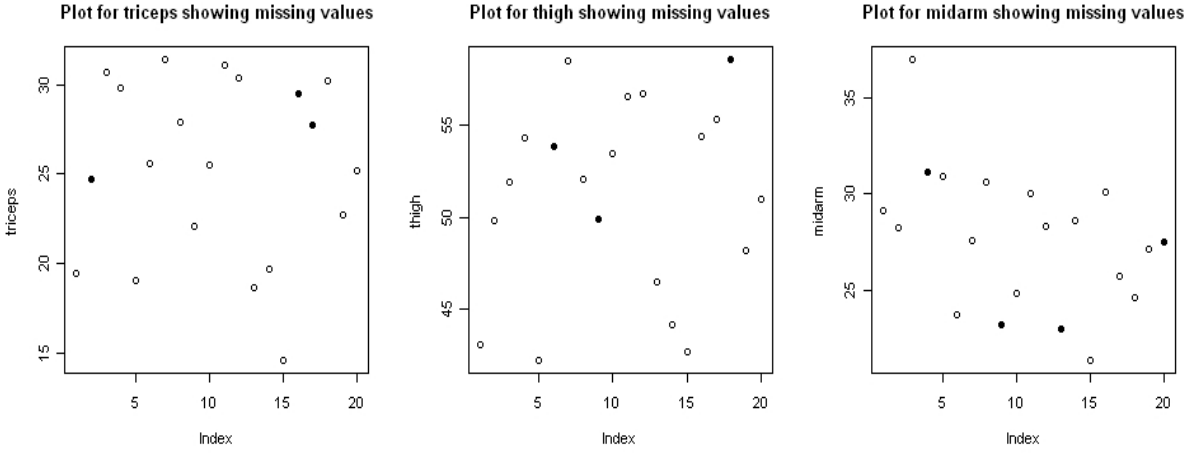


Figure 3.2: Body fat data set: 10 values ($\sim 17\%$) *missing completely at random*

For complete case analysis, the subjects with numbers 2, 4, 6, 10, 13, 16, 17, 18 and 20 would all need to be removed from the analysis. Therefore, just ten missing values (17%) would result in 9 of the original 20 subjects (45%) being excluded for the purposes of the analysis. This is a large loss of data. Multiple linear regression can now be used and a model produced for this data set once the subjects with missing values have been removed. The resulting model is,

```
Residuals: #model results
    Min      1Q  Median      3Q     Max
-3.3129 -2.1982  0.2178  1.1987  3.6903

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.5488   145.4026   0.128    0.902
Xcomp[, 2]    1.3455     4.4042   0.306    0.769
Xcomp[, 3]   -0.2621     3.7491  -0.070    0.946
Xcomp[, 4]   -0.6769     2.3511  -0.288    0.782

Residual standard error: 2.909 on 7 degrees of freedom
Multiple R-squared: 0.8347,     Adjusted R-squared: 0.7639
F-statistic: 11.78 on 3 and 7 DF,  p-value: 0.003991

                  2.5 %     97.5 % #confidence intervals
(Intercept) -325.273847 362.371416
Xcomp[, 2]    -9.068735  11.759705
Xcomp[, 3]    -9.127365   8.603097
Xcomp[, 4]    -6.236443   4.882682
```

Once models have been produced for all the different methods, these will be compared to one another and the original model which has no missing values, see §3.7.

3.3.2 Available Case Analysis/Pairwise Deletion

Pairwise deletion is a similar idea to casewise or listwise deletion as mentioned above. However, rather than deleting an entire subject if they have one or more value missing, this method simply uses the subjects which are complete in terms of the variables required for the calculation. For example, if there are 20 variables in total and for the first calculation only 10 variables are required, then all subjects who have values for these 10 variables will be used. However, if in the next calculation, only 8 variables are required, then there may be more subjects which can contribute to the calculation. This can result in different calculations having different sample sizes. This method can be suitable if there are,

1. A larger number of subjects,
2. Very few missing values, and
3. Not high correlations between the variables.

However, this method can still produce unreliable results and it does assume the data to be missing completely at random (Little & Rubin, 2002). Also, as the sample size can change from one calculation to another, this does not allow the analyst to compare the results across the variables (Little & Rubin, 2002). For this reason, a fixed sample size would be far more convenient.

The example data set regarding body fat may not be suitable here, since there are only a few subjects with a reasonably large amount of missing values and with high correlations between the variables. Therefore, the results obtained from this method, for this data set, are expected to be poor. Firstly, if all the variables were to be used in the analysis, the amended data set would be same as that used for the complete case analysis. However, if just two predictors are used for illustration, say triceps and thigh, then a new amended data set can be used. This results in 6 of the original 20 subjects being deleted from the analysis. The resulting model and output are given below. Note how these differ again from the previous two models.

```
Call: #model results
lm(formula = Xcomp[, 1] ˜ Xcomp[, 2] + Xcomp[, 3])

Residuals:
    Min      1Q  Median      3Q     Max
-3.0085 -2.1147  0.1564  1.0852  4.0237

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.17246   11.15710  -2.077   0.0715 .
Xcomp[, 2]    0.08408    0.42033   0.200   0.8464
Xcomp[, 3]    0.81003    0.40624   1.994   0.0813 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.737 on 8 degrees of freedom
Multiple R-squared: 0.8327,     Adjusted R-squared: 0.7909
F-statistic: 19.91 on 2 and 8 DF,  p-value: 0.0007827

                  2.5 %   97.5 % #confidence intervals
(Intercept) -48.9007866 2.555873
Xcomp[, 2]   -0.8851945 1.053364
Xcomp[, 3]   -0.1267491 1.746816
```

3.4 Single Imputation

“Imputation is a procedure for entering a value for a specific data item where the response is missing or incomplete” (UNECE, 2000). Alternatively, Statistics Canada (2003)[pp 43] states that “imputation is the process used to determine and assign replacement values for missing, invalid or inconsistent data that have failed edits. This is done by changing some of the responses or assigning values when they are missing on the record being edited, to ensure that estimates are of high quality and that a plausible, internally consistent record is created”. Therefore, imputation is a method by which the missing values are replaced by an estimate. There are two mains types of imputation, namely single imputation and multiple imputation. Single imputation is where one value is imputed for each of the missing items (Little & Rubin, 2002). Whereas multiple imputation imputes more than one value, which allows for the additional uncertainty of imputation (Little & Rubin, 2002). For more information on multiple imputation, see §3.5.

Single "imputations are means or draws from a predictive distribution of the missing values" (Little & Rubin, 2002). Therefore, there must be a way of choosing the predictive distribution for the imputed values, which must be based on the recorded data. Generally, there are two methods which can be used to select the predictive distribution; explicit modelling or implicit modelling. Explicit modelling is where a standard statistical model is chosen, such as the multivariate normal, and so the assumptions are explicit. Whereas implicit modelling is where the predictive distribution is formed using an algorithm, hence the assumptions are implicit (Little & Rubin, 2002).

Some of the different ways in which the replacement values can be imputed will be described below.

3.4.1 Mean Imputation

Mean imputation is an example of an explicit modelling method (Little & Rubin, 2002). It is commonly used as it is easy to perform, even for large data sets. First, the missing values for the variable are located and set aside. There remains only the subjects with a value for the chosen variable. The mean of these values is then computed and used as replacement values for those missing values. This process is then repeated for each of the variables with missing values. However, this method is not generally statistically valid and can result in serious bias (Sterne *et al*, 2009).

This method of imputation can be carried out using the example data set about body fat. First the mean values for the remaining variables must be calculated. These are shown in table 3.1. These can then be used in place of the missing values, and the regression performed as usual.

Variable	Mean Value (1dp)
Triceps	25.0
Thigh	50.6
Midarm	28.0

Table 3.1: Mean values: Body fat data set

The regression produces the following output,

```
Coefficients: #model results
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.4595    10.8369  -0.873   0.3956
triceps.i     0.4997     0.2739   1.824   0.0868 .
thigh.i       0.4734     0.2622   1.805   0.0899 .
midarm.i     -0.2427     0.2131  -1.139   0.2715
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 2.684 on 16 degrees of freedom
Multiple R-squared: 0.7673,     Adjusted R-squared: 0.7237
F-statistic: 17.58 on 3 and 16 DF,  p-value: 2.557e-05
```

3.4.2 Median Imputation

Median imputation is also an example of explicit modelling and again it is very easy to carry out. The process is the same as that for mean imputation, but the median value is used as the replacement value rather than the mean. As with mean imputation, there is not statistical justification for this method and it can lead to bias (Sterne *et al*, 2009). Single imputation of the missing values has limitations since it results in small standard errors and it does not allow for the uncertainty in the imputed values (Sterne *et al*, 2009).

This type of imputation can be carried out using the example body fat data set, once the median values for the remaining variables have been calculated. These are shown in table 3.2. These can then be used in place of the missing values, and the regression performed as usual.

Variable	Median Value (1dp)
Triceps	25.5
Thigh	51.9
Midarm	28.2

Table 3.2: Median values: Body fat data set

The following output and model are then produced. Note how similar this is to the mean imputation model.

```
Residuals: #model results
    Min      1Q  Median      3Q     Max
-4.0376 -1.8797  0.0655  1.6993  3.7749

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -11.020     11.324   -0.97    0.345
triceps5       0.472      0.287    1.65    0.119
thigh5         0.500      0.273    1.83    0.086 .
midarm5       -0.215      0.213   -1.01    0.328
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 2.59 on 16 degrees of freedom
Multiple R-squared: 0.783,      Adjusted R-squared: 0.742
F-statistic: 19.2 on 3 and 16 DF,  p-value: 1.48e-05
```

3.4.3 Regression Imputation

Regression imputation is a third example of explicit modelling. It predicts the missing values for each subject using multiple regression of the missing item on items observed for that variable (Little & Rubin, 2002). Mean imputation can be be viewed as a special case of regression imputation, whereby the predictor variables are dummy indicator variables for the cells in which the means are imputed (Little & Rubin, 2002). The model results generated in 'mice' with `quickpred`, see §3.8, are shown below.

```
Residuals: #model results
     Min       1Q   Median       3Q      Max
-3.87944 -1.33060 -0.06103  1.26763  4.78081

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -120.327    110.086  -1.093    0.291
triceps.i     -2.855      3.336  -0.856    0.405
thigh.i        3.261      2.845   1.146    0.269
midarm.i       1.657      1.769   0.937    0.363

Residual standard error: 2.646 on 16 degrees of freedom
Multiple R-squared: 0.7738,     Adjusted R-squared: 0.7314
F-statistic: 18.25 on 3 and 16 DF,  p-value: 2.042e-05
```

3.4.4 Stochastic Regression Imputation

Stochastic regression imputation is the final example of explicit modelling. It includes the addition of a random error to the regression prediction obtained using regression imputation, (Little & Rubin, 2002). Stochastic regression imputation compensates for the underestimation of the variance of variables with missing data that is associated with regression imputation. The models results generated using 'mice' and `quickpred`, see §3.8, are shown below.

```
Residuals: #model results
    Min      1Q  Median      3Q     Max
-3.9454 -1.9167  0.1710  1.3670  4.0248

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.61059   97.19035  -0.181    0.858
triceps.i     0.25872    2.87264   0.090    0.929
thigh.i       0.62015    2.45001   0.253    0.803
midarm.i     -0.01773    1.60181  -0.011    0.991

Residual standard error: 2.648 on 16 degrees of freedom
Multiple R-squared: 0.7736,     Adjusted R-squared: 0.7311
F-statistic: 18.22 on 3 and 16 DF,  p-value: 2.061e-05
```

3.4.5 Hot Deck Imputation/Pattern Matching

This is the first example of implicit modelling. Pattern matching is a form of imputation where the missing values are replaced with the value held by another subject. This other subject is similar in respect to the other variables, for which both subjects do have values recorded. This method is most suitable when there are a low number of missing values and the missing values are present in many different variables (Kilne, 1998). This method is very common when using surveys and there are a number of different techniques available for how the 'similar' subject can be selected (Little & Rubin, 2002).

3.4.6 Substitution

Substitution is another example of implicit modelling. It works by replacing a non-responding subject by another which was not initially selected (Little & Rubin, 2002). For example, if a subject cannot be contacted, then another subject with similar characteristics, who was initially not selected, can be substituted. The subject may be similar with respect to a set variable, such as location or age. Note that the substituted subject differs from the initial subject since they responded. This can show a systematic difference between the two subjects and hence the substituted subject should be regarded as an imputation (Little & Rubin, 2002).

3.4.7 Cold Deck Imputation

Cold deck imputation is also an example of implicit modelling. It works by replacing the missing value with a constant value which is obtained from an external source, which may be a similar survey conducted in the past (Little & Rubin, 2002).

3.4.8 Composite Methods

Composite methods can also be classed as implicit. They simply combine ideas from different methods to form a compromise. An example may be combining hot deck and regression imputation by calculating the predicted means using regression but adding a residual (Little & Rubin, 2002).

3.4.9 Comments

Single imputation methods are generally relatively easy to conduct and thus can be popular. However, they can cause the estimated standard errors to be smaller than they should be, since they do not allow for the added uncertainty surrounding the missing values (Cattle *et al*, 2011). For example, mean imputation causes the estimated standard errors to generally be smaller and does not result in as much variability as is likely. Another example is regression imputation, which tends to strengthen correlations between the variables (Cattle *et al*, 2011).

3.5 Multiple Imputation

This information is taken from Sterne *et al* (2009), unless otherwise stated. Multiple imputation is a commonly used approach for tackling problems associated with missing data. It is therefore now widely available as a function in many statistical software packages. It is intended to be an improvement on single imputation by allowing for the uncertainty associated with imputing the missing values. It does this by producing several different possible imputed data sets and combining the set of results.

First, the data set is duplicated M times and each of the data sets has the missing values within it replaced with imputed values. These data sets are then sampled from their predictive distribution based on the recorded data. The procedure needs to allows for the uncertainty involved in imputing the values, hence it must provide the imputed values with a suitable amount of variability. Next, standard statistical methods should be used to fit appropriate models to each of the data sets, to answer the required question. Note that each of the data sets will provide slightly different results as there is the variability introduced to account for the uncertainty involved. Overall estimated relationships can be obtained by averaging over the data sets. Rubin's rules (Rubin, 1987) are used to calculate the standard errors, and hence allow for the variability in the results. "Rubin's rules state that the average of the parameter of interest is the multiply imputed estimator and its sampling variance is the average of the completed data sampling variances, inflated by the between completion variance (Rubin, 1987)" (Cattle *et al*, 2011). This is described mathematically by Cattle *et al* (2011) as follows; let $\hat{\theta}_m, m = 1, ..., M$ be the set of completed data estimates, that is including the imputed values, from the population quantity θ. Also, let $\hat{s}_m^2$ be the estimates of the completed data sampling variances. Then, according to Rubin's rules, the multiply imputed estimator of θ is,

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m. \tag{3.4}$$

Its sampling variance can be calculated using,

$$\hat{s}^2 = \hat{U} + \left(1 + \frac{1}{M}\right) \hat{B}, \tag{3.5}$$

where,

$$\hat{U} = \frac{1}{M} \sum_{m=1}^{M} s_m^2, \tag{3.6}$$

and,

$$\hat{B} = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \tilde{\theta})^2. \tag{3.7}$$

Note that $\hat{U}$ would be the estimate of the sampling variance for $\hat{\theta}$ if the data had not included missing values. The $\hat{B}$ term increases the sampling variance to account for the uncertainty involved in imputing the missing values. Lastly, $\frac{B}{M}$ is a correction used as there are only a finite number of imputations carried out. The conclusions drawn are valid, since the averaging is carried out over the distribution of the missing data conditional on the recorded data.

The number of predictors in an imputation model should ideally be as big as possible, as generally, using more information will result in imputations with less bias (Meng, 1994). The imputation model should contain at least as many predictors as the final model, for example regression model, which will be analysed after the imputation. This is known as congeniality (Meng, 1994).

There are some points which should be considered when carrying out multiple imputation, to try to obtain accurate results. These include;

- The outcome variable should be included when imputing the missing value of a predictor. It is likely that the outcome value for the subject contains information about the predictor value.

- Many of the multiple imputation procedures assume the data to be normally distributed. Therefore problems can arise if this is not the case. One suggestion is to transform the data to be approximately normal before imputation and then transform back after imputing. Note that additional problems arise if any of the data are binary or categorical. It is believed that some methods are more suitable for categorical variables with missing data than others, but this is an area requiring further research.

- The missing at random assumption is a justification for the analysis and not a property of the data. Multiple imputation will only not contain bias if there are enough suitable variables included in the imputation model. For example, if subjects with high social-economic status are more likely to have their systolic blood pressure recorded and also less likely to have high systolic blood pressure, then social-economic status would need to be included in the imputation model, or else the model would underestimate the mean systolic blood pressure, which in turn could affect any associations between variables. For this reason, it is advisable to include a large number of variables in the imputation model.

- There may be data missing not at random, which can affect the validity of the multiple imputation results. For example, in an investigation of causes of depression, those depressed at the time may be more likely to not attend appointments. Therefore, this data would not be missing at random nor missing completely at random. However, the size of this problem is difficult to predict and therefore, it is a point which should be considered and mentioned during the report.

- Multiple imputation involves many computations and some approximations and therefore there may be situations where the algorithm needs to be run over and over. If the algorithm is not repeated enough times or the situation is different to that which the software was developed for, the results may not be as accurate as hoped.

Below is the R output from the multiple imputations using the body fat data set. The first output shows the results for 10 imputations, including the estimates, their standard error and the corresponding 95% confidence interval. The second output shows the same information but for 100 imputations. In this case it can be seen that the signs of the triceps and midarm estimates change. The standard errors of the estimates for 100 imputations are larger than the standard errors of the estimates for 10 imputations, but this is because the method has allowed for the added uncertainty involved in imputing values.

```
#for m=10, ie, 10 imputations
                    est        se       lo 95      hi 95
(Intercept) -41.5513761 96.841654 -256.640059 173.537307
triceps.i    -0.4423796  2.914073   -6.916134   6.031375
thigh.i       1.2170665  2.495808   -4.327399   6.761532
midarm.i      0.3868629  1.553794   -3.064283   3.838009
```

```
#for m=100, ie, 100 imputations
                    est         se      lo 95      hi 95
(Intercept)-10.46100325 105.790122 -246.026157 225.104150
triceps.i    0.48875184   3.201433   -6.638638   7.616141
thigh.i      0.40453702   2.731654   -5.676983   6.486057
midarm.i    -0.08418138   1.702489   -3.875787   3.707424
```

3.6 Other Methods

There are many other methods which have been suggested to deal with the problems associated with missing data. Two main examples are mentioned below, but not investigated any further for this study,

1. Full information maximum likelihood estimation (FIML),
2. The expectation-maximisation (EM) algorithm.

3.7 Summary and Comparison of the Methods

Some of the methods previously mentioned have been carried out using the data set regarding body fat with 20 subjects. Recall that the data set has 10 values, around 17%, missing completely at random (MCAR). The results from these methods are listed in table 3.3. Table 3.3 shows the linear regression models generated once the methods for dealing with the missing data had been used. The ‘original model’ is the linear regression model created from the data before any values were removed to create missingness. This is therefore the model which the others will be compared to. These estimates and confidence intervals were all generated using R, following the steps mentioned previously for each of the methods. The multiple imputations, however, were carried out using a statistical package in R called ‘mice’, see §3.8. As each method was only carried out once, it must be noted that some methods, such as stochastic regression imputation, may produce different results if carried out again. Note that others, such as mean imputation or median imputation, will remain the same however many times they are redone. Therefore, ideally these methods should be carried out a number of times and the results from these combined to produce a more accurate model.

From table 3.3 it can be seen that the original model suggests a positive correlation between the triceps variable and the outcome of body fat. It also suggests a negative association with both the thigh and midarm variables against body fat. However, none of the models which follow show these same associations. It can also be seen that the regression imputation model estimates, for all variables, lie outside the confidence intervals for the original model. Note also how the confidence intervals are generally rather wide, this is likely to be caused by the small number of subjects being studied. This, however, also results in many of the confidence intervals including both positive and negative values, which correspond to the variables not having a significant effect on the outcome of body fat. Therefore, the results seem rather inconclusive as to which variables are positively and which are negatively associated with the outcome of body fat. This may be partly explained by the high correlation between the triceps and thigh variables, which can cause problems associated with multicollinearity. One solution may be to use either the triceps or thigh variable as a predictor of the other and use just one of these variables. The results may be further improved if more information was available.

Method	Variable	Estimate (1dp)	95% Confidence Interval (1dp)
Original Model	Intercept	117.1	(-78.5, 312.7)
(No missing data)	Triceps	4.3	(-1.6, 10.2)
	Thigh	-2.9	(-7.9, 2.2)
	Midarm	-2.2	(-5.3, 0.9)
Complete Case Analysis	Intercept	18.6	(-325.3, 362.4)
	Triceps	1.4	(-9.1, 11.8)
	Thigh	-0.3	(-9.1, -8.6)
	Midarm	0.7	(-6.2, 4.9)
Mean Imputation	Intercept	-9.5	(-3.7, 11.8)
	Triceps	0.5	(0.0, 1.0)
	Thigh	0.5	(0.0, 1.0)
	Midarm	-0.2	(-0.7, 0.2)
Regression Imputation	Intercept	-120.3	(-336.1, 95.4)
	Triceps	-2.9	(-9.4, 3.7)
	Thigh	3.3	(-2.3, 8.8)
	Midarm	1.7	(-1.8, 5.1)
Stochastic Regression	Intercept	-17.6	(-208.1, 172.9)
Imputation	Triceps	0.3	(-5.4, 5.9)
	Thigh	0.6	(-4.2, 5.4)
	Midarm	0.0	(-3.2, 3.1)
Multiple Imputation	Intercept	-41.6	(-256.6, 173.5)
(10)	Triceps	-0.4	(-6.9, 6.0)
	Thigh	1.2	(-4.3, 6.8)
	Midarm	0.4	(-3.1, 3.8)
Multiple Imputation	Intercept	-10.5	(-264.0, 225.1)
(100)	Triceps	0.5	(-6.6, 7.6)
	Thigh	0.4	(-5.7, 6.5)
	Midarm	-0.1	(-3.9, 3.7)

Table 3.3: Estimates and confidence intervals produced using different methods: Body fat data set, 20 subjects, MCAR

Figure 3.3 shows the estimates and confidence intervals from the models produced using different methods for coping with missing data. It also shows the original model, that is, the model produced when no data are missing. Figure 3.3 corresponds to the numbers in table 3.3. It can be seen clearly from figure 3.3 that the complete case analysis model has the widest confidence intervals, since it has the least information available, as some subjects have been removed before analysis. The mean imputation model has the smallest confidence intervals, since any missing values are replaced by the mean value and therefore the variance is likely to be reduced, hence shrinking the confidence intervals. It can be seen that all confidence intervals overlap with the confidence interval for the original model. Also, all estimates except for regression imputation lie within the confidence intervals of the original model. Note, however, that regression imputation is a method which may produce different results if carried out several times, so this may not be so if the imputation was rerun. Also, the results can vary greatly depending upon which particular values are missing. For example, if those near the mean are removed, then mean imputation should cope relatively well, compared with if those far from the mean are removed. This also applies to those points which are more likely to affect a regression line and hence may affect the performance of regression imputation.

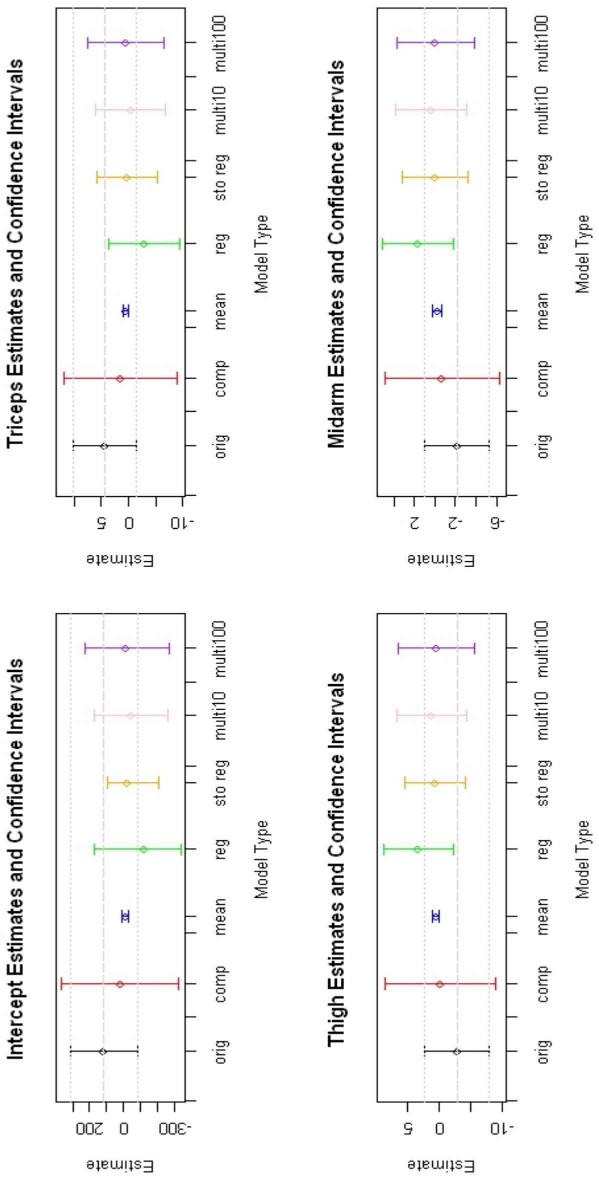


Figure 3.3: Estimates and confidence intervals for the body fat data: 20 subjects, 17 percent MCAR

3.8 The ‘mice’ Package

The ‘mice’ package in R (Van Buuren & Groothuis-Oudshoorn, 2011) will be used to produce all the estimates and confidence intervals for chapter 4. There follows information on this package, choices which need to be made during the calculations and details of any defaults which can be used. Many of the detail which follow are based on the information in Van Buuren & Groothuis-Oudshoorn (2011).

The name of the package, ‘mice’, stands for ‘Multivariate Imputation by Chained Equations’. Recall that multiple imputation is an option for tackling complex data with missing values and note that when the missing values are in more than one variable, this can increase the complexity of the problem. One method for imputing multivariate data is called fully conditional specification (FCS) and this is the method adopted by the ‘mice’ package. FCS specifies the imputation model one variable at a time using a set of conditional densities. There is one set of conditional densities for each incomplete variable. Then, once the initial imputation has been specified, the remaining imputations can be obtained by iterating over these conditional densities. One advantage of this method is that often a low number of iterations will be sufficient, say 10-20.

The ‘mice’ package works in stages for multiple imputation. It first takes the original data, with missing values, and creates as many copies of the data set as required, which have the missing values replaced by plausible, imputed values. Note the imputed values will differ between the data sets, hence there are now several data sets with the same observed values, but different imputed values. The imputed data sets are then saved in the package as a `mids`, that is, a ‘multiply imputed data set’. It then analyses the results from each of the separate imputed data sets. The analysis will differ depending upon the quantity of interest, it may be, for example, a regression coefficient. The analysis method used is often the same as the one which would be used had the data set been complete initially. There is a `with` function in the package which allows the analysis to be completed directly using the imputed data sets. There will be differences between the results from the different imputed data sets since there is uncertainty involved as to which imputed values to use. The results from the analyses are then stored as a `mira`, that is a ‘multiply imputed repeated analysis’. Finally it pools the results from these different data sets using Rubin’s rules, see §3.5, to form one overall result, with an associated variance. The pooled results are lastly stored as a `mipo`, a ‘multiply imputed pooled outcomes’. Figure 3.4 shows these steps more clearly.

There are some problems which may occur from using the above method. Some of these are listed below.

1. For a particular variable with missing values, the other variables used to impute may themselves have values missing,

2. Circular dependencies may occur when variables are correlated, leading to the first variable relying upon the second, and the second on the first,

3. When the number of variables is large and the number of subjects is small, there can be problems associated with collinearity or empty cells,

4. Sometimes there will be restrictions, such as when the data is ordered. One example is when there is longitudinal data,

5. It may be more complicated when the data consists of several different types, such as some categorical, some binary, some continuous, etc,

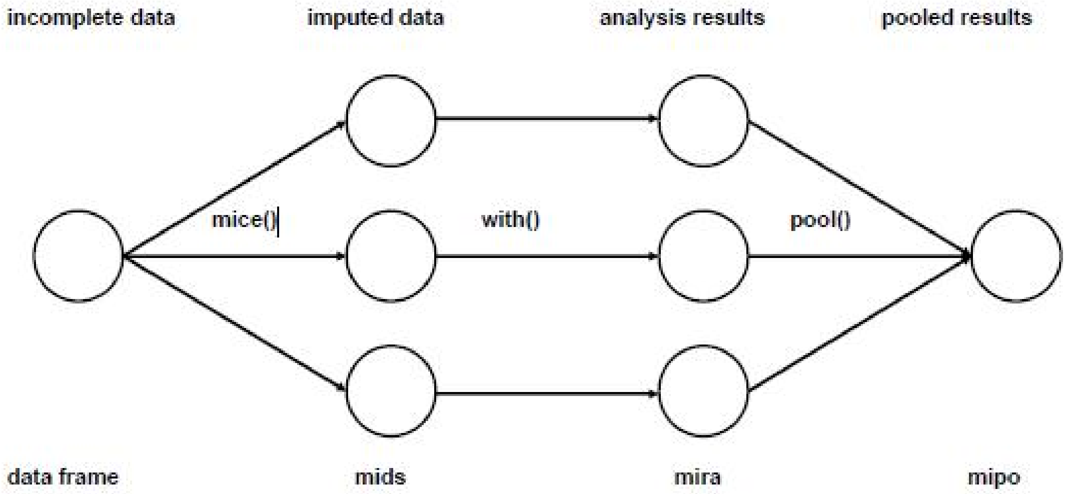


Figure 3.4: The steps used in multiple imputation, taken from Van Buuren & Groothuis-Oudshoorn (2011) Figure 1

6. Imputation can create impossible values, such as negative weights or pregnant males.

Therefore, care must be taken to summarise the imputated data sets and limit these problems where possible. One way to help alleviate some of these problems is by stating a specific imputation model for each column of the data, depending on the nature of the data in that column.

The 'mice' package can be used to summarise the data with missing values before any imputation is carried out. One possibility is to calculate the numbers of rows which are complete, which can then be used to calculate the number of missing data values. An alternative method is to calculate the missingness for pairs of variables. It may be that both the variables are missing, both are present, or one or the other is missing. The package also has many defaults programmed, therefore, it is possible to simply use one command using the data set with missing values, to produce a multiply imputed data set. Of course, these default settings do not need to be used, but allow for a basic `mids` to be produced easily. Once the `mids` is obtained, the imputations should be examined to assess whether they are plausible. Therefore, diagnostic checks should be carried out on the imputed values. Once the imputed values have been deemed satisfactory, they can be combined with the observed values to create an imputed data set. Note that different imputed data sets will have variations in the imputed values but the observed values will remain the same. Next, the analysis of interest can be carried out, in one single step, on all the imputed data sets, and the results pooled using Rubin's rules, see §3.5. Finally summaries can be produced of these pooled results for the overall conclusions to be drawn.

Choosing the imputation model is perhaps the most difficult part of the 'mice' package in R. There are seven stages which should be considered.

1. Whether the MAR assumption is plausible. This can be investigated using summary plots or summary tables.

2. The imputation model form needs to be specified for each column which contains missing values. Both the nature of the data in the column will be need to be considered and the relationship the column has with other columns. It is possible to specify a method for each

column separately, or use one method for all the columns. The package will automatically skip any variables without missing values, but it can also be requested that certain variables with missing values are also not imputed if required. Defaults are available and the package can distinguish between numerical, binary and categorical variables and assign the default method to each if necessary. The defaults are shown in table 3.4. See chapter 5 for more information on these methods.

Method	Description	Scale Type
pmm	Predictive mean matching	Numeric
logreg	Logistic regression	Factor, 2 levels
polyreg	Polytomous (unordered) regression	Factor, $>$ 2 levels

Table 3.4: Table showing the default methods for imputation in 'mice'

3. The set of variables which can be used as predictors should be chosen. It is recommended that as many appropriate variables as possible are included, but this may not be suitable for all data sets. A predictor matrix can be build which is a matrix of zeros and ones which indicate whether one variable is used as a predictor for another variable. The diagonal of this matrix will contain only zeros, since a variable cannot predict itself. There may be cases where the number of variables is very large and including all variables as predictors may become impractical or cause problems associated with multicollinearity. However, including more variables does make the MAR assumption more plausible. However, a suitably selected set of predictors will need to be chosen. Some advise is given on how to choose this subset.

    (a) Include all variables which will be included in the model after the imputation stage.

    (b) Include all variables which are related to the reasons for the missing values.

    (c) Include any variables which explain a large amount of variability.

    (d) Remove any variables from the previous two stage which should not be included as they have a large percentage of missing values.

    There is an option in the package to generate a prediction matrix quickly using the correlations of the variables as a basis. This includes a threshold which decides whether or not the variables should be used as predictors. For more details, see §4.2. Alternatively the default can be used, whereby each variable is used as a predictor for the other variables.

4. Whether to include in the imputation model those variables which are functions of other variables. For example, there may be a column for height, another for weight and a third for body mass index (BMI). It must then be questioned whether it is logical to impute values for all three of these variables, since BMI is calculated as the weight in kilograms divided by the height in metres squared. The package contains a mechanism, called passive imputation, to help deal with such circumstances.

5. Which order the variables should be imputed in. The default for the package is to impute the variables in order from left to right.

6. Where the starting imputation should be and the number of iterations. The default number of iterations is 5.

7. The number of multiply imputed data sets which should be generated. The default number of imputations is 5.

Other issues may need to be considered in addition to the ones listed above. Such as, it may be necessary to set limits on some of the imputed values. For example, there may be some implausible values. Say one variable is weight, there may be values imputed for weight which are negative. In this case, one solution would be to insert a lower bound on the weight variable, so any negative weights are pushed up to a certain minimum value, say zero. These restrictions will usually depend upon prior knowledge of the variables in the context they are in.

Once the imputed values have been obtained, the complete data sets, that is the observed values along with the imputed values, can be viewed for each of the imputed data sets. Summaries for each of the complete data sets can also be produced. Alternatively, if required, the imputed values only can be called in R and summaries of these studied. Next, the subjects who have had their outcome imputed must be removed from the analysis. They are used in the imputation stage to aid the imputation of the other variables, but must be removed from the final analysis stage. If they were to be kept in, they could add error to the results, (Cattle *et al*, 2011). Once the imputed data sets have been considered, they can be pooled using Rubin's rules, see §3.5, to give more accurate estimates and standard errors overall. Summaries of the pooled estimates can also be generated. Further example of using the 'mice' package are shown in chapter 5. There are also examples of the R code in appendix B.

Chapter 4

Missing Data: Body Fat Data Set Example (1000 subjects)

The body fat data set used in chapter 3 only contained 20 subjects. This is a useful size for a data set so the methods so far can be clearly explained. However, to thoroughly compare these methods, a larger data set is required. Therefore, the 20 subjects will be used as a basis on which to simulate a larger data set. The larger data set will contain 1000 subjects, which should be sufficient to compare the different methods.

4.1 The Data

First, the summary statistics from the original 20 subjects will be taken. These are shown in table 4.1. Next, the distribution of the data was investigated. A multivariate normal distribution was assumed, since the pairs plots showed approximately elliptical patterns, see figure 3.1 and each of the individual variables has approximately a normal distribution as shown in figure 4.1. It can be seen that there are no major departures from normality for each of the variables.

R was used to generate 1000 subjects from a multivariate normal distribution with the same mean and variance values as those listed in table 4.1. This was done using the vector of means from the variables of the original 20 subjects, along with their covariance matrix. The seed was also set for the random number generation, to allow the same 1000 subjects to be generated each time the command was run. The summary statistics from the 1000 subjects were then generated to ensure they were similar to those in table 4.1, see the R output below. First shown are the correlation matrices for both the 1000 subjects (`sample`) and the 20 subjects (`body`), then the mean values and variance/covariance matrix for easy comparison with table 4.1. The 1000 subjects can be seen in figure 4.2. Figure 4.3 shows the pairs plot for this new data set.

Variable	Mean (1dp)	Variance (1dp)
Fat	20.2	26.1
Triceps	25.3	25.2
Thigh	51.2	27.4
Midarm	27.6	13.3

Table 4.1: Summary statistics of the original body fat data set

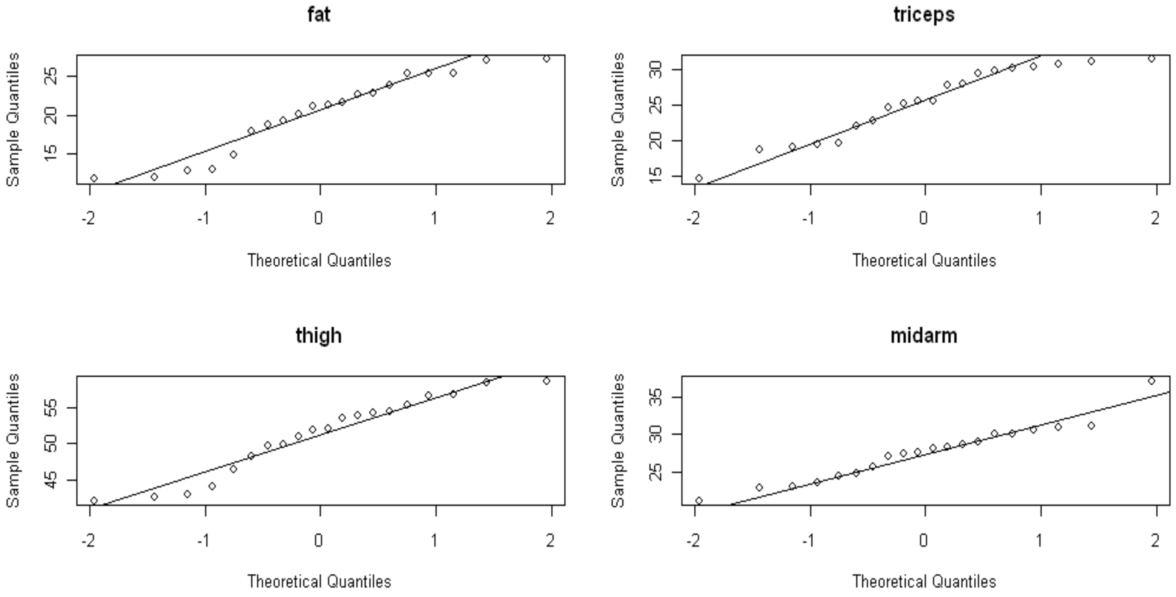


Figure 4.1: Body fat data set: qq-plots

```
> cor(sample) #1000 values      > cor(body) #20 values
       Fat  Tri  Thigh Midarm          Fat  Tri  Thigh Midarm
Fat    1.00 0.84 0.878  0.142  Fat    1.00 0.84 0.878  0.142
Tri    0.84 1.00 0.924  0.458  Tri    0.84 1.00 0.924  0.458
Thigh  0.88 0.92 1.000  0.085  Thigh  0.88 0.92 1.000  0.085
Midarm 0.14 0.46 0.085  1.000  Midarm 0.14 0.46 0.085  1.000
```

```
> mean(sample[,1]) #1000 values   > mean(body) #20 values
[1] 20.2                          Fat Triceps   Thigh  Midarm
> mean(sample[,2])                20.2    25.3    51.2    27.6
[1] 25.3
> mean(sample[,3])
[1] 51.2
> mean(sample[,4])
[1] 27.6
```

```
> var(sample) #1000 values        > var(body) #20 values
         Fat   Tri Thigh Midarm              Fat   Tri Thigh Midarm
Fat    26.07 21.63 23.47   2.65  Fat    26.07 21.63 23.47   2.65
Tri    21.63 25.23 24.29   8.39  Tri    21.63 25.23 24.29   8.39
Thigh  23.47 24.29 27.40   1.62  Thigh  23.47 24.29 27.40   1.62
Midarm 2.65   8.39  1.62  13.30  Midarm 2.65   8.39  1.62  13.30
```

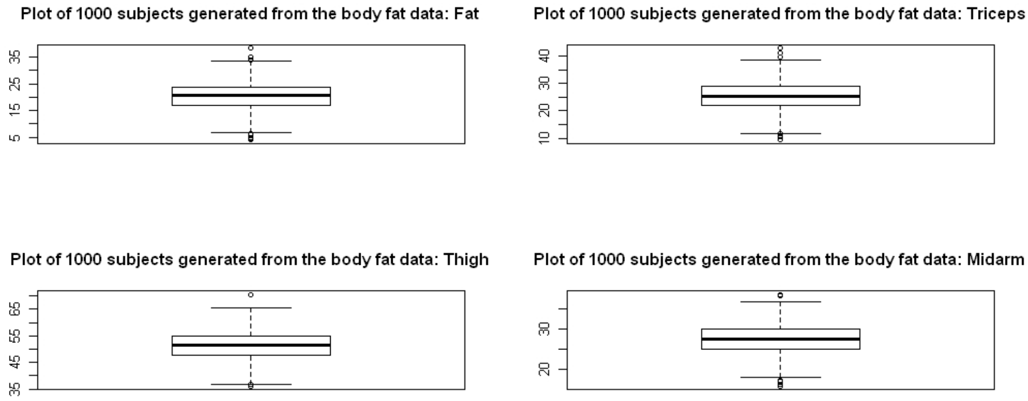


Figure 4.2: Body fat data set: 1000 generated values

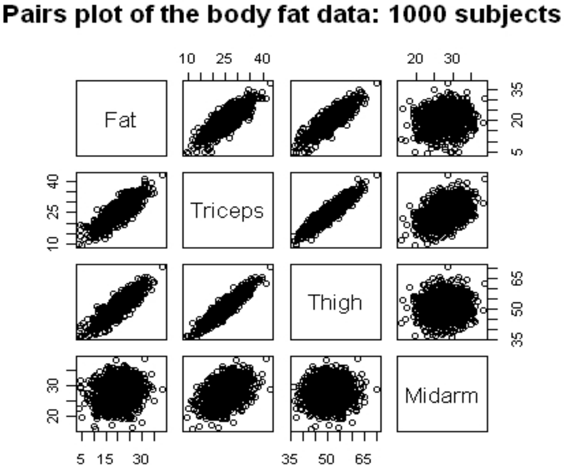


Figure 4.3: Pairs plot for the body fat data: 1000 subjects

Next, some the methods used for the original body fat data set will be carried out for the new larger set and new models produced. These models will then be compared to see how much they differ from one another and also from the model generated using the full data set with no missing values. It can then be judged which methods appear to be more accurate.

4.2 Missing Values

To generate the models, first some data have to be made 'missing'. To remove some of the data, a loop will be written in R for each missingness mechanism, which can remove some of the data values and replace them with NA. The seed will be set for each of the randomly generated sections, to ensure the same observations are deleted each time the loop is run. The R code for these loops can be found in appendix B. Once some values have been set to NA, this can be used as a basis for each of the analysis methods. Each method will be run five times, with different levels of missingness in each variable; $10\%, 20\%, 30\%, 40\%, 50\%$. Tables 4.3, 4.4 and 4.5 will show the number of complete cases once the missingness has been generated. These tables are used because the number of complete cases remaining cannot be easily estimated from solely the percentage of missingness and the method used to generate the missing values. For example, if there are 10% of values missing from each variable, this will usually result in more then 10% of subjects missing from the complete case analysis, since it is unlikely that the same 10% of subjects will be selected for each variable. Note that in this example, missingness will be generated in the all four variables for the data which are MCAR. For MAR and MNAR, missingness will only be generated in the three predictors. Image plots, see example in figure 4.4, will also be shown in figures 4.7, 4.11 and 4.15 which show the missingness generated for each mechanism. The plots show each of the missingness percentages, $10\%, 20\%, 30\%, 40\%, 50\%$, with the x-axis showing the four variables; fat, triceps, thigh and midarm, and the y-axis showing the 1000 subjects. The present subjects will be shown in red and those missing will be shown in white.

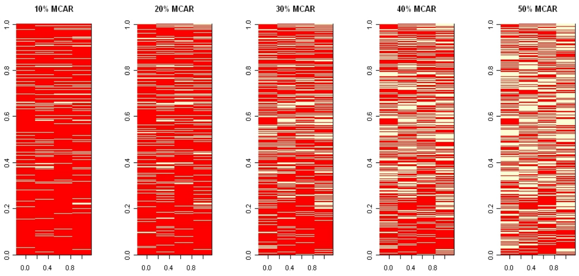


Figure 4.4: An image plot

The plots in figures 4.7, 4.11 and 4.15 can be summarised in a table using the `md.pattern` command in the ‘mice’ package in R. This summarises how many subjects have each combination of missing/present variables. There are five columns; one for each of the four variables plus a final column showing the number of missing variables. In the left margin there is also a list showing the number of subjects which have the given missingness pattern. A value of 1 represents a variable is observed, whilst a value of 0 indicates a missing value. The rows show the different patterns of missingness which can occur. The rows will only show combinations of missingness which occur amongst the subjects. Note there are up to 2^n rows, where n is the number of variables. The final row then shows the totals from the columns. This summary can then be used to assess whether there appears to be any obvious patterns in the data in terms of which variables are missing.

```
> md.pattern(x)  #md.pattern command
   thigh.i fat.i triceps.i midarm.i
84       1     0         1        1    1
79       1     1         0        1    1
.        .     .         .        .    .
.        .     .         .        .    .
```

Another summary method is to investigate pairs of the variables. In a pair of variables for each subject, there are four options; recorded/recorded, recorded/missing, missing/recorded or missing/missing. Studying the pairs of variables offers another way in which to find patterns in the data. The `md.pairs` command in the 'mice' package in R produces four tables giving the number of subjects with each combination of missingness.

```
            fat.i  triceps.i  thigh.i  midarm.i  #md.pairs command
fat.i         490        204      236       196  for recorded/recorded
triceps.i     204        488      238       205
thigh.i       236        238      533       227
midarm.i      196        205      227       475
```

Margin plots, see figure 4.5 for an example, are a graphical summary which can be used to investigate the missingness mechanism. These are shown in figures 4.8, 4.12 and 4.16. The blue dots show the subjects which have observed values for both the variables considered. The red dots show the subjects which have one of these two variables missing. The one observed value for that subject is plotted at the correct point along the observed variable axis. The orange dots simply show those subjects who have missing values for both variables. The box plots along each axis summarise the observed points in blue and missing points in red. For the data to be MCAR it is expected that these two box plots, along each axis, are similar.

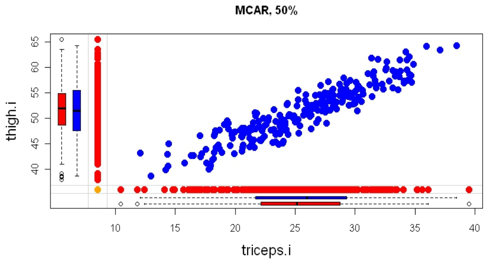


Figure 4.5: A margin plot

To form the complete case analysis, those subjects containing one or more missing value will be excluded from the analysis by removing their entire row. The model can then be run with the remaining complete subjects. The pairwise deletion method will not be conducted as it will produce the same results as the complete case analysis, since all the variables are being used. Next, the missing values will be replaced with the mean of the remaining subjects for the relevant variable. The model will then be generated using this amended data set. Median imputation will not be included as the multivariate normal distribution of the data would cause the results to be very similar to those obtained for mean imputation. Stochastic regression imputation will also be used and will be the basis for multiple imputation, which will be carried out using both 10 and 100 imputations. Note there are other methods which could be selected; these are show in table 4.2. Further details for each of these methods can be found in Van Buuren & Groothuis-Oudshoorn (2011). Predictive mean matching is the same as pattern matching or hot deck imputation whereby the missing value is replaced with the value from another similar subject. The two forms of regression imputation and mean imputation have also previously been described, see chapter 3. Once generated, these estimates can be compared with the 'original model' from the data set with 1000 subjects with no values missing, that is,

$$\text{bodyfat} = 117.09 + 4.33(\text{triceps}) - 2.86(\text{thigh}) - 2.19(\text{midarm}). \tag{4.1}$$

Method	Description	Scale Type
pmm	Predictive mean matching	Numeric
norm	Bayesian linear regression	Numeric
norm.nob	Linear regression, non-Bayesian	Numeric
mean	Unconditional mean imputation	Numeric
2l.norm	Two-level linear model	Numeric
logreg	Logistic regression	Factor, 2 levels
polyreg	Polytomous (unordered) regression	Factor, > 2 levels
lda	Linear discriminant analysis	Factor
sample	Random sample from the observed data	Any

Table 4.2: Table showing the possible methods for imputation in 'mice', taken from Van Buuren & Groothuis-Oudshoorn (2011), Table 1

All calculations will be carried out using the 'mice' package in R, see §3.8 for more details. The prediction matrix which will be used for these calculations is the matrix generated in R using the correlations between the variables. The default setting will be used, which considers the variables to be predictors if their correlation is above 0.1. This results in the following prediction matrix being generated,

```
> pred #quickpredbodyfat - default of 0.1
          fat.i triceps.i thigh.i midarm.i
fat.i         0         1       1        1
triceps.i     1         0       1        1
thigh.i       1         1       0        0
midarm.i      1         1       0        0
```

Note this is easy to generate manually using the correlation matrix as shown below. The default setting of 0.1 can be altered to form different prediction matrices depending upon the nature of the data. The threshold can also be specified for each of the columns if necessary. Alternatively, the prediction matrix can be set to include only predictors with a certain percentage of usable cases. Two example are shown below.

```
> cor(sample)  #correlation matrix for the data
         Fat  Triceps Thigh Midarm
Fat     1.00    0.84 0.878  0.142
Triceps 0.84    1.00 0.924  0.458
Thigh   0.88    0.92 1.000  0.085
Midarm  0.14    0.46 0.085  1.000

> pred  #correlation threshold set at 0.5 for all variables
          fat.i triceps.i thigh.i midarm.i
fat.i         0         1       1        0
triceps.i     1         0       1        0
thigh.i       1         1       0        0
midarm.i      0         0       0        0
```

```
> pred   #minimum usable percentage set at 50 for each variable
          fat.i triceps.i thigh.i midarm.i
fat.i         0         1       1        1
triceps.i     1         0       1        1
thigh.i       1         1       0        0
midarm.i      1         1       0        0
```

Graphs will then be produced for each of the different mechanisms which compare the different methods, see figure 4.6 for an example. Figures 4.9, 4.10, 4.13, 4.14, 4.17 and 4.18 show these summary graphs, two each for MCAR, MAR and MNAR. The first line, in black, shows the original model with no missing data, which the remaining models can be compared to. The circles represent the estimates, while the lines show the 95% confidence intervals for these estimates. Five methods are included; complete case analysis (red), mean imputation (blue), (stochastic) regression imputation (orange), multiple (10) imputation (pink) and multiple (100) imputation (purple). Stochastic regression imputation was used for the multiple imputation. All methods were carried out using the 'mice' package in R, see §3.8. The methods were carried out for $10\%, 20\%, 30\%, 40\%$ and 50% missingness. The five lines for each method show these increasing percentages of missingness.

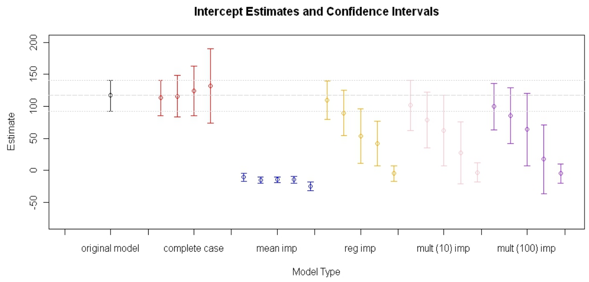


Figure 4.6: A model summary plot

4.3 Missing Completely at Random, MCAR

Figure 4.7 shows the image plot for the data which are MCAR. Note the percentage missing is for each of the variables. For easier interpretation, table 4.3 shows the number of complete cases after the missing values have been generated. As the R code used to produce this missingness was fairly long, the images are a useful method for checking the commands have produced the desired missingness and that no patterns are present amongst the missing values. It can be seen that the missingness appears to be as intended, therefore the models can now be produced using the different methods for coping with missingness.

Below is an example showing the `md.pattern` command which is the summary for the data which are MCAR for 50% missingness. For example, the first row of numbers shows there are 84 subjects from 1000 who have their fat reading missing, but have observed values for thigh, triceps and midarm. The final column shows that this combination of missingness results in one variable from four missing. There are no patterns in this summary, which is as expected as the observations are MCAR.

Percentage Missing	Number of Complete Cases (from 1000)
10	800
20	600
30	400
40	200
50	0

Table 4.3: Table of complete cases after missingness: MCAR

```
> md.pattern(x)  #mcar, percentage missing: 50
   thigh.i fat.i triceps.i midarm.i
84       1     0         1        1    1
79       1     1         0        1    1
71       0     1         1        1    1
86       1     1         1        0    1
64       1     0         0        1    2
50       0     0         1        1    2
46       0     1         0        1    2
68       1     0         1        0    2
71       1     1         0        0    2
47       0     1         1        0    2
81       0     0         0        1    3
81       1     0         0        0    3
82       0     0         1        0    3
90       0     1         0        0    3
       467   510       512      525 2014
```

The `md.pairs` command is shown below as an example for the data which are MCAR for 50% missingness. It can be seen that there are no obvious patterns in the data, which is as expected since the observations are MCAR.

```
> md.pairs(x) #mcar 50
$rr #recorded#recorded
          fat.i triceps.i thigh.i midarm.i
fat.i       490       204     236      196
triceps.i   204       488     238      205
thigh.i     236       238     533      227
midarm.i    196       205     227      475
```

```
$rm #recorded/missing
          fat.i triceps.i thigh.i midarm.i
fat.i         0       286     254      294
triceps.i   284         0     250      283
thigh.i     297       295       0      306
midarm.i    279       270     248        0

$mr #missing/recorded
          fat.i triceps.i thigh.i midarm.i
fat.i         0       284     297      279
triceps.i   286         0     295      270
thigh.i     254       250       0      248
midarm.i    294       283     306        0

$mm #missing/missing
          fat.i triceps.i thigh.i midarm.i
fat.i       510       226     213      231
triceps.i   226       512     217      242
thigh.i     213       217     467      219
midarm.i    231       242     219      525
```

Figure 4.8 shows an example margin plot for the data which is MCAR. It shows the triceps and thigh variables which are MCAR for 50% missingness. It can be seen in the example in figure 4.8 that the box plots are similar and hence this suggests the data are MCAR.

Figures 4.9 and 4.10 show the results from the models produced using data which is MCAR. Notice how for the complete case analysis, there are only four lines, this is because at 50%, which is missing from the plot, there are no complete cases left in the data set, see table 4.3.

It can be seen that as the percentage of missing values increases, the confidence intervals for the complete case analysis estimates increase, as there is less data available. However, the confidence intervals for the mean imputation estimates generally decrease as the percentage of missing values increases. This is because the missing values are replaced with the mean values, which results in small standard error values, hence shrinking the size of the confidence interval. The regression imputation methods, both single imputation and multiple imputation, have varying confidence intervals as the percentage of missingness increases.

The complete case analysis method copes well, even with 50% of the data missing. The estimates remain within the confidence intervals of the original model and they do not change much from the original model. The other methods do not cope as well when the percentage of missing data increases. Note how the mean imputation estimates are far from the original model estimates, with very small confidence intervals in comparison to the other methods, see §4.7 for more information. The regression imputation, both single and multiple, estimates cope relatively well when there is only around 10% missingness, but there is then a pattern away from the original model estimate as the percentage of missingness increases. Multiple imputation aims to allow for the added uncertainty in the imputed values and hence the confidence intervals for the multiple imputation estimates are wider than the corresponding ones for the single regression

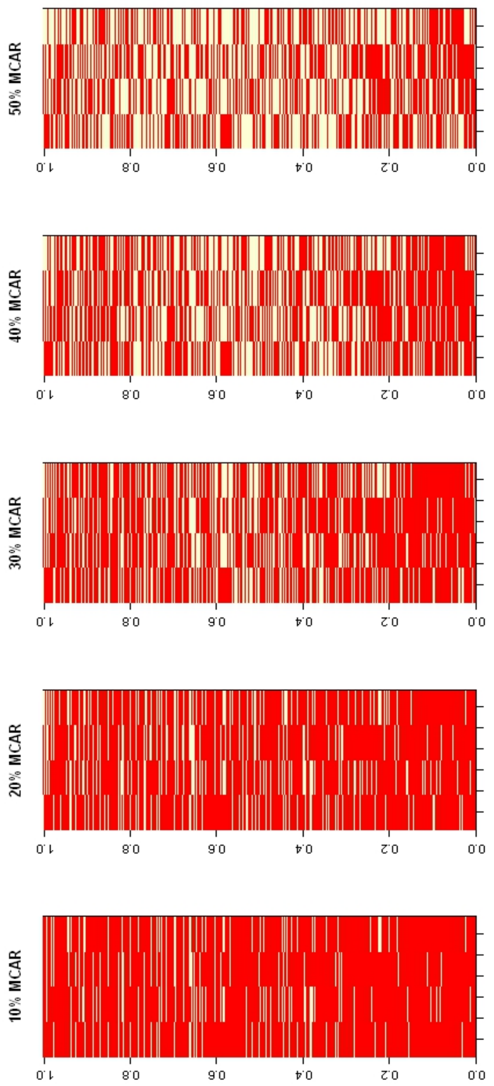


Figure 4.7: Image plot for the body fat data, 1000 subjects, MCAR

imputation estimates, for example when comparing 10% single regression imputation with 10% multiple regression imputation. However, the 50% regression estimates, for both single and imputation, have very small confidence intervals. This is due to the particular values which were made to be MCAR. If this process was repeated with different missing values, it would be expected that the confidence interval for 50% missingness would be larger than the one for 40% missingness, to allow for the added uncertainty in the increasing number of imputed values.

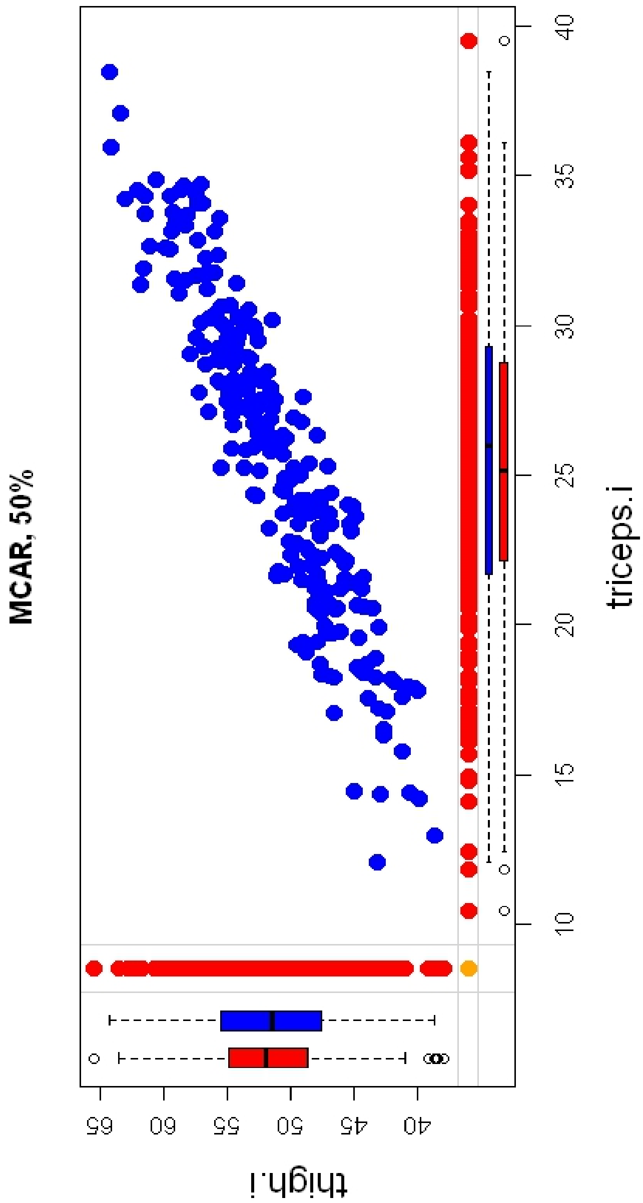


Figure 4.8: Triceps and thigh margin plot for the 1000 subjects, 50% *MCAR*

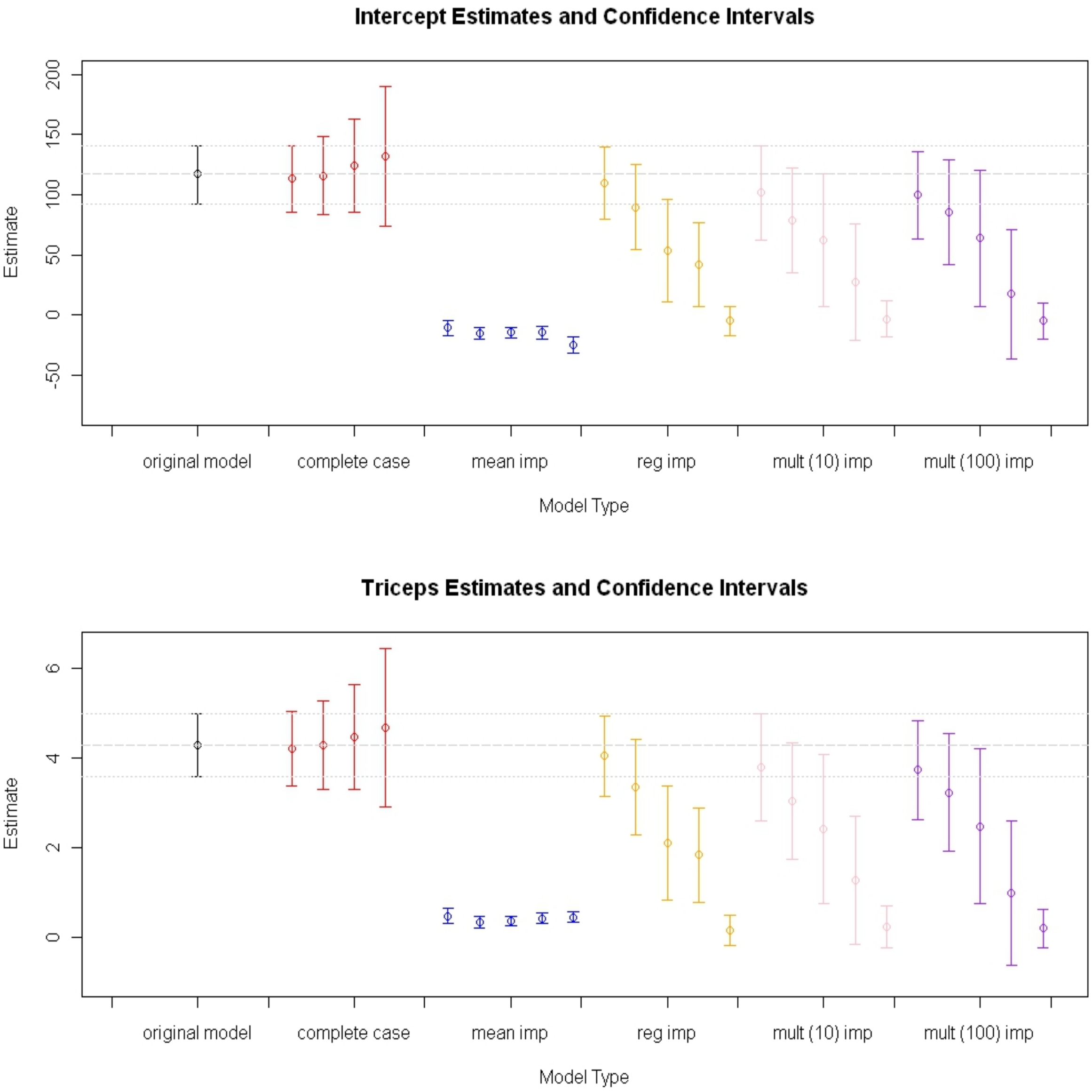
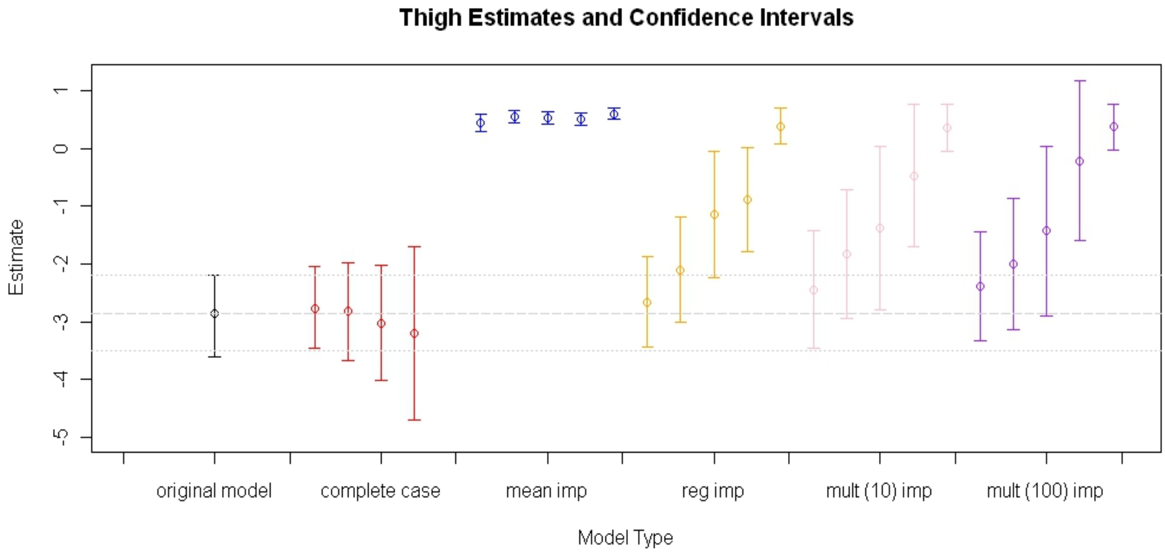


Figure 4.9: Estimates and confidence intervals for the intercept and triceps: 10-50 percent MCAR per variable



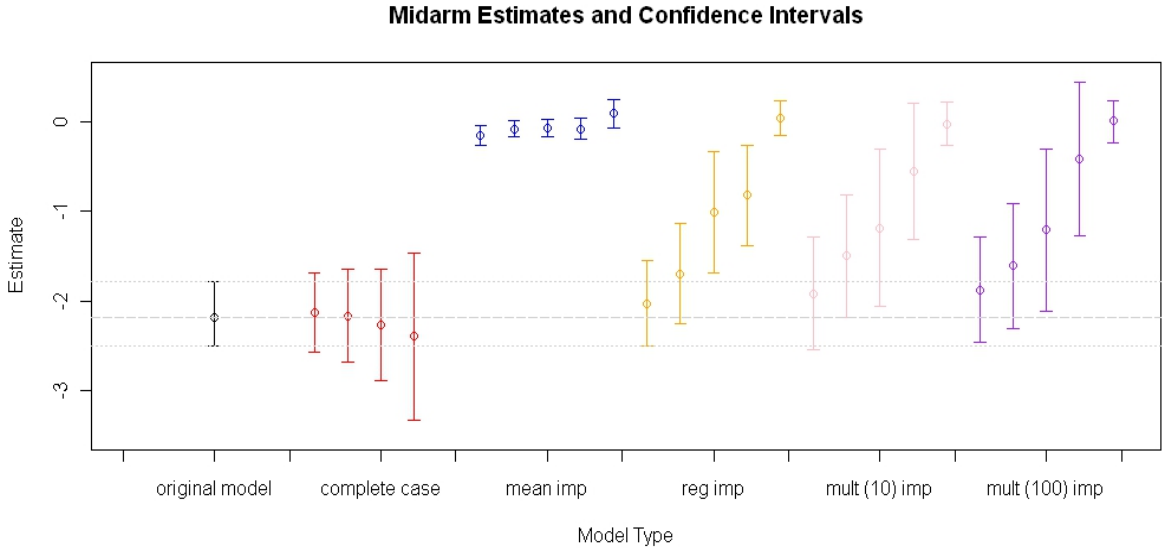


Figure 4.10: Estimates and confidence intervals for thigh and midarm: 10-50 percent MCAR per variable

4.4 Missing at Random, MAR

Next, it can be explored how the different methods cope when the data are no longer MCAR. Figure 4.11 shows the image plot for the same data but with some values missing at random, MAR. For values to be MAR, there must be relationships between the variables in the data. One example which can be easily understood is that relating to age and blood pressure. For example, if there are some blood pressure readings missing, it may be that the mean for recorded values is higher than the mean for the true values. However, this pattern may be explained through another variable, that is age. Younger people tend to have lower blood pressure than older people, and younger people are generally less likely to have their blood pressure recorded. Therefore, the missing values in blood pressure can be explained through another variable, age. Now, to have data which is MAR in the body fat data set, there must be relationships present, such as the relationship between the blood pressure and age variables. However, there are no specific reasons to draw such relationships between the measurements of different body parts. Hence, for illustrative purposes, these relationships will be constructed without any reasoning. The relationships used are as follows:

- Order the subjects by their thigh measurement, small to large, and remove the top $x\%$ of triceps values,
- Order the subjects by their thigh measurement, small to large, and remove the top $x\%$ of midarm values,
- Order the subjects by their triceps measurement, small to large, and remove the top $x\%$ of thigh values.

These relationships result in all three predictor variables having $x\%$ of their values removed. As with the MCAR plots, the MAR plots will start with 10% missingness and end in 50% missingness, with increases of 10% missingness. Table 4.4 shows the number of complete cases after the missing values have been generated. The plots in figure 4.11 have all subjects ordered by increasing body fat, hence the missingness in each of the variables is not in clear blocks, but there are patterns resulting from the correlations between the predictors and the outcome. The plots help to check the R code used and show the required patterns for the data to be MAR from the relationships used.

Percentage Missing	Number of Complete Cases (from 1000)
10	876
20	757
30	644
40	545
50	432

Table 4.4: Table of complete cases after missingness: MAR

The plots in figure 4.11 can also be summarised in a table showing how many subjects have each combination of missing/present variables. Below is an example which is the summary for the data which are MAR for 50% missingness. This summary can be used to assess whether there appears to be any obvious patterns in the data in terms of which variables are missing. There are 68 people who have both triceps and midarm measurements missing, recall these are from the people with the top 50% of thigh measurements. There are also 68 people who have their thigh measurement missing only, recall these are from the people with the top 50% of triceps measurements. Finally, there are 432 people who have their triceps, thigh and midarm measurements missing. There are, therefore, patterns in this data, hence the missingness is not MCAR.

```
> md.pattern(x) #MAR, missingness percentage: 50
    fat.i triceps.i thigh.i midarm.i
432     1         1       1        1    0
 68     1         1       0        1    1
 68     1         0       1        0    2
432     1         0       0        0    3
        0       500     500      500 1500
```

Next the pairs of variables can be investigated. It can again be seen that there are patterns in the data as the only values present in the table are 0, 68, 432 and 500. These numbers are also arranged in patterns, hence the data values are not MCAR.

```
> md.pairs(x)   #mar, percentage missing: 50
$rr #recorded/recorded
          fat.i triceps.i thigh.i midarm.i
fat.i      1000       500     500      500
triceps.i   500       500     432      500
thigh.i     500       432     500      432
midarm.i    500       500     432      500

$rm #recorded/missing
          fat.i triceps.i thigh.i midarm.i
fat.i         0       500     500      500
triceps.i     0         0      68        0
thigh.i       0        68       0       68
midarm.i      0         0      68        0

$mr #missing/recorded
          fat.i triceps.i thigh.i midarm.i
fat.i         0         0       0        0
triceps.i   500         0      68        0
thigh.i     500        68       0       68
midarm.i    500         0      68        0
```

```
$mm #missing/missing
           fat.i triceps.i thigh.i midarm.i
fat.i          0         0       0        0
triceps.i      0       500     432      500
thigh.i        0       432     500      432
midarm.i       0       500     432      500
```

Figure 4.12 shows an example margin plot for the data which are MAR. It shows the triceps and thigh variables which are MAR for 50% missingness. For the data to be MAR it is expected that the two box plots, along each axis, are different. It can be seen in the example in figure 4.12 that this is true for the body fat data set.

Figures 4.13 and 4.14 show the estimates and confidence intervals for the data which are MAR. It can be seen that the complete case analysis copes relatively well as all the estimates, even for 50% missingness, are close to the estimates for the original model. The confidence intervals simply widen as the percentage of missing values increases, as there is less information available for the complete case analysis. The mean imputation estimates are, as with the MCAR estimates, far from the original model estimates, see §4.7. The regression imputation estimates, both single and multiple imputation, seem to cope relatively well with the data MAR. However, as the missingness increases, the estimates move further away from the original model estimate and the confidence intervals increase as more imputations are carried out. This is expected, since the estimates are likely to be less accurate when more values are missing and the wider confidence intervals allow for the added uncertainty when more values are imputed. It can be seen that the multiple imputations cope better than the single regression imputation. This is because the single regression imputation can change each time the method is run in R and hence different values imputed. Therefore, multiple imputation can be more useful as it runs the imputation many times and combines the results to produce one summary output, whilst allowing for the added uncertainty involved. As the number of imputations increases, from one in the regression imputation method to 10 and then 100 in the multiple imputation methods, it can be seen that the estimates and their confidence intervals become more stable. The estimates from 10% to 50% missingness begin to form a pattern and the confidence intervals slowly increase to allow for the increasing number of imputations. Note that the performance of all methods will be dependent on the particular values which are made 'missing'. For example, mean imputation will cope better if values close to the mean are removed, opposed to values far from the mean. Mean imputation may, therefore, cope better in data sets where the variance of each variable is relatively small. Regression imputation may also be affected by which particular values are made missing. It is likely to cope better if the values removed are ones which will not affect the regression line, rather than those which will cause the line to be moved far from the original regression line.

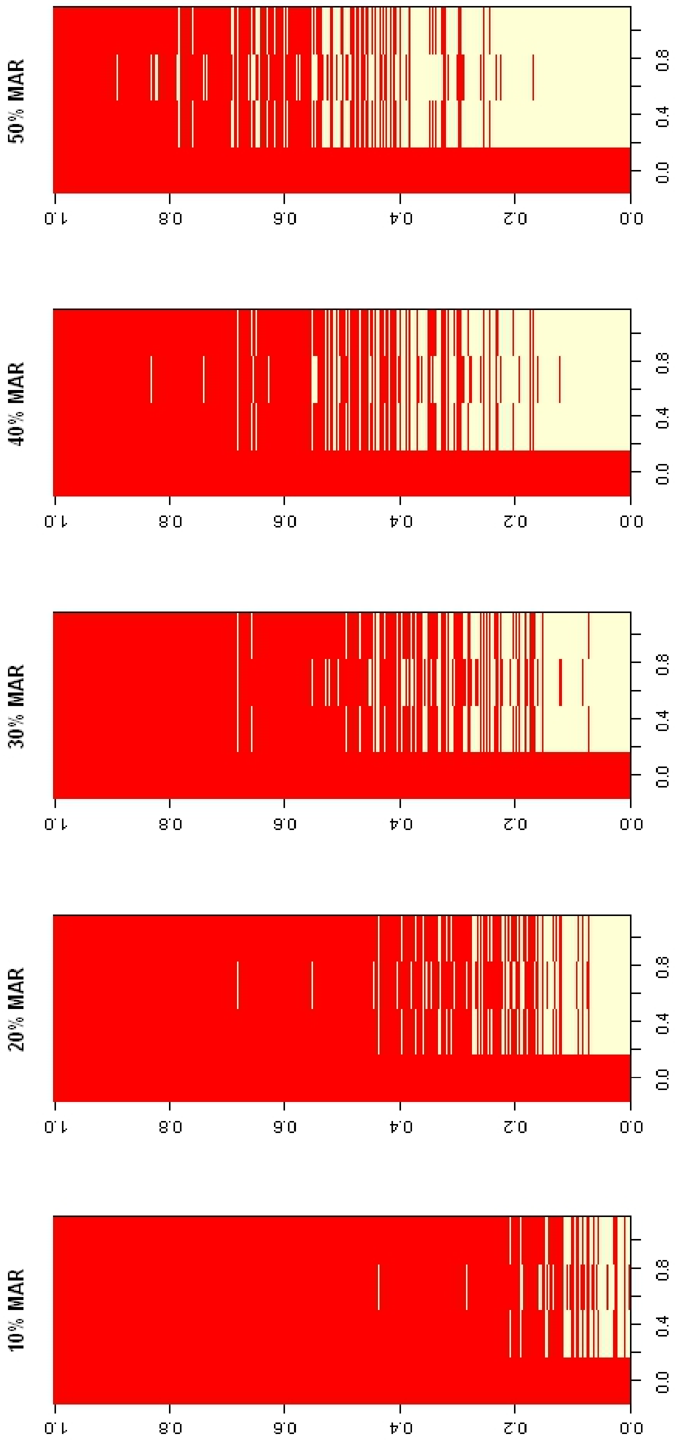


Figure 4.11: Image plot for the body fat data, 1000 subjects, MAR

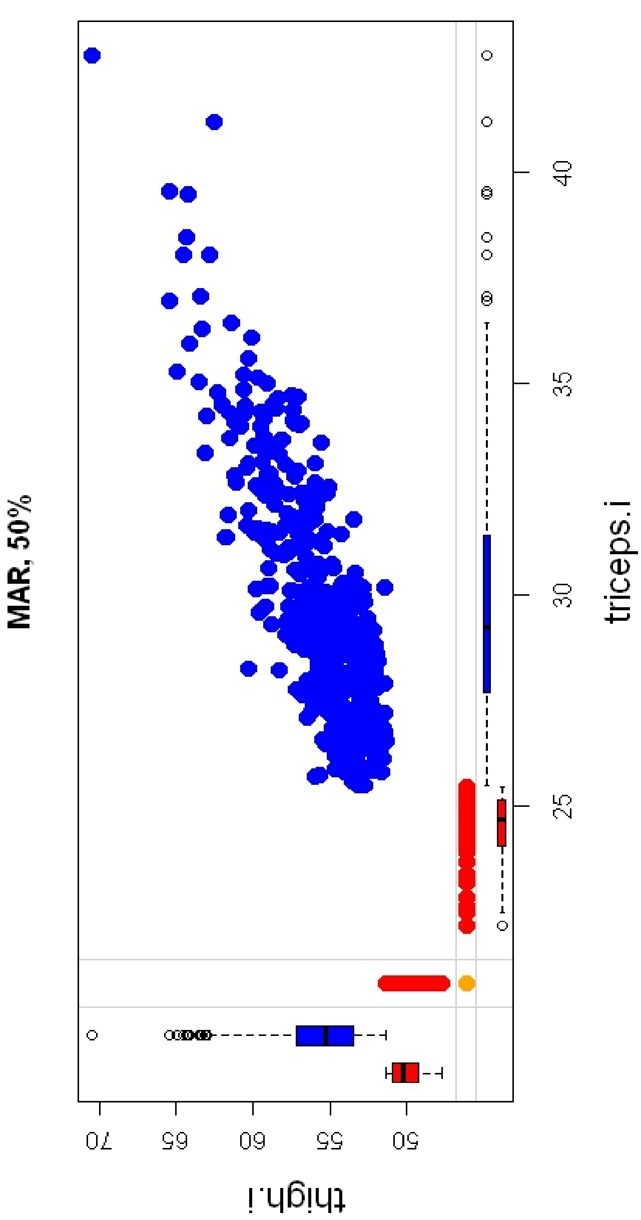
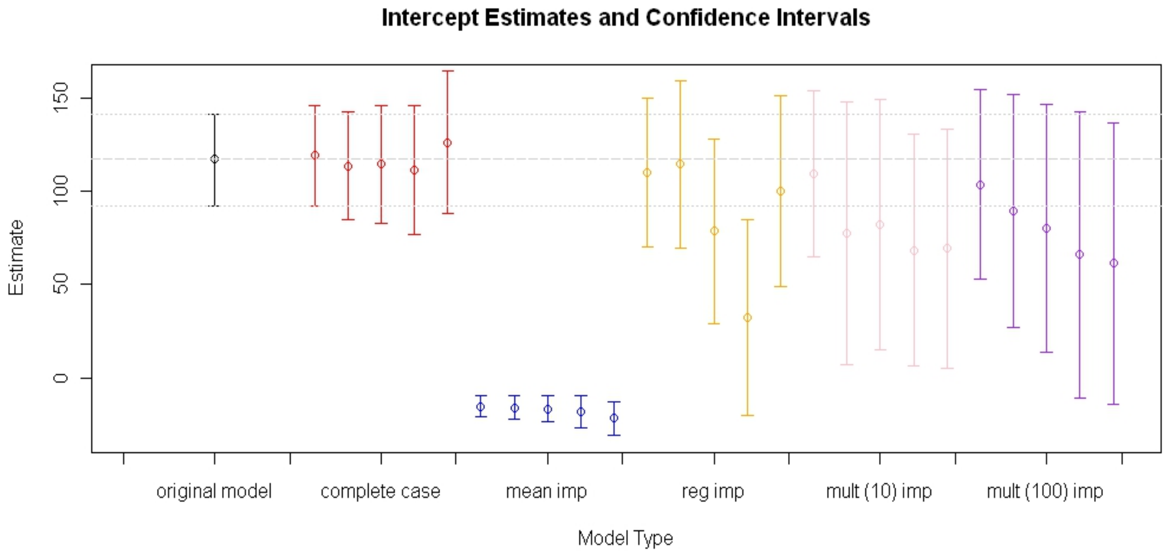


Figure 4.12: Triceps and thigh margin plot for the 1000 subjects, 50% *MAR*



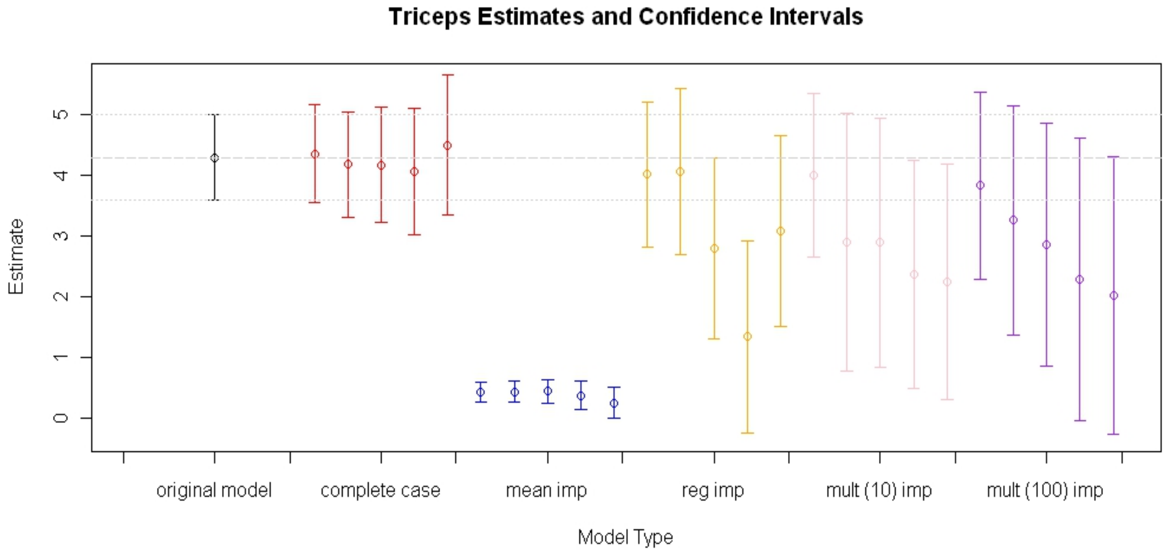


Figure 4.13: Estimates and confidence intervals for the intercept and triceps: 10-50 percent MAR for 3 relationships

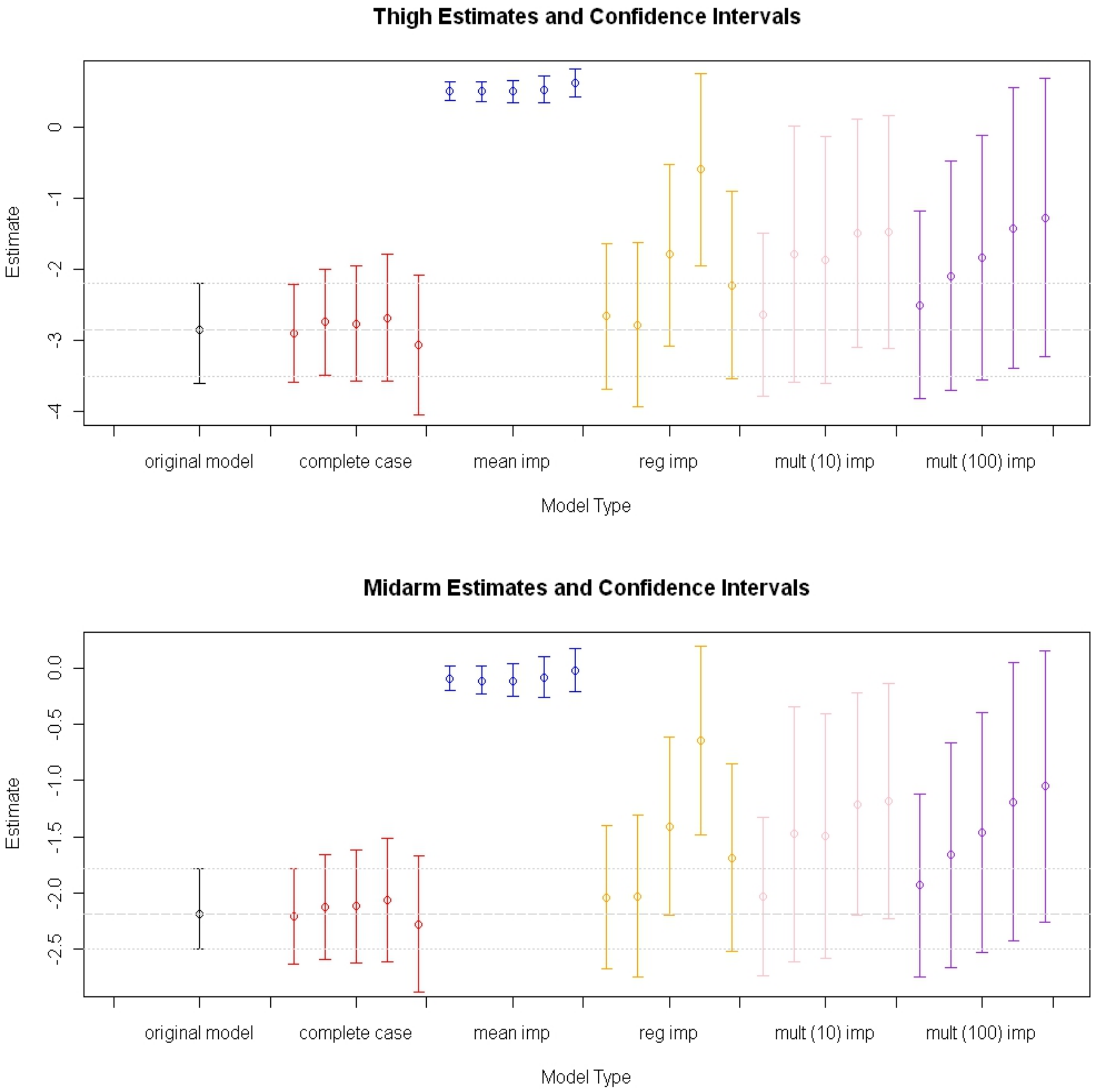


Figure 4.14: Estimates and confidence intervals for thigh and midarm: 10-50 percent MAR for 3 relationships

4.5 Missing Not at Random, MNAR

Finally, it can be investigated how the different methods cope when the data are missing not at random. For the data values to be missing not at random, there must be values missing which cannot be explained through another variable and the missingness must not be completely random. One example involves blood pressure readings. The recorded blood pressure readings from a group of people, where some values are missing, may be lower than the true readings for that group of people. This may be because those with high blood pressure may have problems associated with high blood pressure, such as headaches, and therefore may be more likely to miss their appointments. Reasoning like this is not as easy for the body fat data set, but patterns like this can still be constructed for the purposes of creating data values which are missing not at random. The patterns used are as follows:

- Order by triceps measurements, then remove the smallest $x\%$ of triceps measurements,
- Order by thigh measurements, then remove the smallest $x\%$ of thigh measurements,
- Order by midarm measurements, then remove the smallest $x\%$ of midarm measurements.

This will result in each predictor having $x\%$ removed, with x starting at 10 and increasing, in stages of 10, up to 50. Figure 4.15 shows the data with values MNAR, and table 4.5 shows the number of complete cases for each percentage of missingness. The subjects in figure 4.15 are in increasing order of body fat, hence the predictors are not quite as expected. However, there are patterns in the figure as there are correlations between the outcome and predictors. The plots help to show the missingness is as required, hence checking the R commands used.

Percentage Missing	Number of Complete Cases (from 1000)
10	805
20	641
30	492
40	371
50	266

Table 4.5: Table of complete cases after missingness: MNAR

The plots in figure 4.15 can also be summarised in a table showing how many subjects have each combination of missing/present variables. Below is an example which is the summary for the data which are MNAR for 50% missingness. This summary can be used to assess whether there appears to be any obvious patterns in the data in terms of which variables are missing. It can be seen that there are obvious patterns in the values in the margin to the left of the table; 266, 68, 166, 166, 68, 266. As with the data which were MAR, the data which are MNAR have patterns which, in this case, correspond to the methods used to create the missingness. As there are patterns in this data, the missingness is not classed as MCAR.

```
> md.pattern(x) #MNAR, percentage: 50
    fat.i triceps.i thigh.i midarm.i
266     1         1       1        1    0
 68     1         1       0        1    1
166     1         1       1        0    1
166     1         0       0        1    2
 68     1         0       1        0    2
266     1         0       0        0    3
        0       500     500      500 1500
```

Now the pairs of variables can be investigated one at a time. It can be seen that there are patterns in the data as there are only a few values present in the tables and they are arranged in patterns. As there are patterns in the data, the missingness is not considered to be MCAR.

```
> md.pairs(x) #MNAR, percentage: 50
$rr #recorded/recorded
          fat.i triceps.i thigh.i midarm.i
fat.i      1000       500     500      500
triceps.i   500       500     432      334
thigh.i     500       432     500      266
midarm.i    500       334     266      500

$rm #recorded/missing
          fat.i triceps.i thigh.i midarm.i
fat.i         0       500     500      500
triceps.i     0         0      68      166
thigh.i       0        68       0      234
midarm.i      0       166     234        0

$mr #missing/recorded
          fat.i triceps.i thigh.i midarm.i
fat.i         0         0       0        0
triceps.i   500         0      68      166
thigh.i     500        68       0      234
midarm.i    500       166     234        0
```

```
$mm #missing/missing
           fat.i triceps.i thigh.i midarm.i
fat.i          0         0       0        0
triceps.i      0       500     432      334
thigh.i        0       432     500      266
midarm.i       0       334     266      500
```

Figure 4.16 shows an example margin plot for the data which are MNAR. It shows the triceps and thigh variables which are MNAR for 50% missingness. For the data to be MNAR it is expected that the two box plots, along each axis, are different. It can be seen in the example in figure 4.16 that this is true for the body fat data set. Note that is it not possible to distinguish between the data which are MAR in figure 4.12 and those which are MNAR in figure 4.16 from these plots.

Figures 4.17 and 4.18 show the model estimates and 95% confidence intervals produced when the data are MNAR. It can be seen from figures 4.17 and 4.18 that the complete case analysis again performs well, even with 50% missingness. Although, as the percentage of missing values increases, the confidence intervals become wider. This is expected since there are less subjects included in the complete case analysis as the percentageof increases. The mean imputation estimates, as with MCAR and MAR, also give estimates far from the original model estimates and with small confidence intervals. More information on mean imputation can be found in §4.7. It can be seen that the regression imputations, both single and multiple, perform less well compared with MCAR and MAR. Previously, with the MCAR and MAR estimates, the estimates for 10% and sometimes 20% missingness were included within the confidence intervals of the original model esimate. However, with the MNAR plots, none of the regression imputation estimates, single or multiple imputation, are included within the original model confidence intervals. Therefore, it would appear that these methods cope less well when the data are MNAR. Even when multiple imputation is used, with 100 imputations, the estimates are still not as close to the original model, even when missingness is just 10%. Note, however, how the multiple imputation estimates become more stable than the single regression imputation estimates. The multiple regression imputation estimates form a pattern, as they are the combination of many different regression imputations, to give one more accurate estimate. As with the MAR values, the performance of the different methods will depend upon which particular values are MNAR.

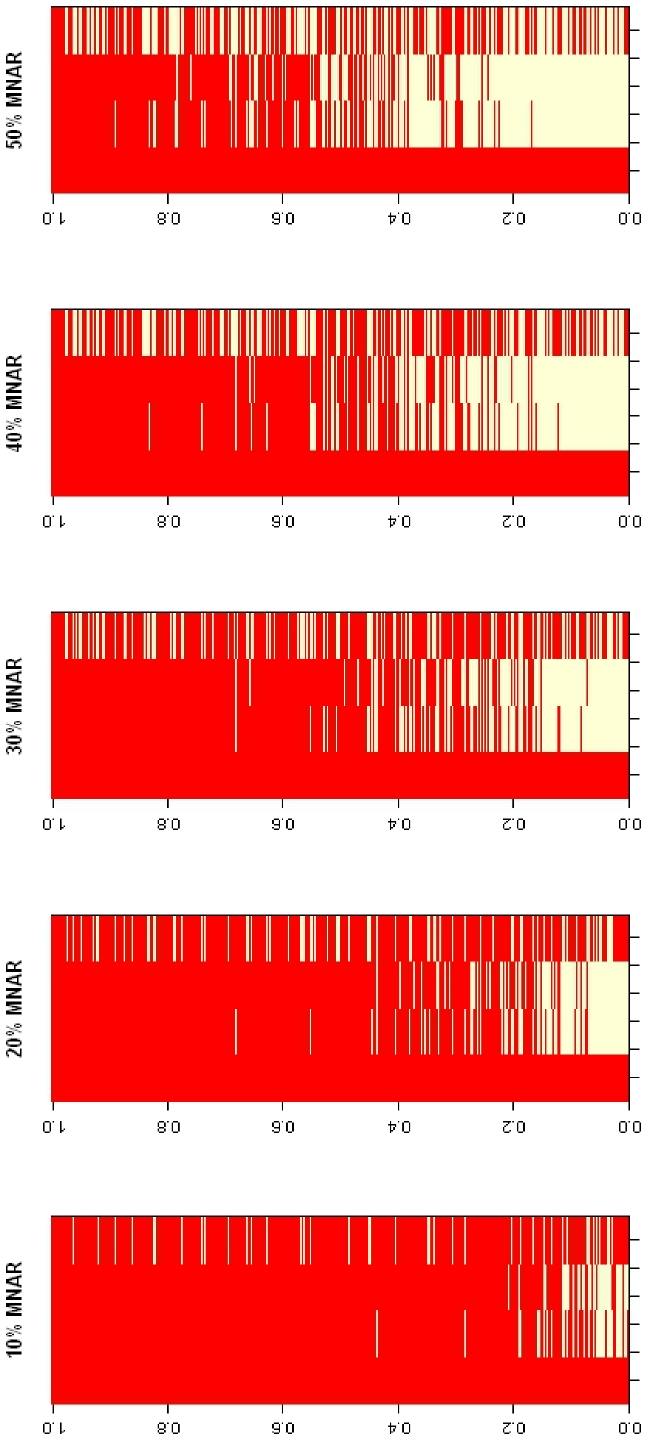


Figure 4.15: Image plot for the body fat data, 1000 subjects, MNAR

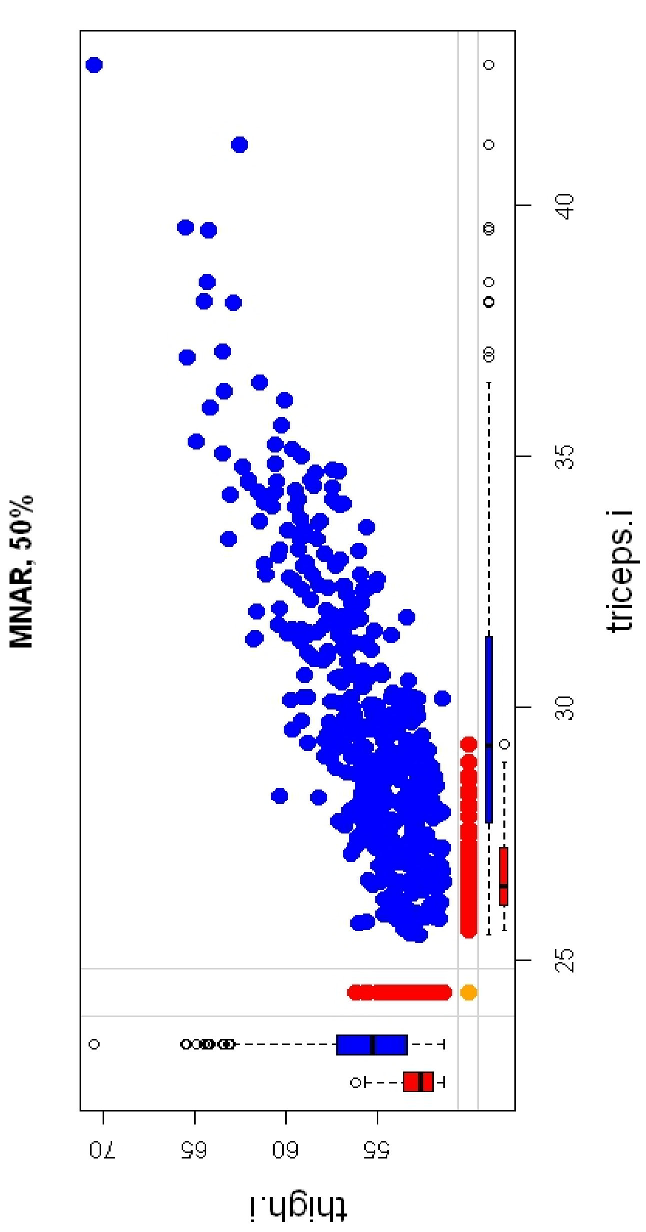


Figure 4.16: Triceps and thigh margin plot for the 1000 subjects, 50% *MNAR*

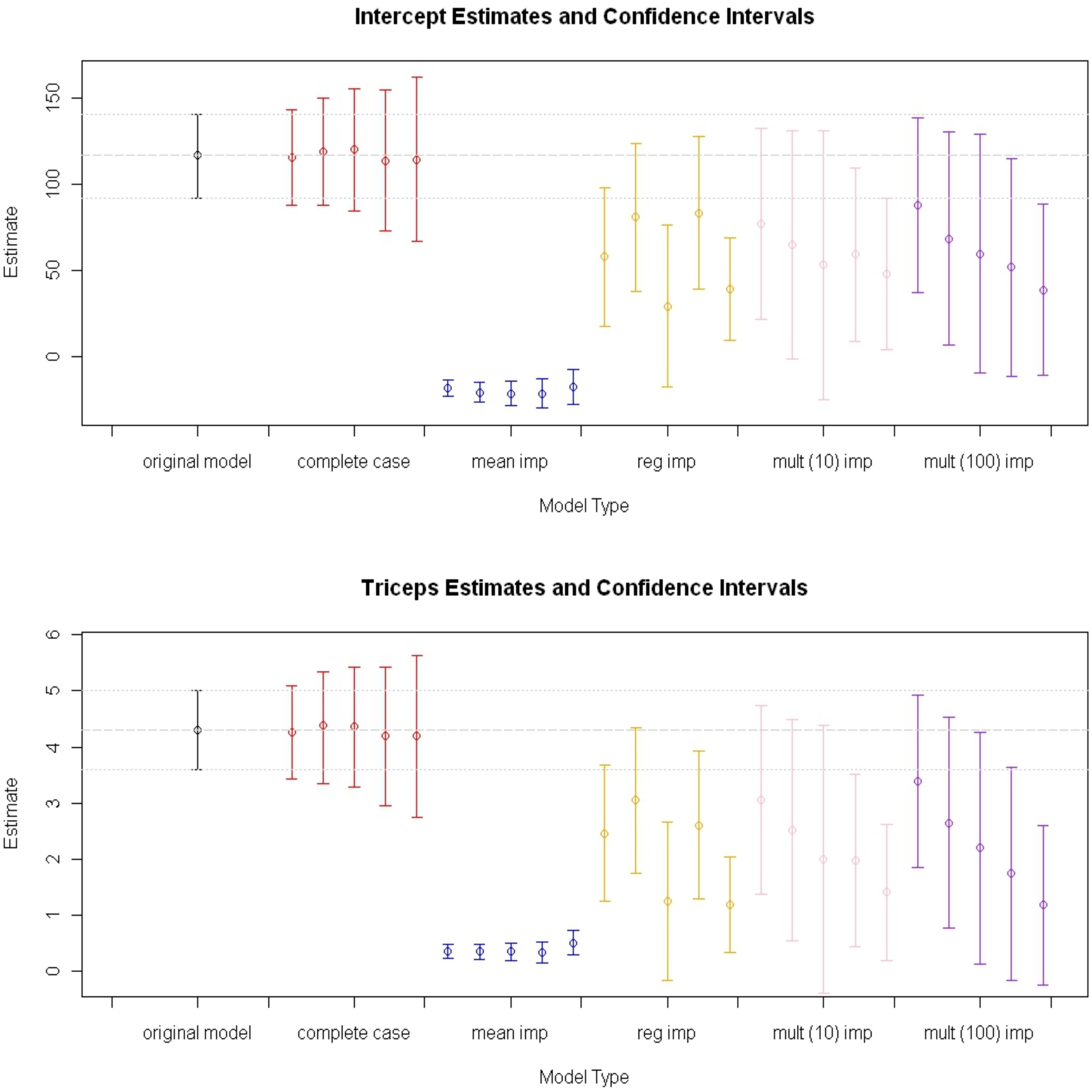
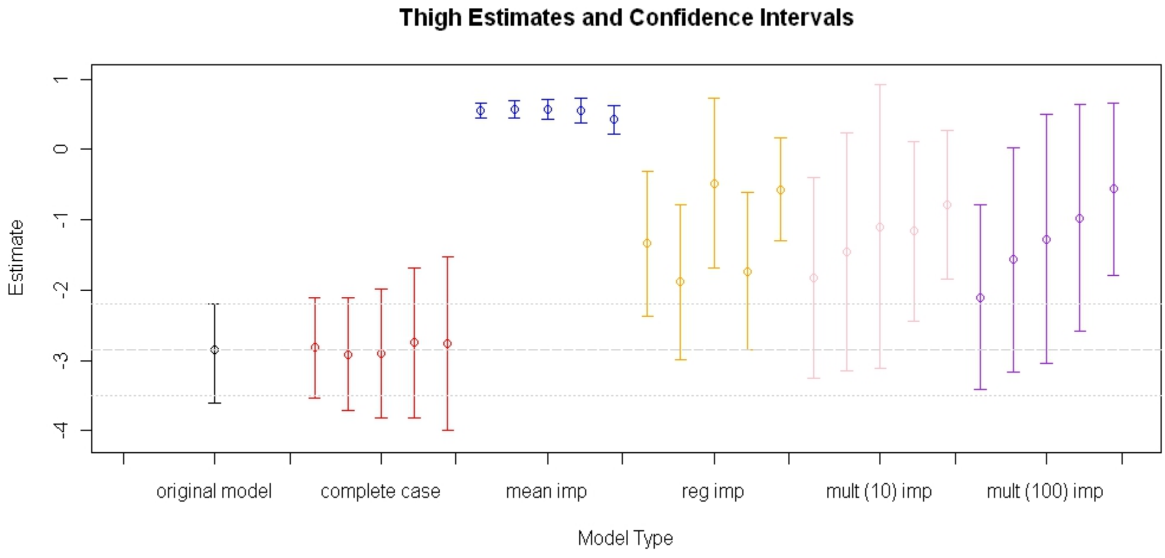


Figure 4.17: Estimates and confidence intervals for the intercept and triceps: 10-50 percent MNAR for each predictor



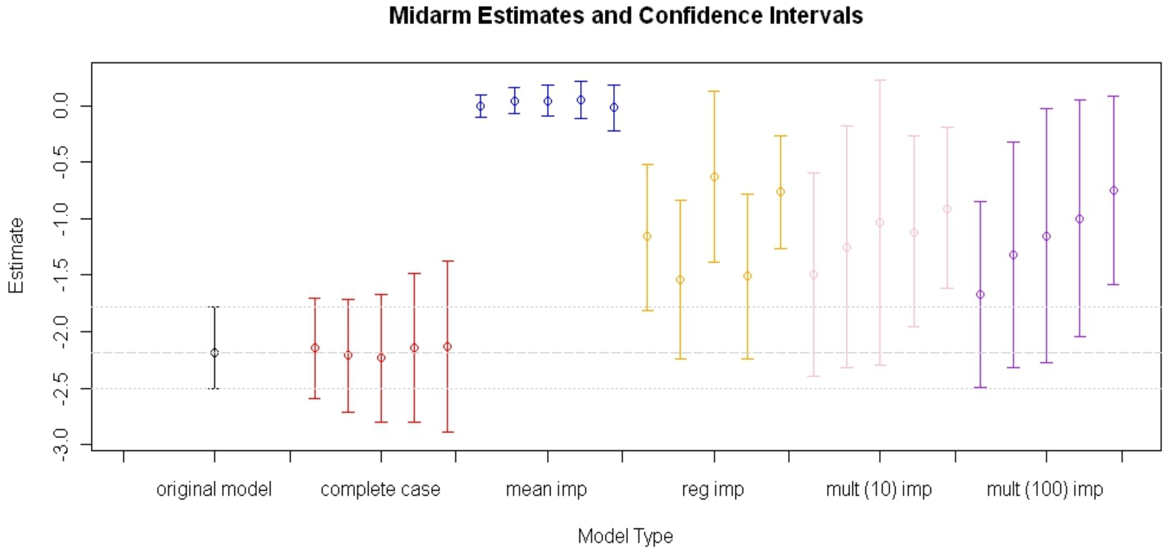


Figure 4.18: Estimates and confidence intervals for thigh and midarm: 10-50 percent MNAR for each predictor

4.6 Summary: Different Methods and Missingness Mechanisms

The body fat data set with 1000 subjects has been used to compare five different methods for dealing with missing data, across three different missingness mechanisms; MCAR, MAR and MNAR. The original data set, with no missing values, provided a basis against which to compare these five different methods.

It has been seen that the complete case analysis appears to be the method which copes most well throughout all types of missingness. Mean imputation appears to be the method which copes least well. An investigation into why the mean imputation estimates were so poor is carried out in §4.7. The complete case analysis copes better than expected, as in theory, it should only cope well when the data are MCAR. The good performance for the MAR and MNAR data may be unexpected. However, there may be reasons for this, which lie in the nature of the data. The data is multivariate normally distributed and some variables, such as triceps and thigh, are highly correlated; 0.92 (2dp). Consider figure 4.19 which is a simple graph showing just the triceps and thigh variables for the 1000 subjects. The graph contains all the 1000 subjects, that is, there are no missing values. The red line shows the regression line through these variables.

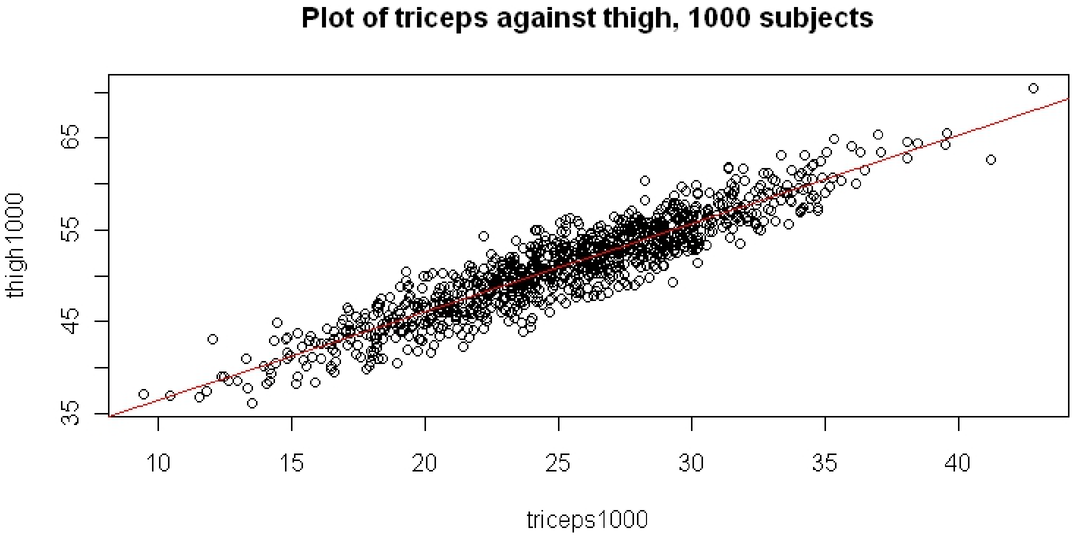


Figure 4.19: Body fat data, 1000 subjects, triceps against thigh

Now, if the data are MCAR, then the points removed should be such that they do not affect the regression line, since there are not any points which would significantly alter the regression line, as the two variables are highly correlated. Figure 4.20 gives an example where the points are MCAR. The black points represent the points which are kept and the red points show those which are missing. The black line shows the original regression line and the red line shows the regression line after 20% of the points are MCAR. Therefore, as the regression lines are very similar, complete case analysis could be expected to cope well when the data values are MCAR.

Next, if the data values are MAR, this means that one variable has been ordered and the say, bottom section of the other variable have been removed. In this case, it may be that the subjects are ordered by triceps measurements, and the 20% of subjects with the smallest triceps measurements have their thigh measurements removed. Note that in practice there would be an underlying reason why the two variables would be connected in such a way. This would

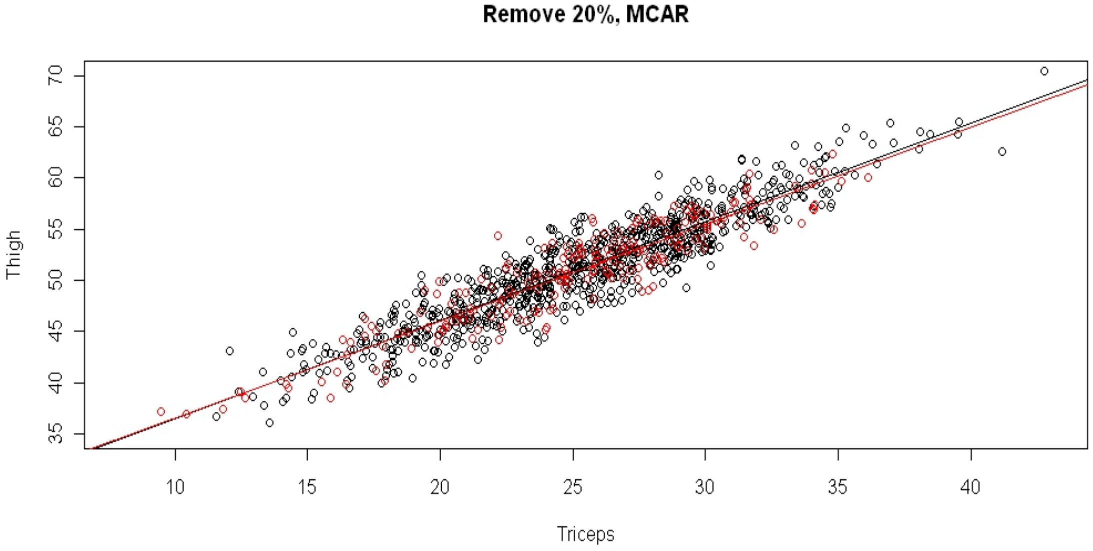


Figure 4.20: Body fat data, 1000 subjects, 20% *MCAR*

result in points from the bottom left corner of the plot being removed, especially since the two variables are so highly correlated, see figure 4.21. Removing these points has very little impact on the regression line through the rest of the data. This holds true even if half of the values are removed, see figure 4.22. Both the spread of the data and the high correlation of the variables help to maintain the original regression line through the variables. This basic idea can then be extended to include all the variables. Therefore, when the points are MAR, the complete case analysis still copes well. The main change as the percentage of missingness increases, is the width of the confidence intervals associated with the estimates, since more subjects are excluded from the analysis.

The same idea can be applied to the data values which are MNAR. Since the two variables are so highly correlated, it is likely that the subjects with say, the smallest 20% triceps measurements which then have their thigh measurements removed, for MAR, are likely to be a very similar group of people to those who simply have their thigh measurements removed as they are in the smallest 20% of thigh measurements, for MNAR. Therefore, for this data set, the complete case analysis would also be expected to cope well when the data values are MNAR, see figure 4.23.

The regression imputation estimates appear to cope relatively well when the data values are MCAR, when the missingness is around 20% or less, but estimates poorly when the missingness is above this percentage. The regression imputation estimates appear to be most accurate when the data are MAR, with the estimates closest when the missingness is around 20% or less. However, the regression imputation estimates do not cope well when the data values are MNAR. The regression imputation estimates are likely to cope well when a small percentage of data values are MCAR since the regression line through the data, see figure 4.19, is unlikely to be affected greatly. Hence the imputed values are likely to be close to the true values. However, as the percentage of missingness increases, the regression line may change depending upon which particular values are missing. If the missing values are those which can affect the regression line, which is more likely as the percentage increases, then the imputed values may not be as accurate. When the data values are MAR, it may be expected that the method copes less well

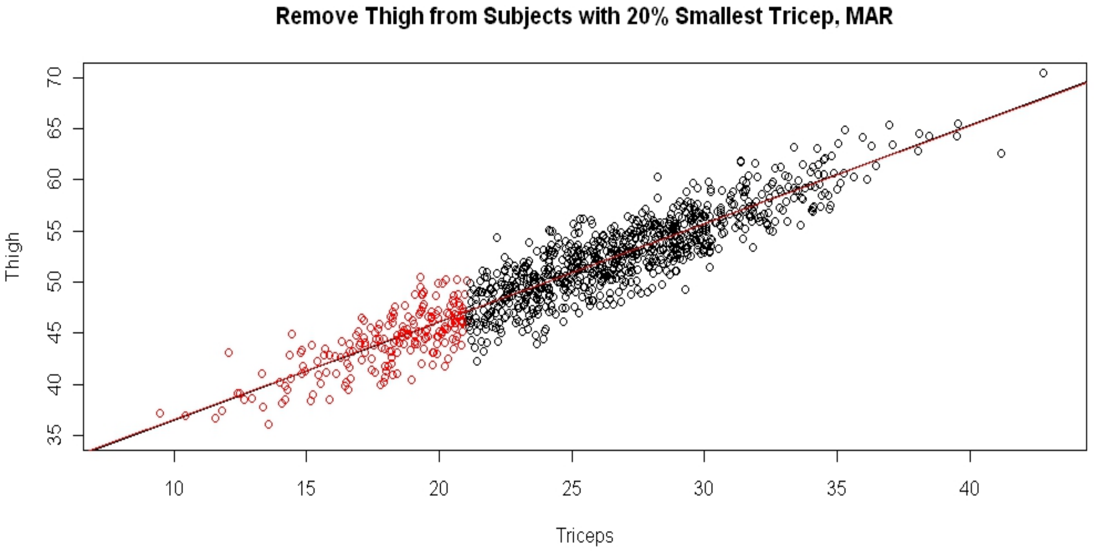


Figure 4.21: Body fat data, 1000 subjects, 20% *MAR*

than it does. However, since some of the variables are highly correlated and well spread, the effect of the missing values is reduced. For example, if a small percentage of the data values in figure 4.19 are MAR and replaced by regression imputed values, the resulting linear regression line is unlikely to be greatly affected, since many of the data points already lie close to the regression line. However, as the percentage of missing values increases, the more likely the regression line is to differ from the original regression line through the data points. This results in the imputed values being less accurate and as a consequence, the final linear regression line may not be as accurate. When the data values are MNAR, the chances of the regression line differing from the original regression line are increased, since the values are removed from one area of the regression line, which makes the new line for imputation more likely to move. This is more likely once the percentage of missingness increases or if the data values are less correlated. However, it can affect all levels of missingness and strengths of correlations. Therefore, it is understandable that the regression imputation estimates do not cope well when the data are MNAR.

The multiple imputation estimates, which were calculated using regression imputation, can be compared with both the original model estimates and the single regression imputation estimates. The multiple imputation methods performed reasonably well for the MCAR and MAR data values, but only when the missingness was around 10% or so. Previous simulation studies by Rubin (1987) suggest that for missingness up to 20%, the number of imputations can be as low as 3. Therefore, using 10 and 100 imputations should be more than sufficient. As the missingness percentage increased, these methods were less accurate. They did not, however, perform well when the data were MNAR. The reasoning behind the different performance levels for the different missingness mechanisms can be extended from the reasoning for the single regression imputation estimates, since the same method was used. However, the multiple imputation estimates can be compared with the single regression imputation estimates within each missingness mechanism. For the MCAR data values, all numbers of regression imputations perform in a similar way. This is because the regression line through the data is less likely to be affected when the data are MCAR, hence the regression lines used for imputation are likely

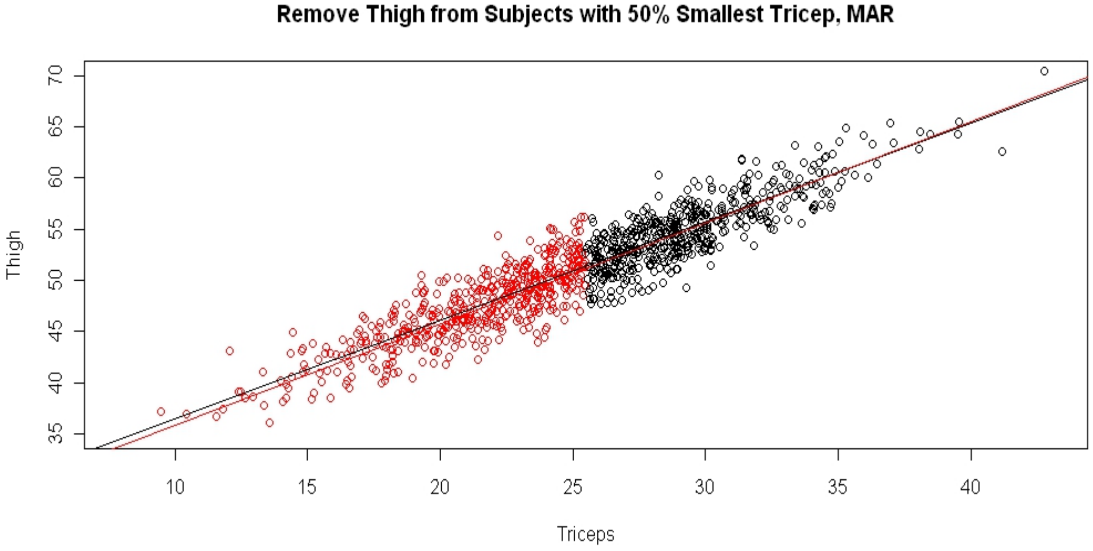


Figure 4.22: Body fat data, 1000 subjects, 50% *MAR*

to all be similar. However, when the data are MAR or MNAR, the regression lines may differ considerably each time the imputation regression line is drawn. This can result in the single imputation estimates being less consistent. Therefore, it is advantageous to use multiple imputation, so an estimate can be taken using a summary of several different imputations. It can be seen from the graphs that the estimates and confidence intervals appear to be far more stable for the multiple imputations that the single regression, with the 100 imputations more so than the 10 imputations.

It has been seen that the different methods cope with varying success over the different missingness machanisms for particular predictors. However, the overall model should also be considered. It appears that for each method, if it over/under-estimates the intercept and triceps coefficient, then it will under/over-estimate the thigh and midarm coefficients. Note how similar the intercept/triceps graph patterns are and the thigh/midarm graph patterns are. Note also how the intercept/triceps estimates on the original model are positive and the thigh/midarm estimates are negative. It appears the methods either cause all estimates to be more positive/negative than they should be, or less so. This assumes that the original model can be considered to be ‘most correct’ solely for the purpose of these comparisons. That is, the estimates are either all further from zero than they should be, or all closer to zero than they should be. For all mechanisms, regression imputation, both single and multiple, and mean imputation, have given all estimates closer to zero than they should be. This is clearer as the percentage of missingness increases. The complete case analysis has generally remained close to the original model estimates. Generally, however, the estimates for the intercept and triceps are positive, as in the original model, and the thigh and midarm estimates are negative. The exceptions are the mean imputation estimates, which are considered separately in §4.7 and the regression imputations for the 50% MCAR data. The signs of the regression imputation estimates, both single and multiple, are affected when the missingness is 50%, but it must be considered just how high this percentage of missingness is. It has already been discussed how removing just a few significant points can greatly affect the performance of the regression imputation method. Therefore, any conclusions drawn from the data sets with missing values, regarding which variables are positively or

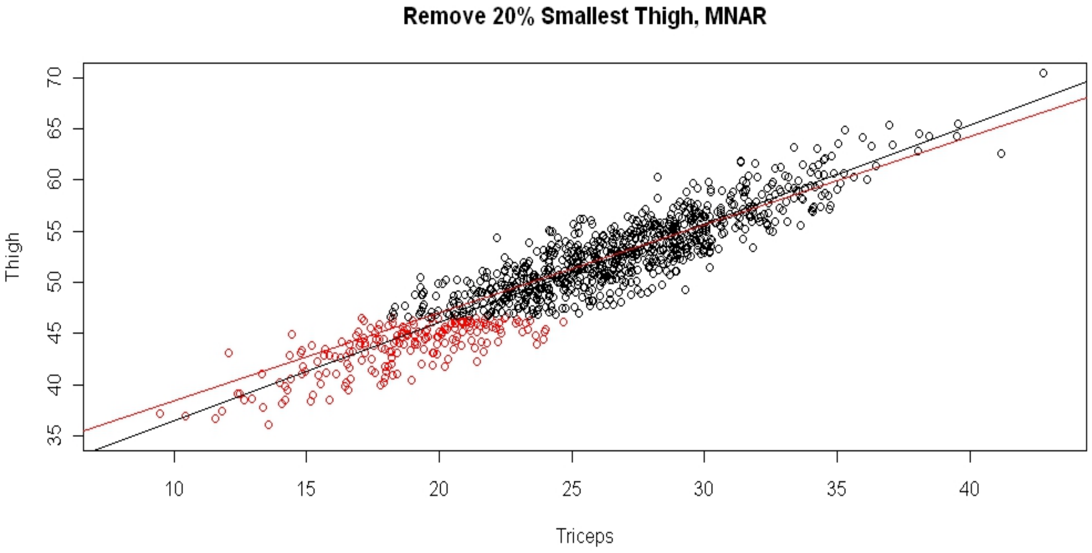


Figure 4.23: Body fat data, 1000 subjects, 20% *MNAR*

negatively associated with the outcome, are generally correct.

Similar simulation studies have also been carried out by Little (1979), Chen *et al.* (2007) and Rubin (1976). These studies used different imputation methods but their results showed how, generally, it is beneficial to use a form of imputation rather than complete case analysis. The final results are usually less biased and the methods allow for less loss of information. Some of the results seen in this study have seemed unusual, for example, how well the complete case analysis performed when the data were MAR and MNAR. One explanation may be that in this study, there were no missing values in the outcome. Missingness was generated in the outcome in the similar simulations studies mentioned, and well explained in Little (1979). If missingness was also generated in the outcome, then different results may have been seen.

4.7 Mean Imputation

The mean imputation estimates for all the missingness mechanisms were far from those for the original model. This section aims to further investigate the reasons for this. It has already been discussed that the confidence intervals for the mean imputation methods are smaller. The confidence intervals are likely to be smaller as the imputed values are the mean and hence the variance of the variables is reduced. The confidence intervals and accuracy of the estimates will both depend upon which particular values are missing. If there are a small percentage of missing values and they are close to the mean, then mean imputation will cope reasonably well and the confidence intervals will not change much from those for the original model. However, it is more likely that the missing values will not all be close to the mean, nor the percentage of missingness will be small enough.

Figure 4.25 shows data which are MCAR for different percentages of missingness. The smallest percentage considered previously has been 10% and the method has not coped well under any of the missingness mechanisms. MCAR has been selected since this is the mechanism under which most methods should cope best. Figure 4.25 considers from 0.1% missingness up to the 10% used previously. The percentages in between are; 0.5%, 1% and 5%. It can be seen from the plots that as the percentage of missingness increases, the estimates drift further away from the original model estimate, shown in black. This suggests that, for data MCAR, the mean imputation estimates may be acceptable for very small amounts of missing data. The numbers of missing values from these plots are shown in table 4.6. Note these number are out of 4000; 1000 subjects × 4 variables. Mean imputation therefore only appears to cope, for this data set, when the missingness is around 0.1%, which corresponds to 3 missing values across the variables. For the other percentages shown, the estimates and confidence intervals are entirely outside the confidence intervals for the original model.

Percentage of Missingness, MCAR	Number of Missing Values
0.1	3
0.5	20
1.0	43
5.0	208
10.0	410

Table 4.6: Numbers of missing values, MCAR

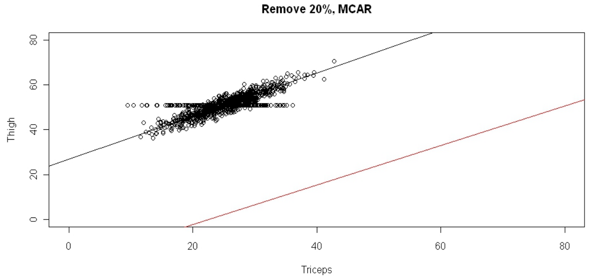


Figure 4.24: Body fat data, 1000 subjects, mean imputation

The performance of the mean imputation method will depend heavily on which particular values are missing. Figure 4.19 can be used to help explain this. The red line shows the regression line through all the data points for the two variables; thigh and triceps. Note there are no missing values. Now, in figure 4.24 there are some data values which are MCAR in thigh, which have been replaced with the mean value for the thigh variable. This would result in the data values becoming more focused around the mean of thigh, that is, around the centre of the graph. This therefore considerably alters the regression line through the data, shown in red, hence amending the linear regression model for the output. Mean imputation, would in theory, be most useful if the missing values were those close to the mean. However, this is likely to be rare in practice, and certainly would not occur if the data values are MCAR, unless the variance of the variables is very small indeed, that is, all values are close to the mean value. If the values were MAR or MNAR, it is possible that just those values close to the mean could be missing. However, there still needs to be a reasoning for these values to be missing. If the values are MNAR, often the largest or smallest values will be missing, which would again greatly affect the regression line through the data points once these missing values had been replaced with the mean value for that variable. Replacing even just one or two large or small values by the mean could change the regression line enough to provide an overall result very different from that for the original data set.

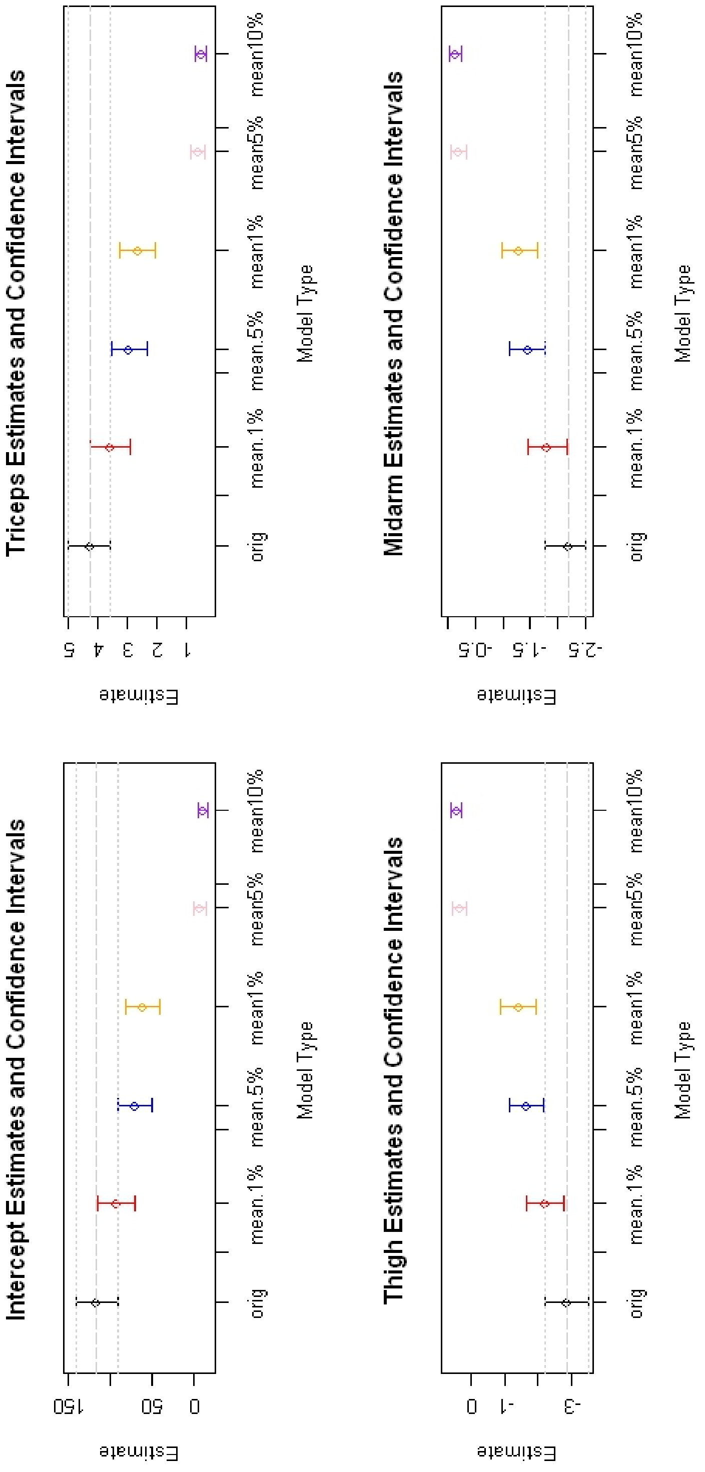


Figure 4.25: Estimates and confidence intervals for mean imputation: Differing % MCAR per variable

4.8 The ‘mice’ Package: Body Fat Data

Several decisions had to be made in order to carry out the calculations on the body fat data set using the ‘mice’ package. These are specified below.

- The data were assumed to be MAR. Comparisons were also drawn between those known to be MAR and MNAR.
- All columns were given the same method of imputation, so comparisons between methods could be made. All the variables are continuous, hence only certain options are available. The methods used were,
  - MEAN: for mean imputations,
  - NORM.NOB: for regression imputation,
  - NORM: for stochastic regression imputation,
  - NORM: for the multiple imputations.
- The predictor matrix was generated using correlations, since there was no prior information available or reasoning to suggest another matrix would be more suitable.
- The imputations were carried out from left to right, using the default ordering.
- The number of imputations was set to one for the mean, regression and stochastic regression methods and increased to 10 then 100 for the multiple imputations.
- Any subjects who had their outcome variable imputed, that is, body fat, were removed from the analysis, as required, so the results were as accurate as possible.
- A number of different percentages of missingness were tried in R to see whether the results were affected.
  - Originally $10\% - 50\%$ missingness was generated across all the variables, before $10\% - 50\%$ missingness was generated in each of the variables.
  - Missingness was also tested at $5\% - 25\%$, in 5% increments.
  - Missingness was also tested at $20\% - 80\%$, in 20% increments.
- Transformations of the data set were investigated to aid understanding of the results observed.

Chapter 5

Missing Data: AAA Surgery Data

This chapter aims to carry out multiple imputation on the data set introduced in chapter 2. First, summaries of the data set will be generated to investigate the missingness. Next, the 'mice' package in *R*, see §3.8 for more details, will be used to carry out the multiple imputation steps.

5.1 Missingness Summaries

This section will produce summaries of the missingness to try to investigate the missingness mechanism and any obvious patterns. Both tables and plots will be used to summarise the missingness before the imputation steps are carried out.

The first table tried was the one generated using the `md.pattern` command in the 'mice' package in *R*. However, since the data set is so large, the full table generated was 1047 by 24, which could not be displayed in *R*, nor would it be useful for analysis. Since the information cannot easily be summarised from the output, especially as the totals for each row are displayed in the margin and not in the matrix itself, it is difficult to use this command for this data set. Possible alternatives may be to use either a subset of the variables which may be of particular interest, or to use a random sample of the subjects to carry out this stage of the analysis on. However, for this study, the alternative summary method of investigating each pair of variables will be used. These are shown in appendix A; there is one table for the number of subjects with both variables present, another for both variables missing and a third for the first variable present and the second missing. Note that the fourth option where the first variable is missing and the second is recorded, can be calculated from the information provided in appendix A. It can be seen from these tables that there appears to be no obvious patterns in the numbers. This therefore suggests that the values may be MAR.

Next, plots will be used to investigate any missingness patterns within the data. Figure 5.1 shows a margin plot for the age and haemoglobin variables. The blue points show the subjects who have a recorded value for both age and haemoglobin. The red dots show the subjects who have a recorded value for one variable but not the other. The subject is placed along the axis of the recorded variable. The orange dot shows the subjects who do not have a recorded value for either variable. The number of subjects without a recorded value for either variable is also listed in the bottom left corner of the plot. It can be seen that there are 66 subjects who are missing both age and haemoglobin values. The red box plots along the axes show the distribution of the missing values, whilst the blue box plots show the distribution of the recorded values. If the data are MCAR, these two box plots are expected to be the same. It can be seen from figure 5.1 that the box plots are very similar, which does not dismiss the possibility of the data being MAR.

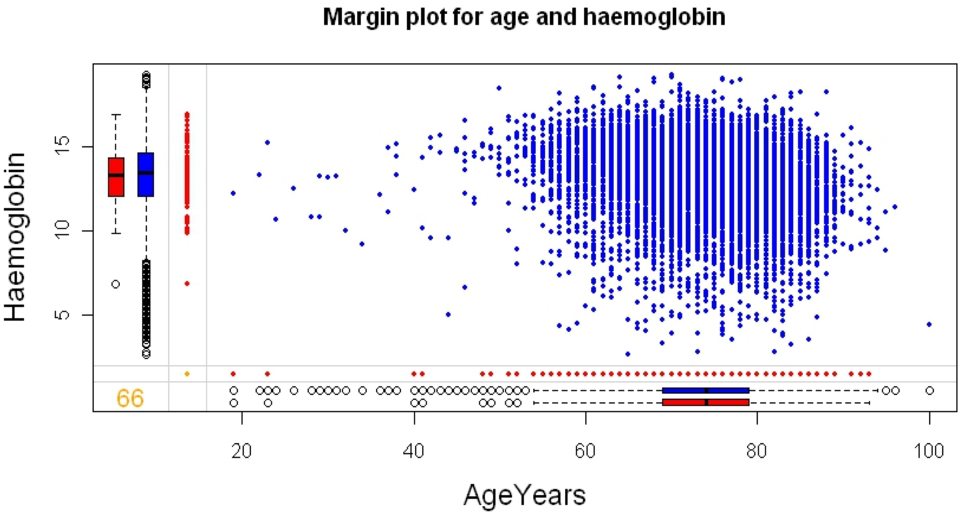


Figure 5.1: Margin plots of the AAA variables: Age and haemoglobin

Each of the pairs of variables were checked to see whether the assumption of MAR could be dismissed or if any patterns emerged. The majority of the plots were similar to figure 5.1 where the red and blue box plots looked very similar and there are no obvious patterns in the scatter plots. However, there were some box plots which differed a little, mainly when haemoglobin was one of the variables in the pair. The most extreme case is shown in figure 5.2 where blood pressure is plotted against pulse. It can be seen that the box plots are not as they are in figure 5.1. However, since this is the most extreme case and all the other box plots seem satisfactory, the data will be assumed to be MAR, which is not an unrealistic assumption.

An alternative plot which can be used to investigate any patterns in the missingness of the data is a parallel box plot. This plot generates two box plots for each variable; one for the missing values and one for the recorded values, similar to the margin plots above. It uses a basis variable which is a variable of interest, against which to compare the other variables. The first box plot in white shows the standard box plot for the variable of interest. There are then blue and red box plots for each of the remaining variables against the variable of interest. Figure 5.3 shows an example, which uses age as the variable of interest. The blue boxes represent the observed values while the red boxes represent the missing values. These are show in pairs and from left to right correspond to; Haemoglobin, WhiteCellCount, Urea, Sodium, Potassium, LowestSystolicBP_Intraoperative and HighestPulse_Intraoperative. The white box is the standard box plot for the age variable. Since the box plots are all very similar, this suggests that the data are not MNAR.

All the different parallel box plots were produced, with each of the different continuous variables used as the variable of interest. Each of the parallel box plots were similar to that in figure 5.3 where the box plots are all very similar. There were slight differences for the haemoglobin and blood pressure plots, but these were only minor differences with the red plots slightly higher/lower than the blue plots. The most noticeable differences were in the pulse parallel plot, shown in figure 5.4. It can be seen that the red plots are all higher than the blue plots, for each of the variables. The variables are again in the order of; Haemoglobin, WhiteCellCount,

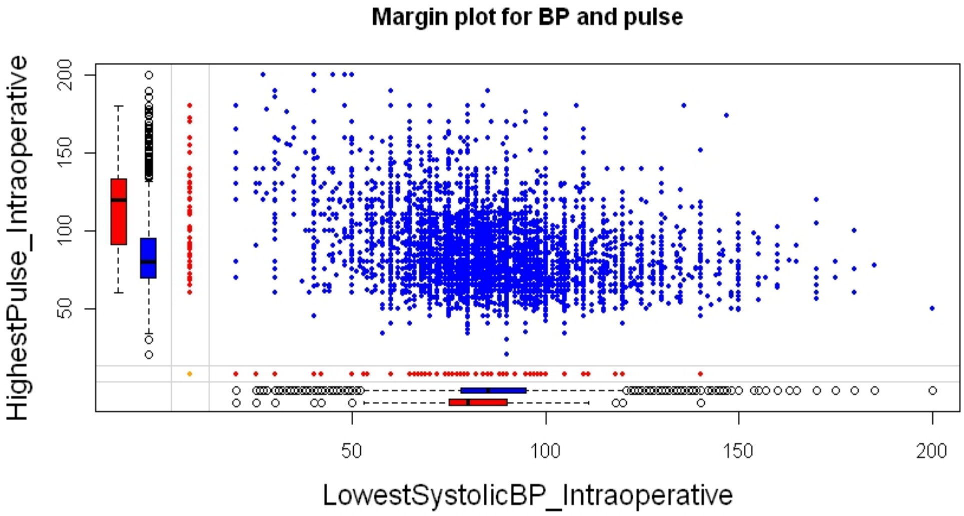


Figure 5.2: Margin plots of the AAA variables: BP and pulse

Urea, Sodium, Potassium, LowestSystolicBP_Intraoperative and HighestPulse_Intraoperative. It can be seen, as with the margin plot in figure 5.2, that the biggest difference is between the blood pressure and pulse variables. However, since all the other plots were satisfactory and this plot is not vastly different, the data will be assumed to be MAR.

Recall also that additional information regarding missingness in the AAA surgery data set was also provided in chapter 2, including the percentages of missingness in the different variables.

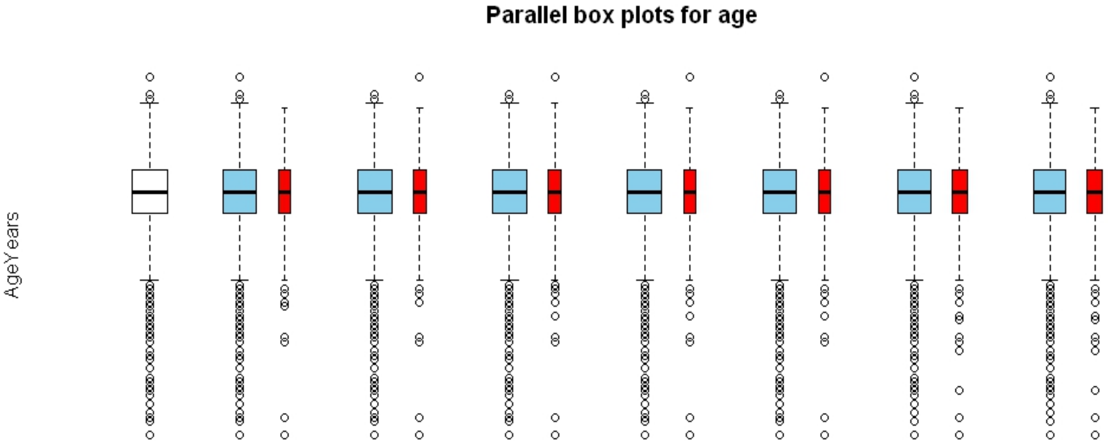


Figure 5.3: Parallel box plot of the AAA variables: Age

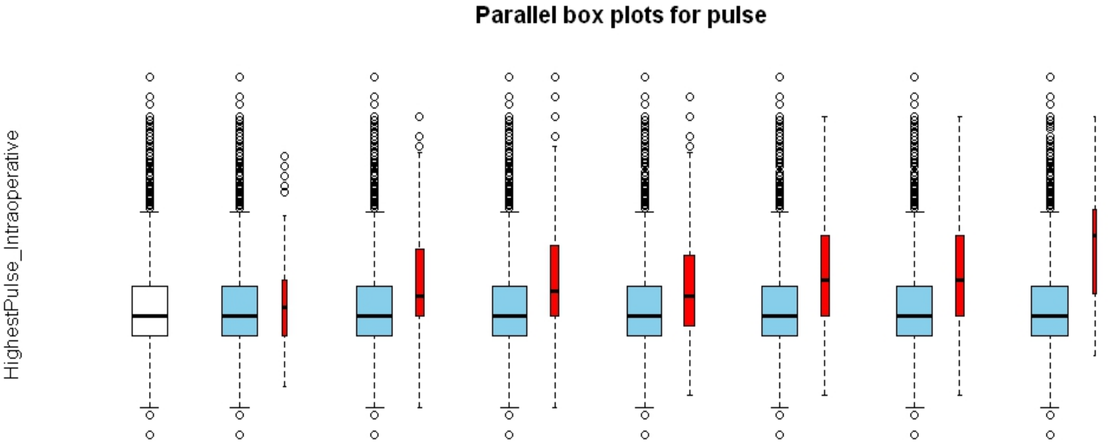


Figure 5.4: Parallel box plots of the AAA variables: Pulse

5.2 Multiple Imputation Method

There are seven steps which should be considered before carrying out the imputation process, as mentioned in §3.8. These are briefly listed below, along with the decisions made for each,

1. From the plots in §5.1, it can be seen that there is not sufficient evidence to conclude that the missing values are not MAR, therefore the assumption of MAR can be assumed.

2. The following methods were chosen for each column of the data, depending on the nature of the data. Generally, logistic regression was chosen for the binary variables, polytomous regression was chosen for the categorical variables and predictive mean matching was chosen for the continuous variables. Predictive mean matching is simply hot deck imputation or pattern matching as previous described in chapter 3, and was chosen so no implausible values would be imputed. Logistic regression and polytomous regression use the same method as the regression imputation methods mentioned previously in chapter 3, but simply with the form of regression suitable for the variable being imputed. Hospital ID was excluded since it had no missing values. Table 5.1 lists the methods used.

Variable	Method
Gender	”logreg”
admissionMode	”polyreg”
Diabetes	”logreg”
CurrentSmoker	”logreg”
RenalDialysis	”logreg”
RenalTransplant	”logreg”
PreviousAorticSurgeryStent	”logreg”
AAASurgery	”polyreg”
Haemorrhage	”logreg”
Stroke	”polyreg”
MyocaridalInfarct	”logreg”
CardiacFailure	”logreg”
Hypotension	”logreg”
dischargeStatus	”logreg”
AgeYears	”pmm”
Haemoglobin	”pmm”
WhiteCellCount	”pmm”
Urea	”pmm”
Sodium	”pmm”
Potassium	”pmm”
LowestSystolicBP_Intraoperative	”pmm”
HighestPulse_Intraoperative	”pmm”
hospitalID	””

Table 5.1: Imputation methods used for the AAA surgery data

3. The predictor matrix could not be selected using the `quickpred` function in the ‘mice’ package in R. Recall that the `quickpred` function generates the prediction matrix by considering the correlations of the variables, and there are settings which can be edited, such as the threshold for the correlations, for two variables to be considered as predictors.

This cannot be used for the AAA surgery data as there are many categorical and binary variables, where the correlations would not be sensible. Therefore, prior knowledge must be used to create a predictor matrix which is plausible. The predictor matrix chosen was developed using the help of clinicians and is shown below in table 5.2.

Gen	aM	D/b	Smk	D/l	T/pl	Pre	AAA	H/h	Str	Myo	Car	Hyp	Sta	Age	H/g	WC	Ure	Sod	Pot	BP	Pul	ID
0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	1	1	0
0	0	1	0	0	1	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0
1	0	0	0	0	0	0	1	0	1	1	1	1	1	1	0	0	0	1	0	1	1	0
1	0	0	1	0	0	0	1	1	0	1	1	1	1	1	0	1	0	1	0	1	1	0
1	0	1	1	0	0	0	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	0
1	0	1	1	0	0	0	1	1	1	1	0	1	1	1	0	0	0	1	0	1	1	0
0	0	0	0	0	0	0	1	1	1	1	1	0	1	1	0	1	0	1	0	1	1	0
1	0	1	1	0	0	0	0	1	1	1	1	1	0	1	0	0	0	1	0	1	1	0
1	1	0	1	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	1	1	0	0	1	0	1	0	0	1	0	0	0	1	0
0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0
0	0	0	0	1	1	0	1	1	1	0	0	0	1	1	0	0	0	0	1	1	1	0
0	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	1	0
0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0
0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5.2: Table showing the predictor matrix used for the AAA surgery data

4. Passive imputation is not necessary in this example, as there are not variables which are functions of the other variables. For instance, BMI would be a function of height and weight.

5. The visiting scheme was set to the default, that is, the order in which the variables were imputed was simply from left to right.

6. The number of iterations was set to the default of 5. More imputations were desired but were limited by the computation time since the data set was so large and there were many values to impute.

7. The number of imputed data sets was set to 10. Again, more were desired but were limited by the computation time.

Once all the steps had been considered, the `mice` command in R could be run. Once finished, the subjects with imputed values for their outcome, dischargeStatus, needed to be removed. The subjects with no recorded outcome need to be involved in the imputation stage, to assist with the imputation of other variables, but must be removed once the imputation stage is complete so as not to bias the results. There were 443 such subjects. Once removed, the remaining subjects could have their imputed values checked to see whether they were plausible. Figure 5.5 shows examples of stripplots which can be used to check the distributions of the recorded and imputed data sets. The plot shows the ten imputed data sets, 1-10, and the data before imputation, 0. It can be seen that, for both the example variables, the observed data (blue) and the imputed data (red) appear to be similar. Under MCAR, the univariate distributions are expected to be identical, but under MAR they may differ. Plots were generated for all the variables with imputed values and all the imputed values seemed reasonable, that is, the imputed values were similar to the recorded values. Next, regression models for each of the 10 imputed data sets could be formed. The regression models could then be pooled using Rubin's rules and summarised to form one overall regression model with estimates and confidence intervals.

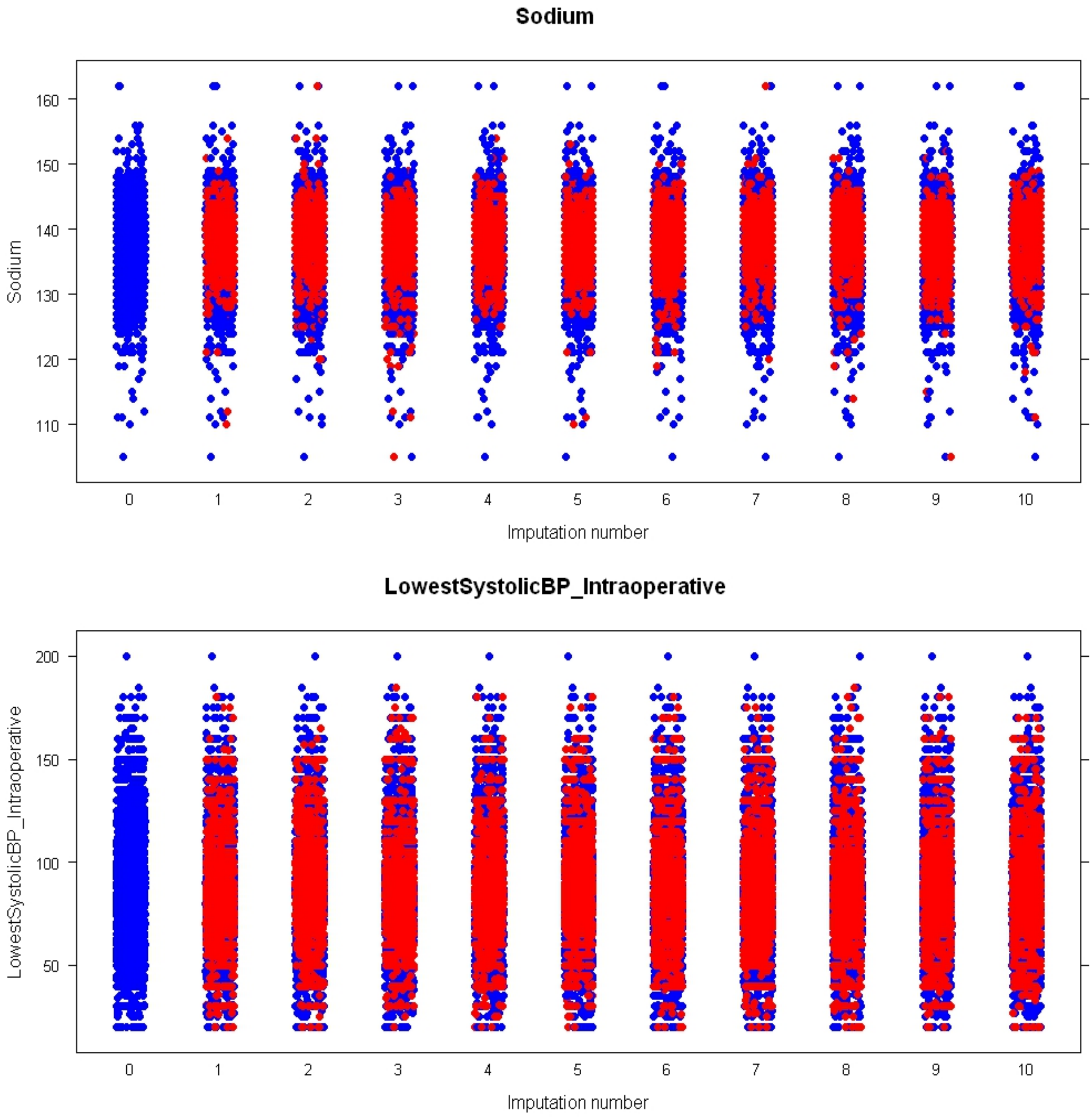


Figure 5.5: Stripplots used to check the distributions of the recorded and imputed data sets

5.3 Results

The variables used in the regression model were those suggested by Tang *et al* (2007). Tang *et al* (2007) carried out a study which focused on developing a model using the same variables and the same outcome. Their study used data collected in the UK between 2002 and 2004. However, Tang *et al* (2007) used only the 2718 complete cases to form the final regression model with the selected variables. These variables are listed below,

- Gender,
- admissionMode,
- AgeYears,
- Urea,
- Sodium,
- Potassium,
- Haemoglobin and
- WhiteCellCount.

These variables were used to form the regression models from the imputed data sets, with dischargeStatus, that is whether or not the subject survived the surgery, as the outcome. The regression models were then combined using Rubin's rules, see §3.5, to form one overall regression model. This is shown in table 5.3. Table 5.3 shows the estimates for each variable along with their standard error and 95% confidence interval.

	est	se	lo 95	hi 95
(Intercept)	-16.42	228.78	-464.88	432.04
GenderM	-0.32	0.14	-0.58	-0.05
admissionMode1	12.00	228.77	-436.44	460.44
admissionMode2	13.37	228.77	-435.07	461.80
AgeYears	0.04	0.01	0.02	0.05
Urea	0.02	0.01	0.01	0.03
Sodium	0.00	0.02	-0.03	0.03
Potassium	0.03	0.04	-0.05	0.11
Haemoglobin	-0.19	0.03	-0.25	-0.14
WhiteCellCount	0.06	0.01	0.03	0.09

Table 5.3: Table showing the results from the overall model for the AAA surgery data

The estimates in table 5.3, considering they are based on a different data set, are close to those given by Tang *et al* (2007), with most estimates inside or slightly outside the confidence intervals. The results from Tang *et al* (2007) are shown in table 5.4.

The estimates from this model can be compared to the ones generated using just the complete cases for the same subset of variables. These are shown in table 5.5. It can be seen that the estimates are generally similar, but the confidence intervals for the complete case analysis are

	Estimate
(Intercept)	-2.257
GenderM	0.1511
admissionMode2/3	0.9940
AgeYears	0.05923
Urea	0.001401
Sodium	-0.01303
Potassium	-0.03585
Haemoglobin	-0.2278
WhiteCellCount	0.02059

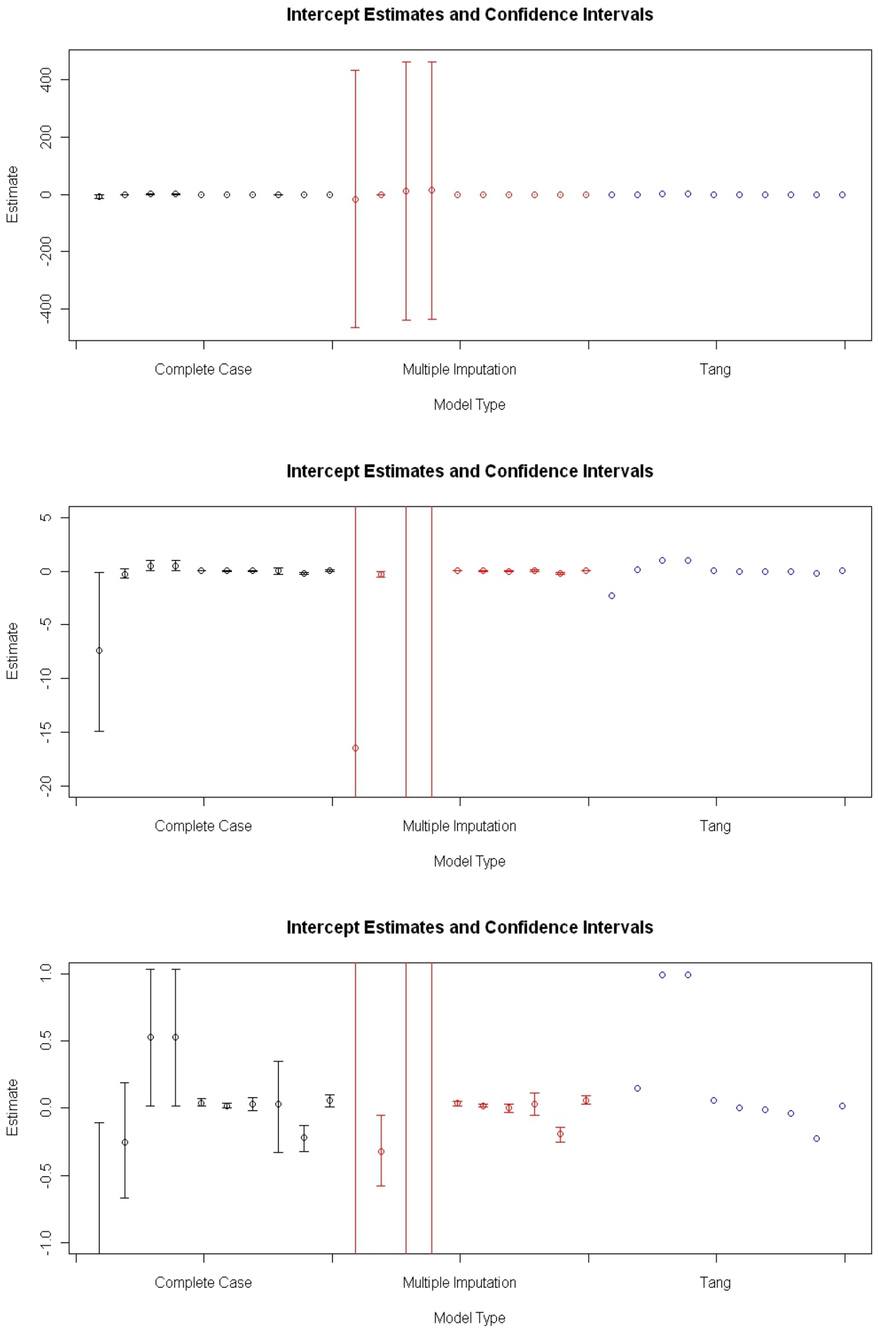
Table 5.4: Table showing the results from Tang et al (2007)

wider. This is as expected, since the complete case analysis has less information, in fact only 1488 subjects compared to the 14010 subjects used for the multiple imputation analysis. There is not a great difference between the width of the confidence intervals as the multiple imputation analysis allows for the added uncertainty in imputing the missing values hence has fairly wide confidence intervals. Notice that the signs of the estimates remain the same for the two models. Therefore, the interpretation of each variable in the model with respect to the outcome of dischargeStatus, remains the same. Thats is, whether the variable being considered increases or decreases the chances of a patient surviving the surgery, remains the same.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.3882	3.7831	-1.95	0.0508
GenderM	-0.2513	0.2190	-1.15	0.2511
admissionMode2	0.5333	0.2566	2.08	0.0377
AgeYears	0.0447	0.0136	3.30	0.0010
Urea	0.0179	0.0099	1.81	0.0703
Sodium	0.0299	0.0250	1.20	0.2319
Potassium	0.0276	0.1759	0.16	0.8753
Haemoglobin	-0.2204	0.0481	-4.58	0.0000
WhiteCellCount	0.0561	0.0241	2.33	0.0198

Table 5.5: Table showing model estimates for the complete cases analysis for the AAA surgery data

Figure 5.6 shows these three models compared graphically. The different colours show the different models, whilst the different lines within each colour show the estimates in the order given in the tables. The first graph shows all the variables, whereas the second and third graphs zoom further in. It can be seen how they all produce similar results.



Intercept Estimates and Confidence Intervals

Intercept Estimates and Confidence Intervals

Intercept Estimates and Confidence Intervals

Figure 5.6: Graphs of all three model types

5.4 Interpretation

The model in table 5.3 shows the selected variables and their estimated effect on mortality. It can be seen that there are several variables which increase the chances of mortality, these are;

- admissionMode: elective or unplanned,
- being older,
- higher urea levels,
- higher potassium levels and
- higher white cell count.

There are also variables which are estimated to decrease the chances of mortality, which are;

- being male and
- a higher haemoglobin level.

However, some of the confidence intervals, such as those for admissionMode, Sodium and Potassium, include both positive and negative values, showing these variables may not be significant to mortality. Note the estimate for Sodium is also listed as 0 (2dp). Recall that, due to computation restrictions, fewer iterations and imputations than desired were carried out. Increasing these may have had an effect on the estimated values shown in table 5.3.

Regression models were used as the final stage of analysis, but note these are based on linear associations and not more complicated relationships. There are more sophisticated variable selection methods and models which could have been chosen, but these areas were not the focus for this study. However, regression models are still a useful form of analysis for future subjects and could be of use to others considering AAA surgery.

Chapter 6

Conclusions

This study has investigated several different missingness mechanisms, how they occur and how summary plots can be used to try to predict the missingness mechanism. It has also considered many different solutions for coping with missing data, from simple complete case analysis to multiple imputation as shown in chapter 5. Examples have been given throughout for the different mechanisms and methods and some of these methods have been tested to see how they perform under the different missingness mechanisms. Reasoning has also been sought for the behaviour seen during these comparisons. Lastly, all the theory and practice was used to tackle a large and complicated data set which contained missing values. Different methods were combined to produce a more suitable method for the analysis with missing data.

It has been noted that missing values occur for a number of different reasons and these help to determine the missingness mechanism. The missingness mechanism is important when choosing a method to deal with the missing values, although it cannot be easily determined from the data alone. The different methods have been shown to be useful in certain circumstances, with some more desirable due to their simplicity, others due to their ability to cope with complicated data; such as binary, categorical and numerical data all in the same data set. Whenever a data set has missing values, it should be carefully considered why the data values are missing and if they are MAR or not. The method then selected should be carefully chosen, with consideration taken for the missingness mechanism, the nature of the data and the required outcome from the analysis.

A range of different sources have been used throughout this study to obtain the information required. Where possible, the original journal articles have been used for information rather than summaries in books or on websites. Unfortunately, many articles, books and websites referred back to Rubin (1987) but this book could not be sourced. If it had been available, more information regarding Rubin's rules and associated work could have been provided.

The data sets in this study have been based on medical data, but missing data can occur in other areas too. The methods described can be used not only in medical research, but wherever missing data occurs, giving more suitable methods and usually an analysis with less bias and less loss of information.

Bibliography

Cattle, B.A., P.D. Baxter, D.C. Greenwood, C.P. Gale & R.M. West, (2011). Multiple imputation for completion of a national clinical audit dataset. *Statistics in Medicine*. To appear.

Chen, Q., D. Zeng & J.G. Ibrahim, (2007). Sieve maximum likelihood estimation for regression models with covariates missing at random. *Journal of the American Statistical Association*, **102**, pp 1309-1317.

Economic Commission for Europe of the United Nations (UNECE), (2000). Glossary of terms on statistical data editing. *Conference of European Statisticians Methodological Material, Geneva*.

Kline, R.B. (1998). Principles and Practices of Structural Equation Modeling. New York:Guilford.

Little, R.J.A. (1979). Maximum likelihood inference for multiple regression with missing values: a simulation study. *Journal Royal Statistics Society, B*, **41**, pp 76-87.

Little, R.J.A. & D.B. Rubin, (2002). Statistical Analysis with Missing Data, second edition. New Jersey: Wiley.

Meng, X.L., (1994). Multiple imputation inferences with uncongenial sources of input. *Statistical Science*, **9**, pp 538-558.

Neter, J., M.H. Kutner, C.J. Nachtsheim & W. Wasserman. Applied Linear Statistical Models (Fourth Edition). Boston:McGraw-Hill.

Rubin, D. (1976). Inference and missing data. *Biometrika*, **63**, pp 581-592.

Rubin, D. (1987). Multiple Imputation for Non-Response in Surveys. Wiley:New York.

Statistics Canada (2003). Statistics Canada Quality Guidelines, 4th edition, Canada.

Sterne J.A.C., R. White, J.B. Carlin, M. Spratt, P. Royston, M.G. Kenwood, A.M. Wood & J.R. Carpenter (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, **338**, pp 23932397.

Tang T., S.R. Walsh, D.R. Prytherch, T. Lees, K. Varty & J.R. Boyle in Association with the Audit and Research Committee of the Vascular Society of Great Britain and Ireland (2007). VBHOM, a data economic model for predicting the outcome after open abdominal aortic aneurysm surgery. *British Journal of Surgery*, **94**, pp 717-721.

Van Buuren S. and K. Groothuis-Oudshoorn (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, forthcoming. Available from: http://www.stefvanbuuren.nl/publications/MICE in R - Draft.pdf.

Wothke, W. (1998). Longitudinal and Multi-Group Modelling with Missing Data. In T. D. Little, K. U. Schnabel, and J. Baumert [Eds.]. Modelling Longitudinal and Multiple Group Data: Practical Issues, Applied Approaches and Specific Examples. Mahwah, NJ: Lawrence Erlbaum Publishers.

Appendix A

AAA Data: Tables of Variable Pairs, Recorded/Missing Values

There follows three tables, which contain the numbers of subjects who have,

- Both the variables in the pair recorded,
- One variable in the pair missing and the other recorded,
- Both variables in the pair missing.

The results can then be used to investigate any patterns which may occur in the data set. For example, there may be a large amount of subjects who have a certain pair of variables missing. These patterns can lead to conclusions as to whether or not the data appear to be MAR.

	Gend	adMode	Diab	Smk	Dial	T/plant	Prev	AAA	Haemor	Str	Myo	Card	Hypo	Stat	Age	Haemog	WCC	Urea	Sod	Pot	LowBP	HighPulse	hospID
Gend	14009	10849	13187	10999	12093	11842	12413	10328	5107	4587	5239	4841	4850	13566	13840	12550	12291	12031	12487	12434	11661	11582	14009
adMode	10849	10849	10283	8721	9515	9283	9600	7719	3639	3322	3737	3479	3495	10547	10737	9893	9757	9567	9918	9891	9081	8997	10849
Diabetes	13187	10283	13188	10939	12025	11776	12161	10109	5022	4520	5157	4769	4778	12789	13063	12219	11977	11722	12166	12116	11541	11475	13188
Smoker	10999	8721	10939	10999	10659	10647	10375	8082	4009	3853	4019	3989	3982	10622	10908	10207	9994	9801	10158	10120	9493	9434	10999
Dialysis	12093	9515	12025	10659	12093	11809	11284	9116	4494	4271	4542	4402	4361	11702	11981	11250	11025	10810	11185	11137	10596	10617	12093
Transplant	11842	9283	11776	10647	11809	11842	11271	8868	4257	4040	4307	4167	4125	11460	11733	11007	10789	10566	10943	10899	10349	10368	11842
Previous	12413	9600	12161	10375	11284	11271	12414	9694	4641	4177	4775	4421	4448	12043	12298	11492	11295	11005	11442	11392	10985	11012	12414
AAASurgery	10328	7719	10109	8082	9116	8868	9694	10329	4104	3621	4226	3833	3846	10039	10215	9605	9405	9141	9528	9485	9303	9218	10329
Haemorrhage	5107	3639	5022	4009	4494	4257	4641	4104	5108	4526	4979	4650	4662	5068	5067	4599	4476	4459	4568	4543	4453	4481	5108
Stroke	4587	3322	4520	3853	4271	4040	4177	3621	4526	4588	4549	4557	4496	4553	4551	4169	4040	4034	4134	4108	3978	4003	4588
Myo	5239	3737	5157	4019	4542	4307	4775	4226	4979	4549	5240	4790	4797	5203	5193	4731	4608	4592	4700	4669	4586	4608	5240
Cardiac	4841	3479	4769	3989	4402	4167	4421	3833	4650	4557	4790	4842	4744	4807	4796	4395	4265	4250	4357	4327	4208	4233	4842
Hypo	4850	3495	4778	3982	4361	4125	4448	3846	4662	4496	4797	4744	4851	4814	4807	4382	4273	4259	4368	4338	4229	4252	4851
Status	13566	10547	12789	10622	11702	11460	12043	10039	5068	4553	5203	4807	4814	13567	13415	12196	11953	11703	12146	12094	11326	11249	13567
Age	13840	10737	13063	10908	11981	11733	12298	10215	5067	4551	5193	4796	4807	13415	13841	12447	12175	11929	12386	12335	11542	11462	13841
Haemoglobin	12550	9893	12219	10207	11250	11007	11492	9605	4599	4169	4731	4395	4382	12196	12447	12550	12150	11869	12325	12273	11111	11035	12550
WhiteCC	12291	9757	11977	9994	11025	10789	11295	9405	4476	4040	4608	4265	4273	11953	12175	12150	12291	11700	12117	12060	10905	10848	12291
Urea	12031	9567	11722	9801	10810	10566	11005	9141	4459	4034	4592	4250	4259	11703	11929	11869	11700	12032	11971	11916	10646	10561	12032
Sodium	12487	9918	12166	10158	11185	10943	11442	9528	4568	4134	4700	4357	4368	12146	12386	12325	12117	11971	12488	12402	11065	10983	12488
Potassium	12434	9891	12116	10120	11137	10899	11392	9485	4543	4108	4669	4327	4338	12094	12335	12273	12060	11916	12402	12435	11022	10940	12435
LowBP	11661	9081	11541	9493	10596	10349	10985	9303	4453	3978	4586	4208	4229	11326	11542	11111	10905	10646	11065	11022	11662	11467	11662
HighPulse	11582	8997	11475	9434	10617	10368	11012	9218	4481	4003	4608	4233	4252	11249	11462	11035	10848	10561	10983	10940	11467	11583	11583
hospitalID	14009	10849	13188	10999	12093	11842	12414	10329	5108	4588	5240	4842	4851	13567	13841	12550	12291	12032	12488	12435	11662	11583	14010

Table A.1: Table showing pairs of recorded values: AAA data

	Gend	adMode	Diab	Smk	Dial	T/plant	Prev	AAA	Haemor	Str	Myo	Card	Hypo	Stat	Age	Haemog	WCC	Urea	Sod	Pot	LowBP	HighPulse	hospID
Gend	0	3160	822	3010	1916	2167	1596	3681	8902	9422	8770	9168	9159	443	169	1459	1718	1978	1522	1575	2348	2427	0
adMode	0	0	566	2128	1334	1566	1249	3130	7210	7527	7112	7370	7354	302	112	956	1092	1282	931	958	1768	1852	0
Diabetes	1	2905	0	2249	1163	1412	1027	3079	8166	8668	8031	8419	8410	399	125	969	1211	1466	1022	1072	1647	1713	0
Smoker	0	2278	60	0	340	352	624	2917	6990	7146	6980	7010	7017	377	91	792	1005	1198	841	879	1506	1565	0
Dialysis	0	2578	68	1434	0	284	809	2977	7599	7822	7551	7691	7732	391	112	843	1068	1283	908	956	1497	1476	0
Transplant	0	2559	66	1195	33	0	571	2974	7585	7802	7535	7675	7717	382	109	835	1053	1276	899	943	1493	1474	0
Previous	1	2814	253	2039	1130	1143	0	2720	7773	8237	7639	7993	7966	371	116	922	1119	1409	972	1022	1429	1402	0
AAASurgery	1	2610	220	2247	1213	1461	635	0	6225	6708	6103	6496	6483	290	114	724	924	1188	801	844	1026	1111	0
Haemorrhage	1	1469	86	1099	614	851	467	1004	0	582	129	458	446	40	41	509	632	649	540	565	655	627	0
Stroke	1	1266	68	735	317	548	411	967	62	0	39	31	92	35	37	419	548	554	454	480	610	585	0
Myo	1	1503	83	1221	698	933	465	1014	261	691	0	450	443	37	47	509	632	648	540	571	654	632	0
Cardiac	1	1363	73	853	440	675	421	1009	192	285	52	0	98	35	46	447	577	592	485	515	634	609	0
Hypo	1	1356	73	869	490	726	403	1005	189	355	54	107	0	37	44	469	578	592	483	513	622	599	0
Status	1	3020	778	2945	1865	2107	1524	3528	8499	9014	8364	8760	8753	0	152	1371	1614	1864	1421	1473	2241	2318	0
Age	1	3104	778	2933	1860	2108	1543	3626	8774	9290	8648	9045	9034	426	0	1394	1666	1912	1455	1506	2299	2379	0
Haemoglobin	0	2657	331	2343	1300	1543	1058	2945	7951	8381	7819	8155	8168	354	103	0	400	681	225	277	1439	1515	0
WhiteCC	0	2534	314	2297	1266	1502	996	2886	7815	8251	7683	8026	8018	338	116	141	0	591	174	231	1386	1443	0
Urea	1	2465	310	2231	1222	1466	1027	2891	7573	7998	7440	7782	7773	329	103	163	332	0	61	116	1386	1471	0
Sodium	1	2570	322	2330	1303	1545	1046	2960	7920	8354	7788	8131	8120	342	102	163	371	517	0	86	1423	1505	0
Potassium	1	2544	319	2315	1298	1536	1043	2950	7892	8327	7766	8108	8097	341	100	162	375	519	33	0	1413	1495	0
LowBP	1	2581	121	2169	1066	1313	677	2359	7209	7684	7076	7454	7433	336	120	551	757	1016	597	640	0	195	0
HighPulse	1	2586	108	2149	966	1215	571	2365	7102	7580	6975	7350	7331	334	121	548	735	1022	600	643	116	0	0
hospitalID	1	3161	822	3011	1917	2168	1596	3681	8902	9422	8770	9168	9159	443	169	1460	1719	1978	1522	1575	2348	2427	0

Table A.2: Table showing the 1st variable recorded, 2nd variable missing: AAA data

	Gend	adMode	Diab	Smk	Dial	T/plant	Prev	AAA	Haemor	Str	Myo	Card	Hypo	Stat	Age	Haemog	WCC	Urea	Sod	Pot	LowBP	HighPulse	hospID
Gend	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
adMode	1	3161	256	883	583	602	347	551	1692	1895	1658	1798	1805	141	57	504	627	696	591	617	580	575	0
Diabetes	0	256	822	762	754	756	569	602	736	754	739	749	749	44	44	491	508	512	500	503	701	714	0
Smoker	1	883	762	3011	1577	1816	972	764	1912	2276	1790	2158	2142	66	78	668	714	780	681	696	842	862	0
Dialysis	1	583	754	1577	1917	1884	787	704	1303	1600	1219	1477	1427	52	57	617	651	695	614	619	851	951	0
Transplant	1	602	756	1816	1884	2168	1025	707	1317	1620	1235	1493	1442	61	60	625	666	702	623	632	855	953	0
Previous	0	347	569	972	787	1025	1596	961	1129	1185	1131	1175	1193	72	53	538	600	569	550	553	919	1025	0
AAASurgery	0	551	602	764	704	707	961	3681	2677	2714	2667	2672	2676	153	55	736	795	790	721	731	1322	1316	0
Haemorrhage	0	1692	736	1912	1303	1317	1129	2677	8902	8840	8641	8710	8713	403	128	951	1087	1329	982	1010	1693	1800	0
Stroke	0	1895	754	2276	1600	1620	1185	2714	8840	9422	8731	9137	9067	408	132	1041	1171	1424	1068	1095	1738	1842	0
Myo	0	1658	739	1790	1219	1235	1131	2667	8641	8731	8770	8718	8716	406	122	951	1087	1330	982	1004	1694	1795	0
Cardiac	0	1798	749	2158	1477	1493	1175	2672	8710	9137	8718	9168	9061	408	123	1013	1142	1386	1037	1060	1714	1818	0
Hypo	0	1805	749	2142	1427	1442	1193	2676	8713	9067	8716	9061	9159	406	125	991	1141	1386	1039	1062	1726	1828	0
Status	0	141	44	66	52	61	72	153	403	408	406	408	406	443	17	89	105	114	101	102	107	109	0
Age	0	57	44	78	57	60	53	55	128	132	122	123	125	17	169	66	53	66	67	69	49	48	0
Haemoglobin	1	504	491	668	617	625	538	736	951	1041	951	1013	991	89	66	1460	1319	1297	1297	1298	909	912	0
WhiteCC	1	627	508	714	651	666	600	795	1087	1171	1087	1142	1141	105	53	1319	1719	1387	1348	1344	962	984	0
Urea	0	696	512	780	695	702	569	790	1329	1424	1330	1386	1386	114	66	1297	1387	1978	1461	1459	962	956	0
Sodium	0	591	500	681	614	623	550	721	982	1068	982	1037	1039	101	67	1297	1348	1461	1522	1489	925	922	0
Potassium	0	617	503	696	619	632	553	731	1010	1095	1004	1060	1062	102	69	1298	1344	1459	1489	1575	935	932	0
LowBP	0	580	701	842	851	855	919	1322	1693	1738	1694	1714	1726	107	49	909	962	962	925	935	2348	2232	0
HighPulse	0	575	714	862	951	953	1025	1316	1800	1842	1795	1818	1828	109	48	912	984	956	922	932	2232	2427	0
hospitalID	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.3: Table showing pairs of missing variables: AAA data

Appendix B

R Code

This appendix will give samples of the R code used during this study, along with brief descriptions.

B.1 Chapter 2

```
aaa<-read.csv(file="D:/Uni Work/Masters #initial data as a csv file
/Dissertation/aaa_missing.csv", header=TRUE, sep=",")

fisher.test(Gender, Status) #odds ratios

counts1 <- table(adMode,
hospitalID) #stacked barchart
barplot(counts1, main="Summary of Hospital
ID and Admission Mode",xlab="Hospital ID",
col=c("blue","orange","blue","blue","blue",
"orange","orange"),legend =rownames(counts1))
```

B.2 Chapter 3

```
pairs(body) #pairs plot of the data

Xcomp=na.omit(X) #complete cases analysis
modela=lm(Xcomp[,1]~Xcomp[,2]+Xcomp[,3]+Xcomp[,4]) #models
```

B.2.1 The Graphs Coding, eg, figure 3.3

```
names<-c("orig", "comp", "mean", "reg", #name the methods
 "sto reg", "multi10", "multi100")
intercept<-c(117.085,18.55,-9.46, #list the intercepts
-120.33,-17.61,-41.6,-10.46)
x<- 1:7*4-1 #set up the x-axis
CI.up<-c(312.7,362.4,11.78,172.89,
95.44,173.5,225.1) #upper CI
CI.dn<-c(-78.5,-325.3,-30.7,-336.10,
-208.1,-256.6,-264.03) #lower CI
plot(intercept˜x, col=c("black", "red", "blue", #plot graph
"green", "orange", "pink", "purple"), cex=1,
axes=FALSE, frame.plot=TRUE, xlim=c(1,28),
ylim=c(-330,370), xlab="Model Type", ylab=
"Estimate", main="Intercept Estimates and
Confidence Intervals")
Axis(side=1, labels=FALSE) #label axes
Axis(side=2, labels=TRUE)
axis(1, at=c(3,7,11,15,19,23,27), labels=names) #label methods
arrows(x,CI.dn,x,CI.up,code=3,length=0.05, #add CIs
angle=90,col=c("black", "red", "blue", "green"
, "orange", "pink", "purple"))
abline(h=117.085,col="lightgray",lty=5) #add ’original’ lines
abline(h=312.7,col="lightgray",lty=3)
abline(h=-78.5,col="lightgray",lty=3)
```

B.3 Chapter 4

```
sigma=cov(body)
sigma #find covariance matrix of 20 subjects
mu=c(mean(fat),mean(triceps),mean(thigh),mean(midarm))
mu #find mean vector of 20 subjects
is.vector(mu)
set.seed(1) #set the seed
sample=mvrnorm(n=1000,mu,sigma,empirical=TRUE)
sample #simulate 1000 subjects from 20

library(mice) #use mice package
quickpred(sample) #prediction matrix
image(is.na(t(X))) #image plots of missingness
```

B.3.1 MCAR

```
N <- 1000 #number of subjects
```

```
n <- 4 #number of variables
X <- matrix(sample, N, n) #matrix
pMiss <- 0.40 #percent of missing values (20%)
set.seed(1) #set the seed
idMiss <- sample(1:N, N*pMiss) #sample subjects
nMiss <- length(idMiss) #number missing
m <- 3 # maximum number of missing variables within subject
set.seed(1)
howmanyMiss <- sapply(idMiss, function(x) sample(1:m, 1))
set.seed(1) #set the seed
lapply(howmanyMiss, function(x) sample(1:n, x))
set.seed(1) set the seed
misscols<-lapply(howmanyMiss, function(x) sample(1:n, x))
for (i in 1:nMiss){  #set missing values o=to NA
        for (j in misscols[[i]]){
                X[idMiss[i],j]<-NA
        }
}
```

B.3.2 MAR

```
sample2=sample[sort.list(sample[,3]), ] #order by thigh
sample3=sample[sort.list(sample[,2]), ] #order by triceps
pMiss <- 50 #percent of missing values
N <- 1000 #number of subjects
sample2[,2][1:(N * pMiss / 100)]<-NA
                     #order: thigh, remove triceps
sample2[,4][1:(N * pMiss / 100)]<-NA
                     #order: thigh, remove midarm
sample5=sample2[sort.list(sample2[,1]), ]
                     #reorder by fat
sample3[,3][1:(N * pMiss / 100)]<-NA
                     #order: triceps, remove thigh
sample4=sample3[sort.list(sample3[,1]), ]
                     #reorder by fat
sample6<-cbind(sample5[,1],
 sample5[,2], sample4[,3], sample5[,4])
sample6 #combine columns to form new matrix: MAR
```

B.3.3 MNAR

```
sample2=sample[sort.list(sample[,2]), ] #order: triceps
pMiss <- 50 #percent of missing values
N <- 1000 #number of subjects
sample2[,2][1:(N * pMiss / 100)]<-NA #removed small triceps
sample3=sample[sort.list(sample[,3]), ] #order: thigh
sample3[,3][1:(N * pMiss / 100)]<-NA #removed small thigh
sample4=sample[sort.list(sample[,4]), ] #order: midarm
sample4[,4][1:(N * pMiss / 100)]<-NA #removed small midarm
sample5=sample2[sort.list(sample2[,1]), ] #reorder: fat
sample6=sample3[sort.list(sample3[,1]), ] #reorder: fat
sample7=sample4[sort.list(sample4[,1]), ] #reorder: fat
sample8=cbind(sample5[,1],sample5[,2],sample6[,3],sample7[,4])
sample 8 #combine to create new data set: MNAR
```

B.4 Chapter 5

```
marginplot(aaa[,c("AgeYears","Haemoglobin")], #marginplots
 col=c("blue","red","orange"), cex=1.5,
cex.lab=1.5, cex.numbers=1.3, pch=19)

pbox(aaacont,pos=1,int=FALSE,cex=1.2) #parallel boxplots

library(lattice) #use lattice package
com <- complete(imp2, "long", inc=T) #create stripplots
col <- rep(c("blue","red")[1+as.numeric(is.na
(imp2$data$Sodium))],11)
stripplot(Sodium˜.imp, data=com, jit=TRUE, fac=0.8,
 col=col, pch=20,
cex=1.4, xlab="Imputation number",main="Sodium")

imp2=mice(aaa,m=imputations,imputationMethod=c(
  "logreg", "polyreg", "logreg", "logreg", "logreg",
"logreg", "logreg", "polyreg", "logreg", "polyreg",
"logreg", "logreg", "logreg", "logreg", "pmm", "pmm",
"pmm", "pmm", "pmm", "pmm", "pmm", "pmm", "" ),
  predictorMatrix=pred, seed=204, printFlag=T)
   #mice on the AAA surgery data
```

```
imp2$imp$dischargeStatus=matrix(nrow=443,ncol=10,NA)
    #set the imputed outcomes to NA and remove subjects

fit3<-with(imp2,glm(dischargeStatus˜Gender+
admissionMode+AgeYears+
Urea+Sodium+Potassium+Haemoglobin+WhiteCellCount,
data=aaa, family=binomial)) #form regression model
est3<-pool(fit3) #pool: Rubin’s rules
summary(est3) #model summary
```