ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Statistical Methods for Analysis under the presence of missing data

By

Aikaterini L. Stamelakou

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
July 2016

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ

Στατιστικές Μέθοδοι για Ανάλυση υπό την παρουσία ελλιπών στοιχείων

Αικατερίνη Λ. Σταμελάκου

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Ιούλιος 2016

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

To my family…

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. V. Vasdeki for his kindness in accepting me as his advisee. His guidance through frustrating times was really helpful. He has always given directions when I needed to deal with complex issues, in a legitimate way. This project was very challenging, but ended successfully.

I'm also thankful to my family, who have always supported me in both moral and financial ways. Their concern gave me strength and free time to study and write a lot.

Finally, I sincerely thank my friends for their help and understanding all this year.

VITA

I was born in Chalkida, Evia, in 1989. I graduated the 4th High school of Chalkida in 2007 and in the same year I became a student in the Department of Mathematics, in the University of Crete (UOC). I took my degree in 2012. In 2014, I was accepted by the Master of Science in Statistics, in the Department of Statistics of Athens University of Economics and Business (AUEB). At this time I am about to complete my Postgraduate studies.

ABSTRACT

Aikaterini L. Stamelakou

Statistical Methods for analysis under the presence of missing data

July 2016

Missing data are a recurring problem which can cause bias or lead to inefficient analysis, no matter how well a survey questionnaire is designed and no matter how effective is the data collection. These data need a special and meticulous handling in analysis. This is why so many statistical methods have been proposed and developed to address missingness.

Some of them are based on deletion of incomplete cases, others try to predict each missing value and then to include the filled in value in analysis, these are called Simple Imputation Methods. Additionally, there is another method, known as Multiple Imputation, which is based on the creation of many imputed data sets by using Data Augmentation. In this thesis, each of these methods will be mentioned. Specifically, the Multiple Imputation method will be the main topic that will monopolize the interest and will be given special emphasis.

In the context of this thesis included and an application of Linear Mixed Models in repeated measurements with data that are not complete. Applying different mixed effect models on these data we reach in the appropriate model through the Bayesian Information Criterion. In continue, we apply multiple imputation in our data and then fit the same models in the imputed data this time. Our main goal is to examine the similarities or differences that may have these two data sets.

ΠΕΡΙΛΗΨΗ

Αικατερίνη Λ. Σταμελάκου

Στατιστικές Μέθοδοι για Ανάλυση υπό την παρουσία ελλιπών στοιχείων

Ιούλιος 2016

Τα δεδομένα που λείπουν είναι ένα επαναλαμβανόμενο πρόβλημα το οποίο μπορεί να προκαλέσει μεροληψία ή να οδηγήσει σε αναποτελεσματική ανάλυση. Δεν έχει σημασία πόσο καλά ένα ερωτηματολόγιο έχει σχεδιαστεί και δεν έχει σημασία πόσο αποτελεσματική είναι η συλλογή δεδομένων. Τα δεδομένα αυτά χρειάζονται μία ειδική και σχολαστική διαχείριση στην ανάλυση. Αυτός είναι ο κύριος λόγος που έχουν προταθεί και αναπτυχθεί για την αντιμετώπιση των ελλιπών στοιχείων τόσες πολλές στατιστικές μέθοδοι.

Μερικές από αυτές βασίζονται σε διαγραφή των ελλειπόντων περιπτώσεων, άλλες προσπαθούν να προβλέψουν κάθε τιμή που λείπει και στη συνέχεια να έχουμε ένα ολοκληρωμένο σύνολο δεδομένων, αυτές οι μέθοδοι ονομάζονται Simple Imputation. Επιπλέον, υπάρχει και μια άλλη μέθοδος, γνωστή ως Multiple Imputation, η οποία βασίζεται στη δημιουργία πολλών ολοκληρωμένων συνόλων δεδομένων. Συγκεκριμένα, η τελευταία μέθοδος θα μονοπωλήσει το ενδιαφέρον μας και θα δοθεί ιδιαίτερη έμφαση.

Στο πλαίσιο αυτής της διατριβής περιλαμβάνεται και μια εφαρμογή που αφορά μοντέλα μικτών επιδράσεων σε επαναλαμβανόμενες μετρήσεις με δεδομένα που δεν είναι πλήρη. Εφαρμόζοντας διαφορετικά μοντέλα μικτών επιδράσεων σε αυτά τα δεδομένα καταλήγουμε στο κατάλληλο μοντέλο μέσα από κάποιο κριτήριο. Εν συνεχεία, εφαρμόζουμε Multiple Imputation στα δεδομένα μας και εφαρμόζουμε ξανά τα ίδια μοντέλα στα νέα δεδομένα αυτή τη φορά. Βασικός μας στόχος είναι να εξεταστούν οι ομοιότητες ή διαφορές που ενδέχεται να έχουν αυτά τα δύο σύνολα δεδομένων, δηλαδή το ολοκληρωμένο αρχείο και εκείνο που περιλαμβάνει τιμές που λείπουν.

TABLE OF CONTENTS

Page

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

LIST OF TABLES

LIST OF FIGURES

Page

Figure

Chapter 1

1.1 Introduction

Standard statistical methods have been developed to analyze rectangular data sets. Generally, the rows of the data matrix represent units, also called cases, observations or subjects depending on context, and the columns represent variables measured for each unit. The entries in the data matrix are nearly always real numbers, either representing the values of essentially continuous variables or representing categories of response which may be ordered or unordered. When some of these entries in the matrix are not observed, we use to say that we have missing values in our data set. According to Schafer and Graham (2002) data contain various codes to indicate lack of response like “Don’t know”, “Refused”, “Unintelligible” and so on.

Also with rectangular data, there are several important classes of overall missing-data patterns. Consider *Figure 1.1* (a), in which missing values occur on an item Y but a set of p other items $X_1, \dots, X_p$ is completely observed, we call this a *univariate pattern*. The univariate pattern is also meant to include situations in which Y represents a group of items that is either entirely observed or entirely missing for each unit. In *Figure 1.1* (b), items or items groups $Y_1, \dots, Y_p$ may ordered in such a way that if Y_j is missing for a unit, then $Y_{j+1}, \dots, Y_p$ are missing as well. This is called a *monotone pattern*. *Figure 1.1* (c) shows an *arbitrary pattern* in which any set of variables may be missing for any unit.

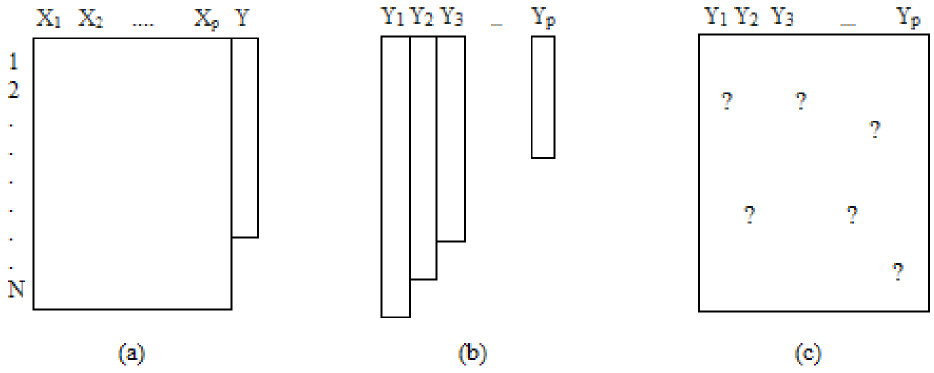

Figure 1.1 : Patterns of nonresponse in rectangular data sets: (a) univariate pattern, (b) monotone pattern, and (c) arbitrary pattern.

Because missingness may be related to the data, we classify distributions according to the nature of that relationship. Rubin (1976) developed a typology for these distributions that is widely cited but less widely understood.

Let us denote the complete data as Y_{com} and partition it as $Y_{com} = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis} are the observed and missing parts, respectively. Rubin (1976) defined missing data to be *missing at random* (MAR) if the distribution of missingness does not depend on Y_{mis}, (we refer to R as the missingness)

$$P(R|Y_{com}) = P(R|Y_{obs}). \tag{1.1}$$

In other words, MAR allows the probabilities of missingness to depend on observed data but not on missing data. An important special case of MAR, called *missing completely at random* (MCAR), occurs when the distribution does not depend on Y_{obs} either,

$$P(R|Y_{com}) = P(R).$$

When Equation 1.1 is violated and the distribution depends on Y_{mis}, the missing data are said to be *missing not at random* (MNAR).

Chapter 2

2.1 Analyzing methods of incomplete data sets

In this chapter, we will discuss briefly about methods for analyzing incomplete data sets. The basic methods are Deletion, Single Imputation and Multiple Imputation. Deletion method contains Complete Case (CC) and Available Case (AC). Single Imputation encloses mean, regression, hot-deck and cold-deck imputation respectively. We will focus on Multiple Imputation (MI) in another section.

2.1.1 Complete Case Analysis

The standard treatment of missing data in statistical packages is *Complete Case Analysis* (CC), where cases with any missing values are simply discarded. This method is also known as Listwise Deletion (LD) and is appropriate only when missing completely at random (MCAR) is a reasonable assumption for the missing data mechanism.

Complete Case (CC) analysis is a very simple method. By using it, we can make valid inference since all univariate statistics are calculated on a common sample base of cases. Moreover, the rejection of incomplete cases is an unnecessary waste of information, such loss of cases reduce statistical power. The loss in sample size can be considerable if the number of variables is large. One recommendation, which can be offered to mitigate the loss of cases, is to drop variables that have high levels of missing data while considering the degree of association between this variable and the others in the analysis. The Complete Cases are effectively a random sub-sample of the original cases, only when the data are missing completely at random (MCAR).

2.1.2 Available Case Analysis

This method, which is also well known as Pairwise Deletion (PD), includes all cases where the variable of interest is present. It is obvious that the sample base changes from variable to variable, which is actually a special disadvantage of this procedure. Also, an appropriate assumption is missing completely at random (MCAR) and at this case too.

Under this rare assumption it is easy to estimate unbiased means and variances, but is more complicated when we have to estimate measures of covariation, such as covariance or correlations. For example, if we want to calculate covariance (or correlation) between two variables Y_j and Y_k, we are based on cases i for which both y_{ij} and y_{ik} (i =1,…,n and j,k = 1,…,p , $j \neq k$) are present. So the estimate of covariance is

$$s_{jk}^{(jk)} = \sum_{(jk)} \frac{(y_{jk} - \bar{y}_j^{(jk)})(y_{ik} - \bar{y}_k^{(jk)})}{n_{jk}} \tag{2.1}$$

where n_{jk} is the number of cases where both Y_j and Y_k are observed, $\bar{y}_j^{(jk)}$ and $\bar{y}_k^{(jk)}$ are the sample means of Y_j and Y_k correspondingly over those n_{jk} cases. With the some procedure, the estimate of correlation between these two variables would be

$$r_{jk} = \frac{s_{jk}^{(jk)}}{\sqrt{s_{jj}^{(j)} s_{kk}^{(k)}}} \tag{2.2}$$

where $s_{jj}^{(j)} = \sum_{(j)} \frac{(y_{jk} - \bar{y}_j^{(jk)})^2}{n_j}$ is the variance of Y_j over n_j cases. A criticism of Equation (2.2) is that, unlike the population correlation being estimated, r_{jk}, which describes Pearson's correlation between Y_j and Y_k, can lie out of the range (-1,1) because $r_{jk} \notin$ (-1,1). As a solution to this problem, we use to compute pairwise correlations, where variances are estimatwd from the sample base as the covariance.

$$r_{jk}^{(jk)} = \frac{s_{jk}^{(jk)}}{\sqrt{s_{jj}^{(jk)} s_{kk}^{(jk)}}} \tag{2.3}$$

2.1.3 Mean Imputation

Refers to the procedure through which we substitute the missing values on a variable with the mean of the observed values for the same variable. In case of categorical data, the mode instead of the mean is used. So the overall respondent mean $\bar{y}_r$ for each variable, is assigned to all missing responses $y_{mis,i} = \bar{y}_r$.

In case there are auxiliary variables (which are fully observed) is preferable to use Mean Imputation within classes. In this procedure, firstly, we divide the total sample into H imputation classes according to values on auxiliary variables, in order to achieve homogeneity within classes. Within each class, the respondent mean for y-variable ($\bar{y}_{rh}$, h=1,2,…,H) is assigned to all the non-respondents in that class, so $\bar{y}_{hmis,i}$ = $\bar{y}_{rh}$ for the i-th non-respondent in class h. The classes may be defined as the cells in the crosstabulation of the (categorized) auxiliary variables, but symmetry is not essential; instead some auxiliary variables may be used for the one part of the sample while others are used for another part, or group of cells may be combined.

However, while this approach may be valid especially if the data are missing completely at random (MCAR), it is argued that mean substitution leads to an underestimation of the true population parameter particularly in simulations, where a segment of the population is more prone to non-response. The variance of the observed and imputed values of Y_j is

$$s^{(*)} = \frac{(n_j-1)}{(n-1)} s_{jj}^{(j)} \tag{2.4}$$

where $s_{jj}^{(j)}$ is the estimated variance from available cases. Under MCAR assumption, $s^{(*)}$ is a consistent estimate of the true variance, so the sample variance from the filled-in datasets underestimates the variance by a factor of $\frac{(n_j-1)}{(n-1)}$. This underestimation is a consistent estimate of imputing missing values at the center of distribution. Similarly, the sample covariance of Y_j and Y_k from the filled-in data is

$$s_{jk}^{(*)} = \frac{(n_{jk}-1)}{(n-1)} s_{jk}^{(jk)} \tag{2.5}$$

where $s_{jk}^{(jk)}$ is the estimated covariance when both Y_j and Y_j are observed. Since $s_{jk}^{(jk)}$ is a consistent estimate of the covariance, again the estimate from filled-in data underestimates the magnitude of the covariance by a factor $\frac{(n_{jk}-1)}{(n-1)}$. Thus, although the covariance matrix from the filled-in data is positive semi definite, the variances and covariances are systematically underestimated.

However, mean substitution has also been criticized on the grounds that it distorts the empirical distribution of the variable and that will be a problem when one wants to examine the shape (e.g. histogram, skewness) of the variable.

2.1.4 Regression Imputation Procedure

In this case, we assume a variable Y with missing data and p auxiliary variables $X_1, X_2, \dots, X_p$. This procedure, which is based on regression analysis, has two different versions: deterministic and stochastic. In fact, deterministic version serves well for estimating means and totals, but it distorts distributional properties of the variable; stochastic version, on the other hand, is less efficient for estimating means and totals but it preserves the variability in the observed data.

- **Predicted Regression Imputation**

This is the deterministic version of the regression imputation method. This method uses respondent data to regress Y on the auxiliary variables $X_1, X_2, \dots, X_p$. Missing Y – values is then imputed as the predicted values from the regression equation

$$Y_{mis,i}= a_0+ \sum_{i=1}^{p} X_i \qquad (2.6)$$

If the Y variable is qualitative, log-linear or logistic models may be used. As in any regression analysis, specific interaction terms may be included in the regression equation and also transformations may be useful. Little (1992) notes that estimated standard errors of the regression coefficients from Ordinary Least Squares (OLS) or Weighted Least Squares (WLS) tend be too small, because imputation error is not taken into account.

A special case of the previous regression model (2.6) is the ratio model. In this case, the regression model is

$$Y_{mis,i}= \frac{\bar{y}_R}{\bar{x}_R} X_i \qquad (2.7)$$

with a single auxiliary variable and an intercept of zero. That is, the value, which is used as a donor, is the ratio of Y variable mean with X variable mean multiplied by value of X in position that Y is missing and we are willing to complete.

- **Random Regression Imputation**

In this case, the imputed values are the predicted values from the regression equation (2.6) plus residual terms $e_{mis,i}$. So, the appropriate model, which describes the stochastic version of regression imputation, is

$$Y_{mis,i} = \hat{a}_0+\sum_{i=1}^{p}\hat{\beta}_i\, X_\iota + e_{mis,i} \tag{2.8}$$

Depending on the assumptions made, the residuals can be determined in various ways, including:

a. If the residuals are assumed to be homoscedastic and normally distributed, a residual can be chosen at random from a normal distribution with zero mean and variance equal to the residuals variance from the regression $\{e_{mis}\sim N(0,\sigma^2)\}$.

b. If the residuals are assumed to come from the same, unspecified distribution, they can be chosen at random from the respondents' residual.

c. As a protection against non-linearity and non-additivity in the regression model, the residuals may be taken from respondents with similar values on the auxiliary variable.

2.1.5 Hot-deck Imputation

Hot-deck procedures are common methods for adjusting data sets for missing values. Because *hot-deck* procedures originated in survey practice with little theory to direct their development, the statistical literature provides few definitions or results about these procedures. Widespread practice in the absence of well-developed theory clouds the subject with ambiguities and inconsistencies.

In general, a *hot-deck* procedure is a duplication process, when a value is missing from a sample a reported value from the same sample is duplicated to represent this missing value. The adjective "*hot*" refers to imputing with values from the current sample.

The most common techniques within the Hot-deck imputation are Flexible Matching Imputation, Nearest Neighbor Hot-deck and Sequential Hot-deck.

2.1.6 Cold-deck Imputation

In this procedure, when a value is missing from a sample, another value from another survey or another sample is used to represent the missing value. In general, two basic types of this substitution procedure are used:

1. Selection of a random substitute.
2. Selection of a specially designed substitute.

With a random substitution procedure, an additional population unit is selected on a probability bases to replace each non-respondent. Usually the substitute for a particular non-respondent is chosen from a restricted population of subgroups. On the other hand, a procedure that uses specially designated substitute units identifies one or more backup units to provide substitutes, if necessary, for each sample unit.

2.3 Multiple Imputation

Multiple Imputation (MI) appears to be one of the most attractive methods for general purpose handling of missing data in multivariate analysis. The basic idea, first proposed by Rubin (1978), is quite simple. This idea is based on creation $m \geq 2$ complete “imputed” datasets. We analyze each one of them, by using standard complete data methods and finally these m complete data inferences can be combined to form one inference that properly reflects uncertainty due to non-response under that model.

Specifically, in this procedure, there are $m \geq 2$ possible values for each missing value, (a vector $m \times 1$), which are ordered in the sense that m complete data sets can be created from the vectors of imputation. Each time, we replace each missing value by one of the components in its vector and we create a complete data set. Standard complete data methods are used to analyze each of m complete data sets. For example, one could perform linear or logistic regression procedures using any standard statistical package. Any model would have to be fitted m times, one for each imputed data set and the results across these data sets will vary as a reflection of missing data uncertainty. So, we obtain an overall set of estimated coefficients and standard errors from these m data sets and then we want to combine the results using certain rules that will be discussed below. In fact, the variability among the results of the m analyses provide a measure of the ordinary sample variation, lead to a single inferential statement about the parameters of interest.

The Multiple Imputation (MI), a simulation based technique, has been developed in an attempt to give solutions to problems because Single Imputation (SI) has two obvious disadvantages. Specifically, single imputation is unable to express the sampling variability under one model for non-response and the uncertainty about the correct model of non-response. Both these disadvantages don't exist in Multiple Imputation, which also shares the advantages of single imputation. That is we can use all standard complete data methods of analysis and also, in many analyses, data collector (e.g. Census Bureau) and data analysts (e.g. a university social scientist) may be different individuals, which is very important because the data collector may have access to more and better information about non-respondents than the data analyst. For example, in some cases, information protected by secrecy constrains (e.g. zip codes of dwelling units) may be available to help impute missing values (e.g. annual incomes).

As we have seen, Multiple Imputation is better than Single, because shares the advantages of Single Imputation and also rectifies disadvantages. The only disadvantage of this procedure is that it requires more work than Single Imputation. The cost of using Multiple Imputation is the computational complexity, space of databases and time required due to the fact that many different sets and samples have to be available at any time.

Finally, we ought to notice that the appropriate number of imputations mostly required is quite small. Usually, we can obtain good results with m as small as 3-5. Why only a few imputations are needed? Actually, this fact is very strange in comparison to the number of repetitions, which are usually required to the EM algorithm or in Data Augmentation. On the other hand, this is quite logic, because firstly, with this procedure, we only desire to solve the missing data aspect of the problem, without decreasing Monte Carlo error, and secondly the rules for combining the m complete data analyses explicitly account for Monte Carlo error.

2.3.1 Rubin's Rules

Rubin (1987) provides a procedure from the m imputations. We use Schafer's (1997) notation. Consider that we want to make inference about a quantity Q in the complete data case. Let $\hat{Q}$

be an estimate of Q that we use if no data were missing and U an estimated variance of Q. Because both these quantities are related.

With m imputations, we calculate m different versions of $\hat{Q}$ and U .

$$\hat{Q}^{(t)} = \hat{Q}(Y_{obs}, Y_{mis})$$

and

$$U^{(t)} = U(Y_{obs}, Y_{mis})$$

be the point and variances estimates using the t -th set of imputed datasets. According to Rubin, the estimate of Q , which combines the m complete data estimates $\hat{Q}^{(t)}$, is the average of these estimates.

$$\overline{Q} = \frac{1}{m}\sum_{t=1}^{m} \hat{Q}^{(t)} \tag{2.9}$$

The variability associated with this estimate has two components:

a. Within imputation variance, which is the average of variance estimates of imputed data.

$$\overline{U} = \frac{1}{m}\sum_{t=1}^{m} U^{(t)} \tag{2.10}$$

b. Between imputation variance, which is the variance of the complete data estimates.

$$B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{Q}^{(t)} - \bar{Q})^2 \tag{2.11}$$

Then, the total variance is defined as :

$$T = \overline{U} + \frac{m}{m+1} B \tag{2.12}$$

which T is equal to $\bar{U}$ plus B corrected for m being finite by the term $\frac{m}{m+1}$ and the inferences are based on the following approximation

$$T^{-\frac{1}{2}}\ (Q-\bar{Q}) \sim t_{df} \tag{2.13}$$

where the t distribution the degrees of freedom (df) are calculated as:

$$df = (m+1)[1+\frac{1}{m+1}\frac{\bar{U}}{B}] \tag{2.14}$$

Equation (2.14) shows that the degrees of freedom are depended by both m and $\frac{\bar{U}}{B}$. That is, as the number of imputations m increases the degrees of freedom increases. Also, as $\frac{\bar{U}}{B}$ increases the df gets larger. According to Schafer and Olsen (1998), if the degrees of freedom are small, less than 10, the estimates will be more accurate when the number of imputations m is large. On the other hand, if the computed value of df is large, greater than 10, little will be gained from a larger m. In fact, when df is large, we may assume that statistic $T^{-\frac{1}{2}}\ (Q-\bar{Q})$ is asymptotically normal. According to Equation (2.13), an 100(1-α)% interval estimate for Q is:

$$\bar{Q} \pm t_{df,1-\frac{\alpha}{2}}\sqrt{T} \tag{2.15}$$

and an appropriate $p-value$ for testing the null hypothesis $Q = Q'$ against a two sided alternative is:

$$p-value = 2P(t_{df} > T^{-\frac{1}{2}}|\bar{Q}-Q'|) \tag{2.16}$$

The ratio $\frac{B}{\bar{U}}$, which indicates how much information is missing, is an estimator of $\frac{\gamma}{1-\gamma}$, where γ is the fraction of information missing for Q due to non-response. If γ is zero then B goes to zero. This quantity γ is defined as:

$$\gamma = \frac{r + \frac{2}{df+3}}{r+1} \tag{2.17}$$

where $r = (1 + \frac{1}{m})\frac{B}{\bar{U}}$ indicates the relative increase in variance due to non-response, because $\bar{U}$ represents the estimated total variance, where is no missing information about Q (B=0). Both γ and r can be used as diagnostic statistics to examine the effect of missing data on estimates of $\bar{Q}$.

Chapter 3

3.1 Introduction

In this chapter we present an overview of linear mixed-effects models. In practice, longitudinal data are often highly unbalanced in the sense that are not equal number of measurements is available for all subjects and/or that measurements are not taken at fixed time points. Due to their unbalanced nature, many longitudinal data sets cannot be analyzed using multivariate regression techniques. A natural alternative arises from observing that subject-specific longitudinal profiles can often be well approximated by linear regression functions.

Many common statistical models can be expressed as linear models that incorporate both fixed effects, which are parameters associated with an entire population or with certain repeatable levels of experimental factors, and random effects, which are associated with individual experimental units drawn at random from a population. Fixed effects factors are generally thought of as fields whose values of interest are all represented in the dataset, and can be used for scoring. Random effects factors are fields whose values in the data file can be considered a random sample from a larger population of values. They are useful for explaining excess variability in the target. A model with both fixed effects and random effects is called a mixed-effects model. Mixed-effects models are primarily used to describe relationships between a response variable and some covariates in data that are grouped according to one or more classification factors. By associating common random effects to observations sharing the same level of a classification factor, mixed-effects models flexibly represent the covariance structure induced by the grouping of the data.

3.2 The General Linear Mixed Model

According to Verbeke and Molenberghs (2000) at the first stage of the two-stage approach assume that the random variable Y_{ij} denote the (possible transformed) response of interest, for the ith individual, measured at time t_{ij} , $i = 1,\dots,N$, $j=1,\dots, n_i$, and let $\mathbf{Y}_i$ be the n_i-dimensional vector of all repeated measurements for the ith subject, that is, $\mathbf{Y}_{\boldsymbol{i}} = (Y_{i1}, Y_{i2}, \ldots, Y_{i\,\mathrm{ni}})'$. The first stage assumes that $\mathbf{Y}_{\boldsymbol{i}}$ satisfies the linear regression model:

$$\boldsymbol{Y_i} = Z_i\beta_i + \varepsilon_i \qquad (3.1)$$

where Z_i is a (n_i x q) matrix of known covariates, modeling how the response evolves over time for the ith subject. Further, β_i is a q-dimensional vector of unknown subject-specific regression coefficients, and $\varepsilon_{\boldsymbol{i}}$ is a vector of residual components ε_{ij} , j=1,…, n_i. It is usually assumed that all $\varepsilon_{\boldsymbol{i}}$ are independent and normally distributed with mean vector zero, and covariance matrix $\sigma^2 I_{ni}$, where I_{ni} is the n_i-dimensional identity matrix. Obviously, model (3.1) includes very flexible models for the description of subject-specific profiles.

In the second stage, a multivariate regression model of the form

$$\beta_i = K_i\beta + b_i \qquad (3.2)$$

is used to explain the observed variability between the subjects, with respect to their subject-specific regression coefficients β_i. K_i is a (q x p) matrix of known covariates, and β is a p-dimensional vector of unknown regression parameters. Finally, the b_i are assumed to be independent, following a q-dimensional normal distribution with mean vector zero and general covariance matrix D.

In practice, the regression parameters in (3.2) are of primary interest. They can be estimated by sequentially fitting the models (3.1) and (3.2). First, all β_i are estimated by fitting model (3.1) to the observed data vector y_i for each subject separately, yielding estimates $\widehat{\beta_\iota}$. Afterward, model (3.2) is fitted to the estimates $\widehat{\beta_\iota}$, providing inferences for β.

This two-stage analysis can be interpreted as the calculation (first stage) and analysis (second stage) of summary statistics. First, the actually observed data vector y_i is summarized by $\widehat{\beta_\iota}$, for each subject separately. Subsequently, regression methods are used to assess the relation between the so-obtain summary statistics and relevant covariates. Other summary statistics frequently used in practice are the area under each individual profile, the mean response for each individual, the largest observation, the half time, and so forth.

As for any analysis of summary statistics, the two-stage analysis obviously suffers from at least two problems. First, information is lost in summarizing the vector y_i of observed measurements for the *i*th subject by $\widehat{\beta_i}$. Second, random variability is introduced by replacing the β_i in model (3.2) by their estimates $\widehat{\beta_i}$. Moreover, the covariance matrix of $\widehat{\beta_i}$ highly depends on the number of measurements available for the *i*th subject as well as on the time points at which these measurements were taken, and this is has not been taken into account in the second stage of the analysis.

In order to combine the models from the two-stage analysis, we replace β_i in (3.1) by expression (3.2), yielding

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i \tag{3.3}$$

where $X_i = Z_i K_i$ is the appropriate (n_i x p) matrix of known covariates, and where all other components are defined earlier. Model (3.3) is called linear mixed effects model with fixed effects β and with subject-specific effects b_i. It assumes that the vector of repeated measurements on each subject follows a linear regression model where some of the regression parameters are population-specific. The b_i are assumed to be random and are therefore often called random effects.

For example, a model with fixed effects β_j and random effects b_i could be written as

$$Y_{ij} = \beta_j + b_i + \varepsilon_{ij}, \qquad i = 1, \dots, 9, \qquad j = 1, \dots, 4,$$

$$b_i \sim N(0, \sigma_b^2), \qquad \varepsilon_{ij} \sim N(0, \sigma^2),$$

or, equivalently,

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\beta + \boldsymbol{Z_i} b_i + \boldsymbol{\varepsilon_i}\,, \qquad i = 1, \dots, 9$$

$$b_i \sim N(0, \sigma_b^2), \qquad \boldsymbol{\varepsilon_i} \sim N(0, \sigma^2 \boldsymbol{I}),$$

where, for i =1,...,9,

$$\boldsymbol{Y_i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix}, \quad \boldsymbol{X_i} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{Z_i} = \boldsymbol{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\varepsilon_i} = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix}.$$

In general, a linear mixed-effects model is any model which satisfies (Laird and Ware, 1982)

$$
\begin{aligned}
&* \quad Y_i = X_i\beta + Z_i b_i + \varepsilon_i \\
&* \quad b_i \sim \mathrm{N}(0, D), \\
&* \quad \varepsilon_i \sim \mathrm{N}(0, \Sigma_i), \\
&* \quad b_1, \dots, b_N, \varepsilon_1, \dots, \varepsilon_N \ \text{ independent,}
\end{aligned}
\tag{3.4}
$$

where Y_i is the n_i-dimensional response vector for subject i, $1 \leq i \leq N$, N is the number of subjects, X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates, β is an p-dimensional vector containing the fixed effects, b_i is the q-dimensional vector containing the random effects, and ε_i is an the n_i-dimensional vector of residual components. Finally, D is a general $(q \times q)$ covariance matrix with (i, j) elements $d_{ij} = d_{ji}$ and Σ_i is a $(n_i \times n_i)$ covariance matrix which depends on i only through its dimension n_i, i.e. the set of unknown parameters in Σ_i will not depend upon i. In some cases, one may wish to relax this last assumption.

It follows from (3.4) that, conditional on the random effect $\boldsymbol{b_i}$, $\boldsymbol{Y_i}$ is normally distributed with mean vector $X_i\beta + Z_i\boldsymbol{b_i}$ and with covariance matrix Σ_i. Further, $\boldsymbol{b_i}$ is assumed to be normally distributed with mean vector 0 and covariance matrix D. Let $f\,(\boldsymbol{y_i}|\boldsymbol{b_i})$ and $f\,(\boldsymbol{b_i})$ be the corresponding density functions. The marginal density function of $\boldsymbol{Y_i}$ is then given by:

$$f\,(\boldsymbol{y_i}) = \int f\,(\boldsymbol{y_i}|\boldsymbol{b_i})\, f(\boldsymbol{b_i})\, \mathrm{d}\boldsymbol{b_i}\,,$$

which can easily be shown to be the density function of a n_i–dimensional normal distribution with mean vector $X_i\beta$ and with covariance matrix $V_i = Z_i D Z_i' + \Sigma_i$. Hence, the marginal model implied by the two-stage approach makes very specific assumptions about the dependence of the mean structure and the covariance structure on the covariates X_i and Z_i, respectively.

Since, model (3.4) is defined through the distributions $f\,(\boldsymbol{y_i}\,|\boldsymbol{b_i})$ and $f\,(\boldsymbol{b_i})$, it will be called the hierarchical formulation of the linear mixed model. The corresponding marginal normal distribution with mean $X_i\beta$ and covariance $Z_i D Z_i' + \Sigma_i$, is called the marginal formulation of the model. Note that, although the marginal model naturally follows from the hierarchical one, both models are not equivalent.

3.3 Estimation the Marginal Model

As we discussed the general linear mixed model (3.4) implies the marginal model

$$\mathbf{Y}_{\boldsymbol{i}} \sim N\,(X_i\beta, Z_iDZ_i' + \Sigma_i). \tag{3.5}$$

Inference is based on this marginal distribution for the response $\mathbf{Y}_{\boldsymbol{i}}$. It should be emphasized that the hierarchical structure of the original model (3.4) is then not taken into account. Indeed, the marginal model (3.5) is not equivalent to the original hierarchical model (3.4). Inferences based on the marginal model do not explicitly assume the presence of random effects representing the natural heterogeneity between subjects.

Let $\boldsymbol{\alpha}$ denote the vector of all variance and covariance parameters (usually called variance components) found in V_i=$Z_iDZ_i' + \Sigma_i$, that is, $\boldsymbol{\alpha}$ consists of the $\frac{q(q+1)}{2}$ different elements in D and of all parameters in Σ_i. Finally, let $\boldsymbol{\theta}$ = (β', α')' be the s-dimensional vector of all parameters in the marginal model for $\mathbf{Y}_{\boldsymbol{i}}$, and let $\Theta = \Theta_\beta \times \Theta_\alpha$ denote the parameter space for $\boldsymbol{\theta}$, with Θ_β and Θ_α the parameter spaces for the fixed effects and for the variance components respectively. Note that $\Theta_\beta = \mathbb{R}^p$, and Θ_α equals the set of values for $\boldsymbol{\alpha}$ such that D and all Σ_i are positive (semi-)definite. The classical approach to inference is based on estimators obtained from maximizing the marginal likelihood function:

$$\mathrm{L}_{\mathrm{ML}}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ (2\pi)^{-\frac{n_i}{2}} |V_i(\boldsymbol{a})|^{-\frac{1}{2}} \times \exp(-\frac{1}{2}(\boldsymbol{Y_i} - X_i\beta)' V_i^{-1}(\boldsymbol{a})(\boldsymbol{Y_i} - X_i\beta)) \right\} \tag{3.6}$$

with respect to $\boldsymbol{\theta}$. Let us first assume $\boldsymbol{\alpha}$ to be known. The maximum likelihood estimator (MLE) of β, obtained from maximizing (3.6), conditional on $\boldsymbol{\alpha}$, is then given by (Laird and Ware, 1982)

$$\hat{\beta}(\alpha) = (\sum_{i=1}^{N} X'_i W_i X_i)^{-1} \sum_{i=1}^{N} X'_i W_i Y_i\,, \tag{3.7}$$

where W_i equals V_i^{-1}.

When $\boldsymbol{\alpha}$ is not known, but an estimate of $\widehat{\boldsymbol{\alpha}}$ is available, we can set $\widehat{V_i}$ = $V_i(\widehat{\boldsymbol{\alpha}})$ = $\widehat{W_i}^{-1}$, and estimate $\boldsymbol{\beta}$ by using the expression (3.7) in which W_i is replaced by $\widehat{W_i}$. In continue, we will see two frequently used methods for estimating $\boldsymbol{\alpha}$, these are the maximum likelihood estimation (MLE) and the restricted maximum likelihood (REML) estimation.

3.3.1 Maximum Likelihood Estimation

The maximum likelihood estimation (MLE) of $\boldsymbol{\alpha}$ is obtained by maximizing the expression (3.6) with respect to $\boldsymbol{\alpha}$, after $\boldsymbol{\beta}$ is replaced by (3.7). This approach arises naturally when we consider the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ simultaneously by maximizing the joint likelihood (3.6).

3.4 Restricted Maximum Likelihood Estimation

In practice, linear mixed models often contain many fixed effects. In such cases, it may be important to estimate the variance components, explicitly taking into account the loss of degrees of freedom involved in estimating the fixed effects. In contrast to simple cases, an unbiased estimator for the vector $\boldsymbol{\alpha}$ of variance components cannot be obtained from appropriately transforming the ML estimator as suggested from the analytic calculation of its bias. However, the error contrasts approach can still be applied as follows. We first combine all N subject-specific regression models (3.4) to one model:

$$\boldsymbol{Y} = X\beta + Z\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{3.8}$$

where the vectors $\boldsymbol{Y}$, $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$, and the matrix X are obtained from stacking the vectors $\boldsymbol{Y_i}$, $\boldsymbol{b_i}$ and $\boldsymbol{\varepsilon_i}$, and the matrices X_i respectively, underneath each other, and where Z is the block-diagonal matrix with blocks Z_i on the main diagonal and zeros elsewhere. The dimensional of $\boldsymbol{Y}$ equals $\sum_{i=1}^{N} n_i$ and will be denoted by n.

The marginal distribution for $\boldsymbol{Y}$ is normal with mean vector $X\beta$ and with covariance matrix V($\boldsymbol{\alpha}$) equal to the block-diagonal matrix with blocks V_i on the main diagonal and zeros elsewhere. The REML estimator for the variance components $\boldsymbol{\alpha}$ is now obtained from maximizing the likelihood function of a set of error contrasts $\boldsymbol{U} = \boldsymbol{A'Y}$ where A is any $(n \times (n-p))$ full-rank matrix with columns orthogonal to the distribution with mean vector zero and covariance matrix $A'V(\boldsymbol{\alpha})A$, which is not dependent on β any longer. Further, Harville (1974) has shown that the likelihood function of the error contrasts can be written as:

$$\mathrm{L}(\boldsymbol{\alpha}) = (2\pi)^{-(n-p)/2} \left|\sum_{i=1}^{N} X'_i X_i\right|^{1/2}$$

$$\times \left|\sum_{i=1}^{N} X'_i\, V_i^{-1} X_i\right|^{-1/2} \prod_{i=1}^{N} |V_i|^{-1/2}$$

$$\times \exp\left\{-\frac{1}{2}\sum_{i=1}^{N} (\boldsymbol{Y_i} - X_i\hat{\beta})'\, V_i^{-1}(\boldsymbol{Y_i} - X_i\hat{\beta})\right\}, \qquad (3.9)$$

where $\hat{\beta}$ is given by (3.7). Hence, the so-obtained REML estimator $\hat{\boldsymbol{\alpha}}$ does not depend on the error contrasts (i.e. the choice of A).

Note that the maximum likelihood estimator for the mean of a univariate normal population and for the vector of regression parameters in a linear regression model are independent of the residual variance σ^2. However, it follows from (3.7) that this no longer holds in the general linear mixed model. Thus, we have that although REML estimation id only with respect to the variance components in the model, the "REML" estimator for the vector of fixed effects is not identical to its ML version.

Finally, mention that the likelihood function in (3.9) equals:

$$\mathrm{L}(\boldsymbol{\alpha}) = \mathrm{C} \left|\sum_{i=1}^{N} X'_i\, W_i(\boldsymbol{\alpha}) X_i\right|^{-1/2} \mathrm{L}_{\mathrm{ML}}\,(\hat{\beta}(\boldsymbol{\alpha}), \boldsymbol{\alpha}) \qquad (3.10)$$

where C is a constant not depending on $\boldsymbol{\alpha}$,where, as earlier, $W_i(\boldsymbol{\alpha})$ equals $\mathrm{V}_i^{-1}(\alpha)$ and where $\mathrm{L}_{\mathrm{ML}}(\beta, \boldsymbol{\alpha}) = \mathrm{L}_{\mathrm{ML}}(\boldsymbol{\theta})$ is the ML likelihood function given by (3.6). Because $\left|\sum_{i=1}^{N} X'_i\, W_i(\boldsymbol{\alpha}) X_i\right|$ in (3.10) does not depend on β, it follows that the REML estimators for $\boldsymbol{\alpha}$ and for β can also be found by maximizing the so-called REML likelihood function

$$\mathrm{L}_{\mathrm{REML}} = \left|\sum_{i=1}^{N} X'_i\, W_i(\boldsymbol{\alpha}) X_i\right|^{-1/2} \mathrm{L}_{\mathrm{ML}}\,(\boldsymbol{\theta}) \qquad (3.11)$$

with respect to all parameters simultaneously ($\boldsymbol{\alpha}$ and β).

3.4.1 Justification of REML Estimation

The main justification of the REML approach has been given by Patterson and Tompson (1971), who proved that, in the absence of information on β, no information about $\boldsymbol{\alpha}$ is lost when inference is based on $\boldsymbol{U}$ rather than on $\mathbf{Y}$. More precisely, $\boldsymbol{U}$ is marginally sufficient for $\boldsymbol{\alpha}$ in the sense described by Sprott (1975). Further, Harville (1974) has shown that, from a Bayesian point of view, using only error contrasts to make inferences on $\boldsymbol{\alpha}$ is equivalent to ignoring any prior information on β and using all the data to make those inferences.

3.4.2 Comparison between ML and REML Estimation

Maximum likelihood estimation and restricted maximum likelihood estimation both have the same merits of being based on the likelihood principle which leads to useful properties such as consistency, asymptotic normality and efficiency. ML estimation also provides estimators of the fixed effects, whereas REML estimation, in itself, does not. On the other hand, for balanced mixed ANOVA models, the REML estimates for the variance components are identical to classical ANOVA-type estimates obtained from solving the equations which set mean squares equal to their expectations. This implies optimal minimum variance properties, and it shows that REML estimates in that context do not rely on any normality assumption since only moment assumptions are involved.

Also with regard to the mean squared error for estimating $\boldsymbol{\alpha}$, there is no indisputable preference for either one of the two estimation procedures, since it depends on the specifics of the underlying model and possibly on the true value of $\boldsymbol{\alpha}$. For ordinary ANOVA or regression models, the ML estimator of the residual variance σ^2 has uniformly smaller mean squared error than the REML estimator when $p = rank(X) \leq 4$, but the opposite is true when $p > 4$ and $n - p$ is sufficiently large ($n - p > 2$ suffices if $p > 12$). In general, one may expect results from ML and REML estimation to differ more as the number p of fixed effects in the

Chapter 4

4.1 The experimental design

The purpose of this Chapter is to evaluate the level of the clinical attachment loss (CAL) and especially the factors, which affect tis level, on patients with early onset or aggressive periodontitis (EOP). Early-onset periodontitis is a type of periodontitis, which is characterized by severe attachment loss and bone destruction in otherwise healthy patients with a tendency to familiar aggregation. Clinical attachment level (CAL) is the measured distance to the nearest mm from the cemento-enamel junction to the deepest probeable pocket point.

Twenty-five patients with a diagnosis of early-onset or aggressive periodontitis (EOP) who received treatment at a private practice limited to periodontics in Athens, Greece, participated in the study. The patients were included in the study only if they had complied with a minimum of 10 supportive periodontal care (SPC) sessions during the 5-year maintenance phase. The group consisted of 11 males and 14 females, from 30 to 39 years old. Also, ten patients were smokers with an average of 22.5 cigarettes/day. For each mouth we have measured the clinical indicator CAL for two teeth.

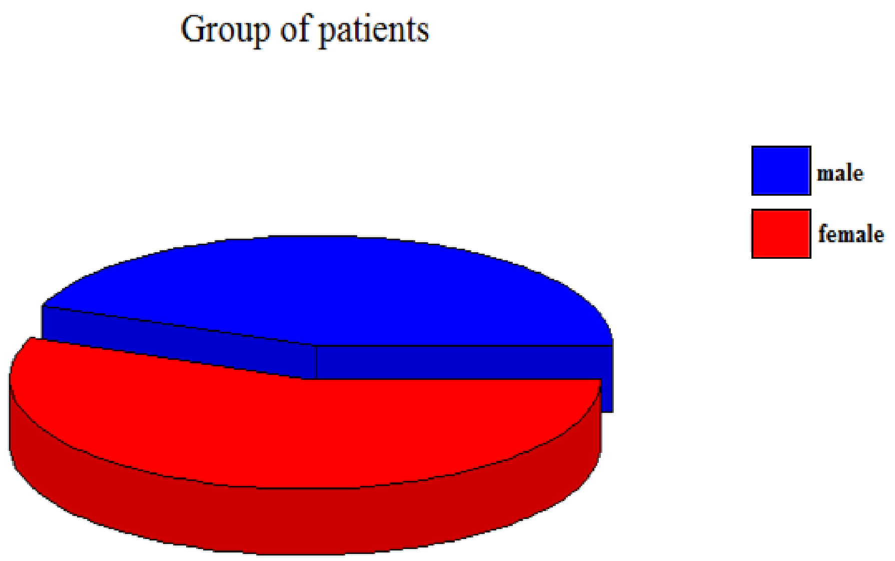

Figure 4.1 : The pie-chart of the patients' sex.

Initially, oral hygiene instructions (bass brushing methods, dental floss and interdental brushes) scaling and root planning were advised to the patients (SRP). SRP was performed under local anesthesia and required about 60–90 min. Periodontal surgery and systemic antibiotics following microbiological testing were performed when indicated. Antibiotics were either ornidazole (Betiral, Roche, Basel, Switzerland) or tinidazole (Fasigyn, Pfizer) for 7 days at each course. More specifically, ten patients received SRP treatment, five patients received SRP treatment and antibiotics, four patients received SRP treatment and periodontal surgery and finally, six patients received SRP treatment, periodontal surgery and antibiotics. Three months later, all patients were recalled. Additionally, they were enrolled in a maintenance care program with annually measurements of the clinical attachment level. All clinical procedures were carried out by the same periodontist with a time limit of ten minutes approximately per tooth.

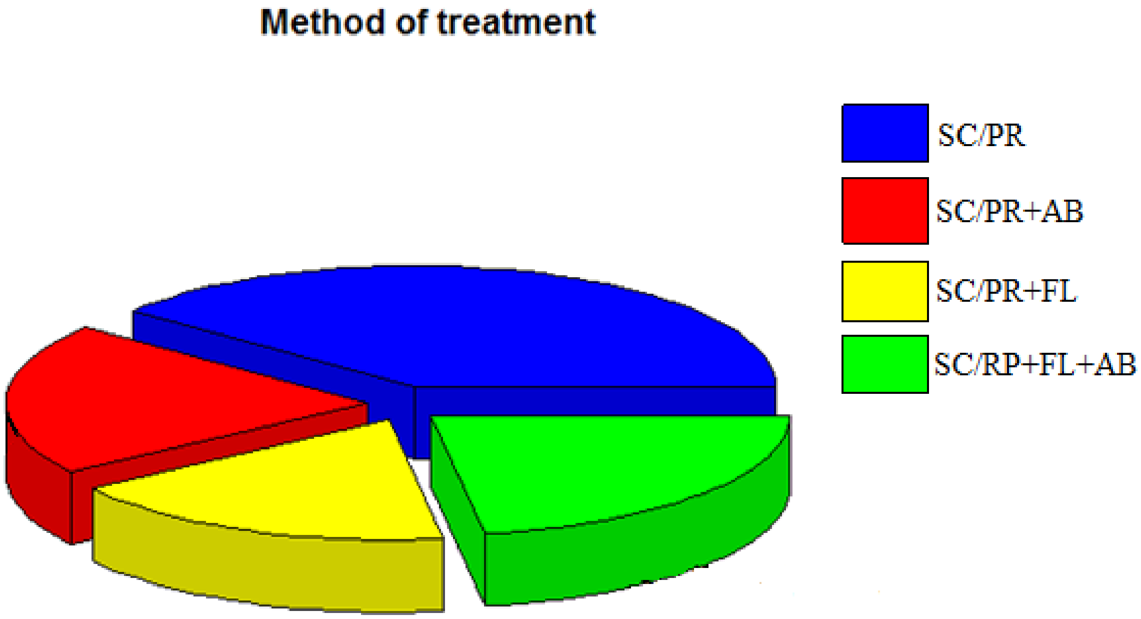

Figure 4.2 : The pie-chart of the therapy the patients received.

The patients were not all measured at the same sets of time points, that is the design is incomplete. (More specifically, in the first tooth we have 25,3% missing values for the clinical index CAL and 12,6% missing values in the second tooth for the same index.) The dependent variable of interest is the level of loss attachment, CAL. Available for other eight independent variables which were related to the response, namely: *time* taking values from 1 to 6 for each annual patient's visit to the periodontist, *id* taking values from 1 to 25 for each patient, *sex* a factor variable taking the value 0 if the patient is male and 1 if the patient is female, *age* denoting the age of the patient, *cigar* denoting the number of cigarettes smoked in a day the patient, *treatment* a factor variable taking the value 0 if the patient received SRP, 1 if the patient received SRP and periodontal surgery (SC/RP+FL), 2 if the patient received SRP and antibiotics (SC/RP+AB) and 3 if the patient received SRP, periodontal surgery and antibiotics (SC/RP+FL+AB). Also, we have the variables *boneloss_beg* and *boneloss_fin* which denoting the percentage of bone loss at the beginning and in the end of the therapy.

4.2 A first approach of the experimental design

An initial step for a researcher when encountering a dataset is to plot the data. By constructing the appropriate plots, one can foresee important information on how to model the data. From the plots, empirical results and some initial tests-estimates for the representation of main effects of factors or the interaction between factors can be obtained.

At first, in *Figure 4.2.1*, a line plot is been presented, in which we have measurements of clinical index CAL of both examined teeth, seperately, over the time.

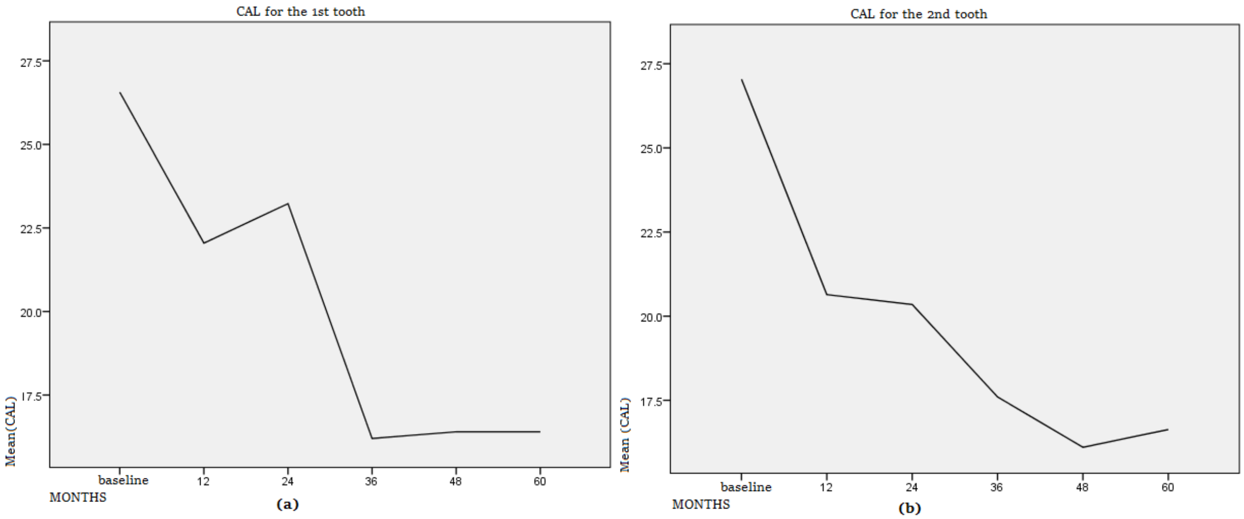

Figure 4.2.1 : (a) line plot of the mean (CAL) over the time for the 1st tooth.
(b) line plot of the mean (CAL) over the time for the 2nd tooth.

The general impression from the figure above is that the CAL index is dropping over the time for both the teeth that we are going to study. Specifically, in *Figure 4.2.1(a)* we can observe that up to 12 months the CAL index is constantly dropping, at 24 months something happens which causes the index to rise, while subsequently at 36 months there is a sharp reduction and up to 60 months we get a relative stability of the index. In *Figure 4.2.1(b*) we can also observe a reduction of the clinical index up to 12 months, a relative stability between 12 and 24 months and then a reduction of the index up to 48 months, with a small increase at the end of the treatment.

There is particular interest in the measurements of the CAL clinical index over the time per treatment group as follows below, separately for each tooth.

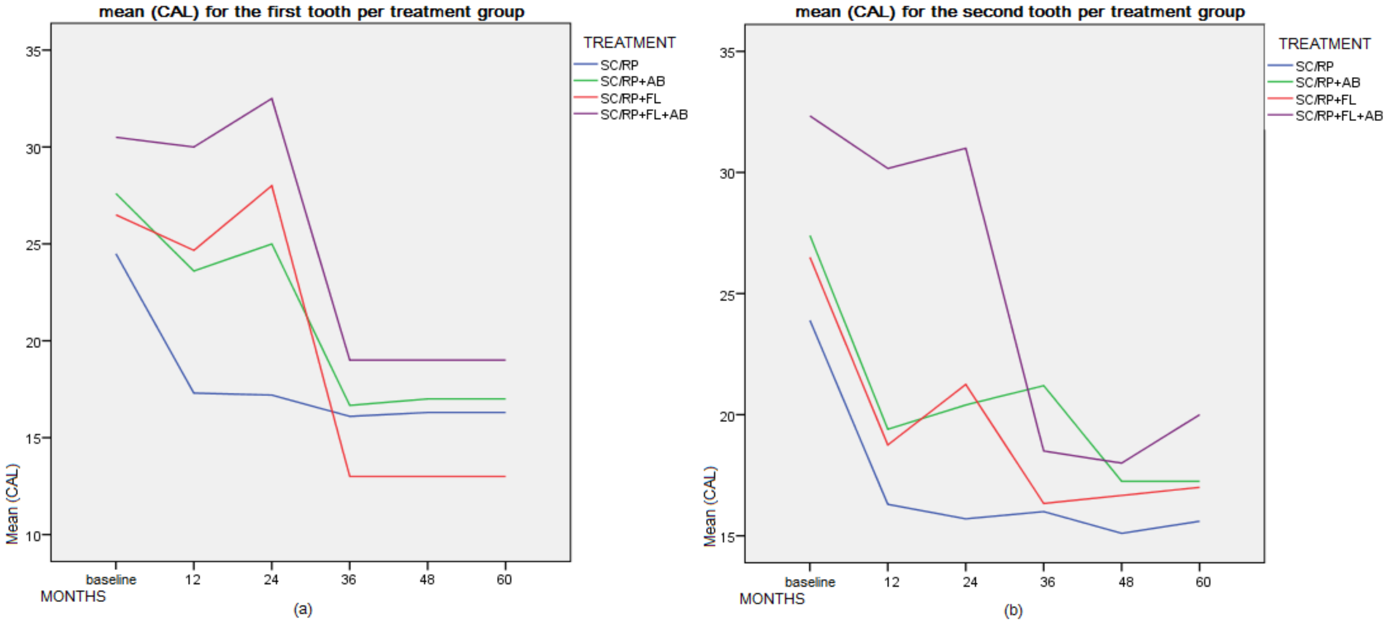

Figure 4.2.2 : (a) line plot of the mean (CAL) over the time, per treatment group, for the 1st tooth.
(b) line plot of the mean (CAL) over the time, per treatment group, for the 2nd tooth.

In *Figure 4.2.2(a)* which concerns the first tooth under study, we can observe that generally in all four treatment groups the mean CAL drops over the time. Specifically, all 4 groups drop up to the 12th month. At the 24th month the group SC/RP presents stability, while all other groups present a relative increase. From the 24th until the 36th month (included) the treatment groups SC/RP+FL, SC/RP+AB, SC/RP+FL+AB present a sharp reduction while the SC/RP has a slight reduction. Until the end of the treatment there is a stability of the CAL index in all four groups. In *Figure 4.2.2(b)* we can observe that the measurement of the CAL index for patients who followed the method SC/RP shows a tendency to drop in relation to time. Patients who followed the method SC/RP+FL are found to present a reduction in the index measurement up to the 12th month, while at the 24th month we can observe an increase which drops until the 36th month and then we can see a gradual increase until the end of treatment. In patients who followed the method SC/RP+AB we find a reduction in the index measurement up to the 12th month, while at the 36th month we can observe an increase which drops until the 48th month and subsequently we see a relative stability up to the 60th month of the measurement. Finally, in patients following the method SC/RP+FL+AB we see that they present a reduction in the index measurement up to the 12th month, while at the 24th month we can observe an increase which drops sharply until the 36th month and subsequently we see an increase from the 48th until the 60th month.

Also worth discussing is the behavior of the CAL index in two sub-groups: men and women. This will be done in continuation with the line plots following for each of the teeth under study which concern us.

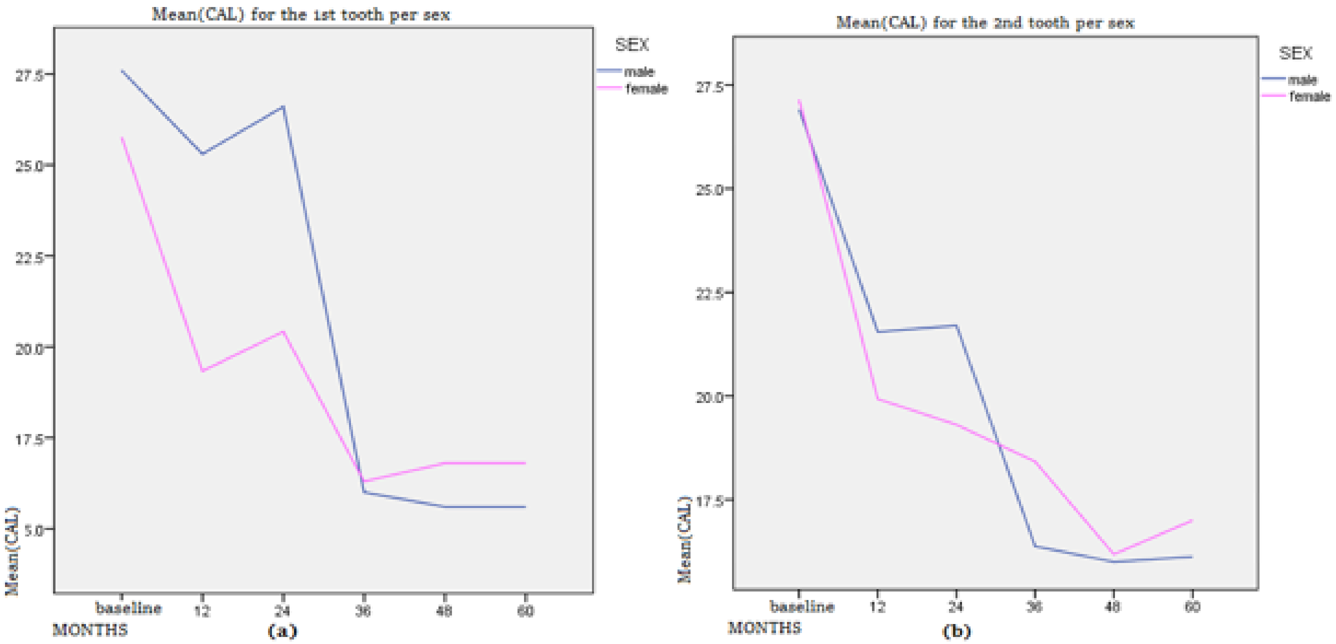

Figure 4.2.3 : (a) line plot of the mean (CAL) over the time, per sex, for the 1st tooth.
(b) line plot of the mean (CAL) over the time, per sex, for the 2nd tooth.

Specifically, in *Figure 4.2.3(a)* we can see that at the beginning of the measurement the CAL index is higher in men in comparison to women. In continuation we can observe that the index drops at the 12th month and increases at the 24th for both men and women. The index then drops sharply and we see that it actually shows almost identical values for both sexes, while subsequently it remains relatively stable until the end of treatment, with the index for women presenting slightly higher values than that for men. In *Figure 4.2.3(b)* we see that the CAL index starts at almost the same value for both sexes and then we have a constant reduction up to the 36th month for men and a constant reduction up to the 48th month for women. In men the index presents a relative stability until the end of treatment, while in women we have a slight increase between the 48th and the 60th month.

We also have at our disposal the measurements of the percentage of bone loss in the tooth at the beginning (*boneloss_beg*) and in the end of the treatment (*boneloss_fin*) for each of the 25 patients. The results have shown that these variables are not affected by sex, as there is no significant statistical difference between the two sub-populations, i.e. between men and women. These differences can easily be graphically represented with the boxplots as is shown indicatively below.

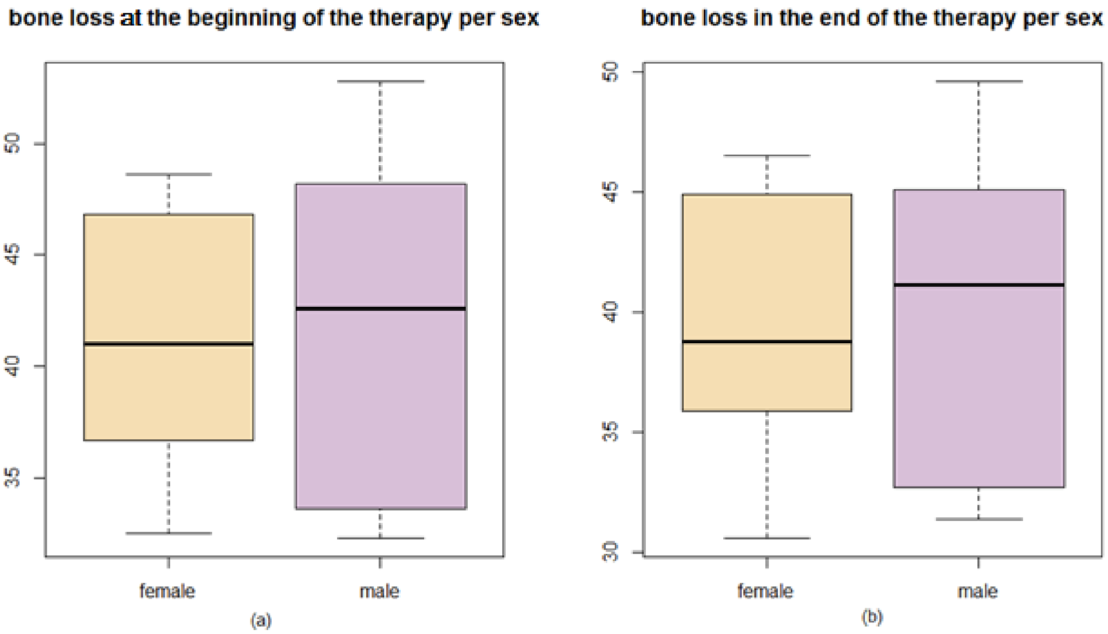

Figure 4.2.4 : Boxplots of the variables “*bone loss at the beginning of the therapy*” and “*bone loss in the end of the therapy”* per sex.

The chart in *Figure 4.2.4* above shows that the differences between men and women in the variables concerning the percentage of bone loss in the tooth at the beginning and in the end of the study are too small and therefore statistically insignificant.

4.3 The general form of the model

As we said in the Second Chapter, the linear mixed effects model is defined as

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{u_i} + \varepsilon_i, \qquad \text{for } i = 1,\ldots,\text{m} \qquad (4.1)$$

where $\boldsymbol{Y_i}$ is a vector of responses of continuous responses for the i-th subject defined by

$$\boldsymbol{Y_{ij}} = \begin{bmatrix} Y_{1i} \\ Y_{2i} \\ \vdots \\ Y_{n_i i} \end{bmatrix}.$$

Note that n_i is dependent on i, hence the number of observations for each subject may differ. We have m subject, in total $n = \sum_i^m n_i$ observations.

The fixed effect design matrix, $\boldsymbol{X_i}$, is a $n_i \times p$ matrix, which represents p covariates corresponding to the fixed effects for each observation of the i-th subject. The fixed effect design matrix is defined as

$$\boldsymbol{X_i} = \begin{bmatrix} x_{1i}^{(1)} & x_{1i}^{(2)} & \cdots & x_{1i}^{(p)} \\ x_{2i}^{(1)} & x_{2i}^{(2)} & & x_{2i}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_i i}^{(1)} & x_{n_i i}^{(2)} & \cdots & x_{n_i i}^{(p)} \end{bmatrix}.$$

The first column of the design matrix is often equal to 1 for all observations to include an intercept term in the model.

The fixed effects matrix, $\boldsymbol{\beta}$, is a vector consisting of p unknown regression coefficients associated with the covariates from the design matrix $\boldsymbol{X_i}$, and is defined as

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

The random effect design matrix, $\boldsymbol{Z_i}$, is a $n_i \times q$ matrix, which represents q covariates corresponding to the random effects for each observation of the i-th subject. The random effect design matrix is defined as

$$\boldsymbol{Z_i} = \begin{bmatrix} z_{1i}^{(1)} & z_{1i}^{(2)} & \cdots & z_{1i}^{(q)} \\ z_{2i}^{(1)} & z_{2i}^{(2)} & & z_{2i}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n_i i}^{(1)} & z_{n_i i}^{(2)} & \cdots & z_{n_i i}^{(q)} \end{bmatrix}.$$

The random effects are effects that vary randomly across subjects. Hence, it includes the individual differences for the subjects. Covariates with random effect are often represented both in the $\boldsymbol{X_i}$ matrix and the $\boldsymbol{Z_i}$ matrix. In the simplest example of the linear mixed effects model, only the intercepts are assumed to vary randomly from subject to subject. Hence, in this case the $\boldsymbol{Z_i}$ matrix is simply reduced to a vector of n_i 1's.

The random effect vector, $\boldsymbol{u_i}$, is a vector consisting of q random effects associated with the covariates from the design matrix $\boldsymbol{Z_i}$, and is defined by

$$\boldsymbol{u_i} = \begin{bmatrix} u_{1i} \\ \vdots \\ u_{qi} \end{bmatrix}.$$

We assume that the random effect vector, $\boldsymbol{u_i}$, follows a multivariate normal distribution,

$$\boldsymbol{u_i} \sim N_q\,(\mathbf{0}, \boldsymbol{D}),$$

where the positive definite symmetric covariance matrix $\boldsymbol{D}$ is defined as

$$\boldsymbol{D} = var(\boldsymbol{u_i}) = \begin{bmatrix} var(u_{1i}) & cov(u_{1i,}u_{2i}) & \cdots & cov(u_{1i,}u_{qi}) \\ cov(u_{1i,}u_{2i}) & var(u_{2i}) & & cov(u_{2i,}u_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(u_{1i,}u_{qi}) & cov(u_{2i,}u_{qi}) & \cdots & var(u_{qi}) \end{bmatrix}. \tag{4.2}$$

Finally, the residual $\boldsymbol{\varepsilon_i}$ vector is defined by

$$\boldsymbol{\varepsilon_i} = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{n_i i} \end{bmatrix}$$

where each element represents the residual associated with each response for the *i*-th subject. Unlike the residuals in standard linear models, the residuals associated with repeated observations on the same subject in a linear mixed effects model can be correlated. We assume that the n_i residuals in the $\boldsymbol{\varepsilon_i}$ vector follow a multivariate normal distribution,

$$\boldsymbol{\varepsilon_i} \sim N_{n_i}(\mathbf{0}, \boldsymbol{R_i}),$$

where the positive definite symmetric covariance matrix $\boldsymbol{R_i}$ is defined as

$$\boldsymbol{R_i} = var(\boldsymbol{\varepsilon_i}) = \begin{bmatrix} var(\varepsilon_{1i}) & cov(\varepsilon_{1i,}\varepsilon_{2i}) & \cdots & cov(\varepsilon_{1i,}\varepsilon_{n_i i}) \\ cov(\varepsilon_{1i,}\varepsilon_{2i}) & var(\varepsilon_{2i}) & & cov(\varepsilon_{2i,}\varepsilon_{n_i i}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\varepsilon_{1i,}\varepsilon_{n_i i}) & cov(\varepsilon_{2i,}\varepsilon_{n_i i}) & \cdots & var(\varepsilon_{n_i i}) \end{bmatrix}.$$

We assume that the vectors of residuals, $\varepsilon_1, \dots, \varepsilon_m$ and the random effects $u_1, \dots, u_m$, are independent of each other.

4.4 Selecting the best Covariance Structure

There are many methods for choosing the most appropriate structure for the covariance matrix of the data. Models with the same fixed effects, but with different covariance structures, can be compared using again statistics based on the likelihood function. In the case of non-nested models, covariance structures can be compared using the Bayesian Information Criterion (BIC) (Schwarz, 1978) and the Akaike Information Criterion (AIC) (Sakamoto, Ishiguro and Kitagawa, 1986).

4.4.1 Bayesian Information Criterion

The Bayesian Information Criterion (BIC) (Schwarz, 1978), which is sometimes called Schwarz's Bayesian Criterion (SBC) is a statistic based on the likelihood function and it is given by:

$$BIC = -2logL + n_{par}logN = -2l\left(\frac{\hat{\theta}}{y}\right) + n_{par}logN$$

where, n_{par} is the number of the parameters in the model and N is the total number of the observations used to fit the model.

If BIC is used to compare two or more models for the same data, the model with the lowest BIC is more preferable.

4.4.2 Akaike Information Criterion

The Akaike Information Criterion (AIC) (Sakamoto, Ishiguro and Kitagawa, 1986) is computed with:

$$AIC = -2logL + 2n_{par} = -2l\left(\frac{\hat{\theta}}{y}\right) + 2n_{par}$$

where, n_{par} is the number of the parameters in the model.

Similarly, when using the AIC the model with the lowest AIC is more preferable.

4.5 Model structure

The purpose of this section is to create the structure of a suitable model and then to select the best mixed effect model based on the existing dataset, i.e. with the missing values in the measurement of the clinical CAL index, for each of the two teeth under study. In continuation, in the next chapter, we will do a multiple imputation in the dataset in order to predict the missing values and to obtain a full observed dataset. Then we will find out whether the model we detected within the scope of the present chapter is still the best choice, always based on the BIC criterion.

Analysis will be done using the statistical package SPSS. We will have to find two suitable models, one for each tooth under study. The Maximum Likelihood (ML) estimation method will be followed. In both cases the models to be studied will have the same structure, so it is sufficient to describe initially the procedure for the 1 tooth.

The variable CAL concerns the measurement of the clinical index of the two teeth over the time, and it will be the dependent variable of our model. We will use the explanatory variables *sex* and *treatment* as factors and the variables *time, age, cigar* and *boneloss_beg* as covariates and the interaction between the variables *treatment* and *time*. The *fixed part* of the model will essentially consist of:

$$\beta_0 + \beta_1\, sex_i + \beta_2\, treatment_i + \beta_3\, time_{ij} + \beta_4\, age_i + \beta_5\, cigar_i + \beta_6\, boneloss\_beg_i +$$
$$+ \beta_7\, time_{ij} * treatment_i\,, \qquad \text{for} \quad i = 1,\ldots,25 \text{ patients and } \; j = 1,\ldots,6 \text{ measures.}$$

To begin with, we will assign a random constant b_{oi} to each subject. The individual characteristic of each subject and that which distinguishes it from another lies in the different individual constant. The model will finally have the following form:

$$CAL_{ij} = fixed + \ b_{0i} + \ \varepsilon_{ij} \qquad (4.3)$$

where $b_{0i} = \ \beta_0 + \ u_{oi}$, $b_{0i} \sim N(0, D)$

The covariance matrix D of the model (4.3) has variance components structure. The value of the BIC criterion for the 1st tooth is 649.433 while for the 2nd tooth it is 782.936.

In continuation, we will use a model where we will assign a random constant and random slope for each subject over the time. Therefore, in the new model we will use, the *random part* will include random intercept b_{0i} and random slope b_{1i} in the variable of time (*time*) for each *i*-subject, where

$$b_{0i} = \ \beta_0 + \ u_{oi} \ , b_{0i} \sim N(0, D)$$

and

$$b_{1i} = \ \beta_1 + \ u_{1i} \ , b_{1i} \sim N(0, D)$$

where D is the covariance matrix. Therefore, the final form of the model we shall subsequently use is

$$CAL_{ij} = fixed + \ u_{0i} + \ u_{1i} time_{ij} + \ \varepsilon_{ij} \qquad (4.4)$$

this means that for the model (4.4) we assume that each patient has his own random intercept and his own slope over the time, with different variables for the two parameters.

In the first case we assume that in the model (4.4) the table D is unstructured, i.e. the model has a random slope and random constant which changes for each patient and the unstructured D matrix is given as

$$D = \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_0 u_1} \\ \sigma_{u_0 u_1} & \sigma_{u_1}^2 \end{bmatrix}.$$

In this case the BIC criterion has value 662.172 for the 1st tooth while for the 2nd tooth is 803.126.

In continuation, with the (4.4) model, we select the covariance matrix D to have a diagonal structure. The diagonal matrix D is given as

$$D = \begin{bmatrix} \sigma_{u_0}^2 & 0 \\ 0 & \sigma_{u_1}^2 \end{bmatrix}.$$

In this case the BIC criterion is 654.151 for the 1st tooth and 787.812 for the 2nd tooth.

Finally, in the same model we try out the covariance matrix D with a scaled identity structure. That means that the random slope and the random constant have the same changes in every patient and the scaled identity matrix D is given as

$$D = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}.$$

The BIC criterion for the 1st tooth is 689.403 while for the 2nd tooth is 846.666.

In continuation, we create a new model which does not include a random part; therefore we have error modelling. This model is as follows

$$CAL_{ij} = fixed + \varepsilon_{ij} \tag{4.5}$$

For model (3.5) and for the positive definite symmetric covariance matrix R we used the diagonal structure. The diagonal structure of the covariance matrix R is the simplest structure, in which the residuals within one subject are assumed to be uncorrelated and have equal variances. Hence, the diagonal structure of the covariance matrix R, is given as

$$R = var(\boldsymbol{\varepsilon_{ij}}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}.$$

In this case the BIC criterion is 707.493 for the 1st tooth while for the 2nd is 829.670.

Then we select the structure of the covariance matrix R as compound symmetry, this assumes that the residuals within one subject have a constant covariance σ_1 and a constant variance, $\sigma^2 + \sigma_1$. Hence, the compound symmetry structure of the covariance matrix R, is given as

$$R = var(\boldsymbol{\varepsilon_{ij}}) = \begin{bmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_1 & \cdots & \sigma^2 + \sigma_1 \end{bmatrix}.$$

In this case the BIC criterion is 651.628 for the 1st tooth and 835.550 for the 2nd tooth.

Finally, we select the structure of the covariance matrix R as AR (1). The covariance matrix R is given as

$$R = var(\boldsymbol{\varepsilon_{ij}}) = \begin{bmatrix} \sigma^2 & \sigma_{\varepsilon_1\varepsilon_2} & \cdots & \sigma_{\varepsilon_1\varepsilon_6} \\ \sigma_{\varepsilon_2\varepsilon_1} & \sigma^2 & & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\varepsilon_6\varepsilon_1} & \sigma_{\varepsilon_6\varepsilon_2} & \cdots & \sigma^2 \end{bmatrix}.$$

The BIC criterion has the value 689.139 for the 1st tooth while for the 2nd tooth is 812.280.

4.6 Selection of the final models

Based on the above models examined in the section 4.5 we conclude that for both teeth the best model is (4.3) which has only random intercept as it presents the lowest BIC value in both cases. The results obtained from the models regarding the dependent variable Clinical attachment level (CAL), which concerns the measurements of the index, are presented in the following tables and then the parameter values are interpreted.

4.6.1 Explanation of parameters for the first tooth

According to the following *Table 4.6.1* about the 1st tooth, the value 25.652 shows the expected value of the CAL variable for female patients who received the therapy SC/RP+FL+AB, have the same time measurement, the same age, the same number smoked cigarettes and the same rate of bone loss at the begin of the therapy.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	23.652685	19.846789	22.929	2.466	.022	5.539911	63.259054
[sex=0]	-3.684874	1.658355	23.489	3.399	.002	2.285836	9.375984
[sex=1]	0[b]	0	.	.	.	.	.
[treatment=0]	-8.522886	2.832954	30.379	-2.944	.006	-12.686823	-2.296522
[treatment=1]	-4.181818	3.017255	31.376	-1.428	.163	-10.665661	1.877277
[treatment=2]	-3.519086	3.736225	29.996	-2.174	.038	-12.153626	-.379070
[treatment=3]	0[b]	0	.	.	.	.	.
time	-1.096786	.068223	93.375	-.672	.503	-.115036	.056860
age	.525708	.526988	23.879	-.407	.687	-.877665	.588483
cigar	.023587	.089584	23.050	1.832	.080	-.011562	.190461
boneloss_beg	.239858	1.499608	23.737	4.404	.000	2.708941	7.493208
[treatment=0] * time	.369875	.072036	92.898	-1.696	.093	-.169963	.013377
[treatment=1] * time	-.462842	.069258	96.561	-.302	.763	-.141679	.104207
[treatment=2] * time	-.425821	.043582	93.709	-1.134	.260	-.160506	.043813
[treatment=3] * time	0[b]	0	.	.	.	.	.

Table 4.6.1 : Coefficients of the final model for the 1st tooth.

The average difference of the dependent CAL variable between the two sexes is -3.68, with the other explanatory variables held constant. The average difference of the CAL variable between patients received the therapy SC/RP+FL+AB and the therapy SC/RP and the other explanatory variables held constant, is -8.52. Keeping constant the explanatory variables relating to age, gender, number of cigarettes and the bone loss in the begin of therapy, apply the following. The average difference between patients taking SC/RP+FL+AB and SC/RP+FL is -4.18 and the average difference between patients taking SC/RP+FL+AB and SC/RP+AB is -3.51. The rate of change of the CAL variable at the reference level, which is the patients taking SC/RP+FL+AB, is -1.096. The value 0.36 indicates the difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP. Specifically, the rate of change for the patients taking SC/RP is -1.096 + 0.36 = -0.736, i.e. the CAL variable drops faster in patients taking SC/RP+FL+AB. The difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP+FL is given by value -0.46. Therefore, the rate of change for patients taking SC/RP+FL is -1.556, which means that it drops at a faster rate in patients taking SC/RP+FL+AB. Also, value -0.42 indicates the difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP+AB. The rate of change for patients taking SC/RP+AB is given by value -1.516 and in this case the CAL variable is shown to be dropping faster in patients forming the reference level who take SC/RP+FL+AB.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		6.488136	.970643	6.684	.000	4.839248	8.698854
Intercept [subject = ID]	Variance	11.870547	3.936956	3.015	.003	6.196742	22.739353

Table 4.6.2 : Estimates of variances for the 1st tooth.

The above *Table 4.6.2* contains the estimates of variances. We note that the variance between subjects is greater than that within subjects. This leads us to conclude that there is a moderate correlation between our observations. This can be calculated using the following formula:

$$Cor\left(Y_{ij}, Y_{ij'}\right) = \frac{{\sigma_{n_0}}^2}{{\sigma_{n_0}}^2 + {\sigma_{n_\varepsilon}}^2} = \frac{11.870547}{11.870547+6.488136} = 0.6465 \approx 65\%$$

As we can therefore see, the correlation between the observations is 65%. It is an average correlation.

4.6.2 Explanation of parameters for the second tooth

The *Table 4.6.3* which follows gives us the values of the parameters from the model analysis (3.3) for the 2nd tooth under study. The value 29.813 indicates the expected value of the CAL variable for female patients taking SC/RP+FL+AB, have the same time measurement, the same age, the same number smoked cigarettes and the same rate of bone loss at the begin of the therapy. The average difference of the dependent CAL variable between the two sexes is -0.249, with the other explanatory variables held constant.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	23.652685	19.846789	23.432	1.674	.107	-7.887156	56.614589
[sex=0]	-3.684874	1.658355	23.673	-.111	.913	-4.256683	5.638469
[sex=1]	0[b]	0	.	.	.	.	.
[treatment=0]	-8.522886	2.832954	27.666	-3.048	.005	-16.402346	-3.211830
[treatment=1]	-4.181818	3.017255	28.393	-2.685	.012	-17.118663	-2.308229
[treatment=2]	-3.519086	3.736225	27.756	-1.936	.063	-14.290509	.407300
[treatment=3]	0[b]	0	.	.	.	.	.
time	-1.096786	.068223	110.480	-1.355	.178	-.121072	.022722
age	.525708	.526988	23.209	-.357	.724	-1.113981	.785999
cigar	.023587	.089584	23.643	-.446	.659	-.164204	.105841
boneloss_beg	.239858	1.499608	23.963	1.437	.164	-.939631	5.250953
[treatment=0] * time	.369875	.072036	109.338	-1.426	.157	-.137410	.022427
[treatment=1] * time	-.462842	.069258	108.339	-.490	.625	-.118710	.071660
[treatment=2] * time	-.465821	.043582	108.328	-.778	.438	-.124631	.054380
[treatment=3] * time	0[b]	0	.	.	.	.	.

Table 4.6.3 : Coefficients of the final model for the 2nd tooth.

The average difference of the CAL variable between patients received the therapy SC/RP+FL+AB and the therapy SC/RP and the other explanatory variables held constant, is -9.807. Keeping constant the explanatory variables relating to age, gender, number of cigarettes and the bone loss in the begin of therapy, apply the following. The average difference between patients taking SC/RP+FL+AB and SC/RP+FL is -9.713 and the average difference between patients taking SC/RP+FL+AB and SC/RP+AB is -6.941. The rate of change of the CAL variable in the reference level which consists of patients taking SC/RP+FL+AB, is -0.049. Value -0.057 indicates the difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP. Specifically, the rate of change in patients taking SC/RP is -0.049 + (-0.057) = -0.106, which means that the CAL variable drops faster in patients taking SC/RP+FL+AB. The difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP+FL is given by value -0.023. That means that the rate of change in patients taking SC/RP+FL is -0.072, which means that it drops faster in patients taking SC/RP+FL+AB. Also, value -0.035 indicates the difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP+AB. The rate of change in patients taking SC/RP+AB is given by value -0.084. In this case we also see that the CAL variable drops faster in patients forming the reference level and taking SC/RP+FL+AB.

Finally, *Table 4.6.4* below, contains the estimates of covariances for the 2nd tooth. We note that the variance between subjects is greater than the variance within subjects in this case, too.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		7.804872	1.080459	7.224	.000	5.950187	10.237667
Intercept [subject = ID]	Variance	22.982563	7.211197	3.187	.001	12.425632	42.508758

Table 4.6.4 : Estimates of covariances for the 2nd tooth.

We calculate once more the correlation between observations using the formula:

$$Cor\left(Y_{ij}, Y_{ij'}\right) = \frac{{\sigma_{n_0}}^2}{{\sigma_{n_0}}^2 + {\sigma_{n_\varepsilon}}^2} = \frac{22.982563}{22.982563+7.804872} = 0.7464 \approx 75\%$$

As we can therefore see, the correlation is 75%, which means that in this case we have strong correlation between observations.

Chapter 5

5.1 Introduction

In this chapter we will do the multiple imputation of our data in order to acquire a fully observed dataset. As mentioned in the previous chapter we must evaluate the level of the clinical attachment loss (CAL) of 25 patients. Two teeth in every patient's mouth have been selected in which we have measured the CAL index at different points in time. As the patients were not all measured at the same sets of time points out design is incomplete. More specifically, in the measure of the CAL index for the first tooth we have 25.3% missing values and 12.6% missing values in the second tooth. As we can see in the following figure (*Figure 5.1*), missing values are presented in black, while observed values are presented in violet color. We can also observe that only two variables present missing values. These two variables concern the measurement of the CAL index in the first and in the second tooth of every patient.

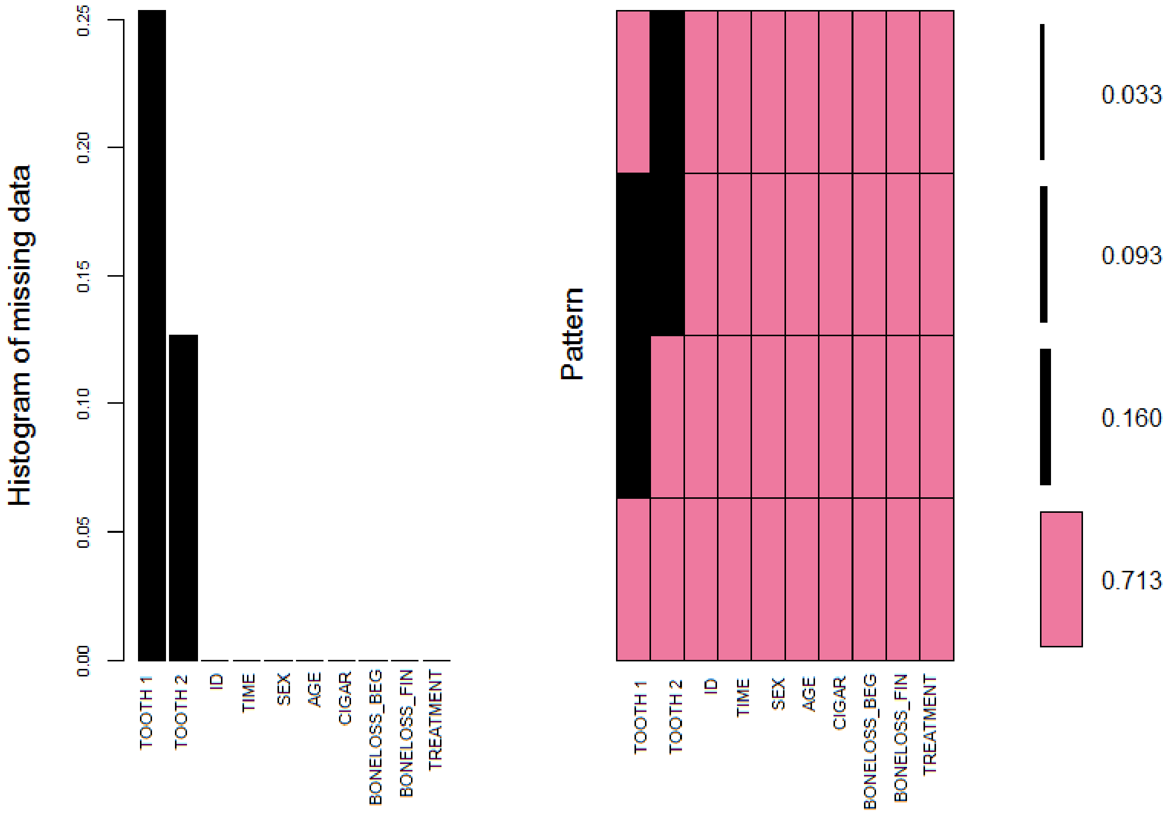

Figure 5.1 : Visual representation of missing values in the variables of the dataset.

5.2 An Approach to Multiple Imputation with the “mice” algorithm

The multiple imputation of data will be done using the programming language R. Specifically, the “mice” package will be used. In the following we provide a brief description of the way the algorithm is generally used in data.

The process can be broken down into four general steps:

- *Step 1*: A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as “place holders.”
- *Step 2*: The “place holder” mean imputations for one variable (“var”) are set back to missing.
- *Step 3*: The observed values from the variable “var” in *Step 2* are regressed on the other variables in the imputation model, which may or may not consist of all of the variables in the dataset. In other words, “var” is the dependent variable in a regression model and all the other variables are independent variables in the regression model. These regression models operate under the same assumptions that one would make when performing (e.g.,) linear, logistic, or Poison regression models outside of the context of imputing missing data.
- *Step 4*: The missing values for “var” are then replaced with predictions (imputations) from the regression model. When “var” is subsequently used as an independent variable in the regression models for other variables, both the observed and these imputed values will be used.
- *Step 5*: *Steps 2–4* are then repeated for each variable that has missing data. The cycling through each of the variables constitutes one iteration or “cycle.” At the end of one cycle all of the missing values have been replaced with predictions from regressions that reflect the relationships observed in the data.

- *Step 6*: *Steps 2* through 4 are repeated for a number of cycles, with the imputations being updated at each cycle. The number of cycles to be performed can be specified by the researcher. At the end of these cycles the final imputations are retained, resulting in one imputed dataset. Generally, ten cycles are performed (Raghunathan, 2002) however, research is needed to identify the optimal number of cycles when imputing data under different conditions.

5.3 Descriptive measures of the imputed variables

In this section, since the multiple imputation in the dataset has been done, we shall present some descriptive measures. In the following *Table 5.3.1* we present the mean value of the clinical CAL index for both teeth under study and for every patient and also the treatment method followed by each one.

Patient's id	Treatment method	mean (CAL) tooth 1	mean (CAL) tooth 2
1	SC/RP	19.666	17
2	SC/RP+FL	14	15.833
3	SC/RP+FL+AB	33.206	31.695
4	SC/RP+FL+AB	25.795	17.833
5	SC/RP+FL+AB	20.666	26.284
6	SC/RP	19.333	17.333
7	SC/RP+AB	25.937	29.246
8	SC/RP	19.333	17
9	SC/RP	19.333	18.166
10	SC/RP+FL	25.805	26.496
11	SC/RP+FL	16.738	18
12	SC/RP+AB	22.166	20.833
13	SC/RP	14	15.833
14	SC/RP+FL+AB	31.679	30.172
15	SC/RP	19.833	17.666
16	SC/RP	14	15.833
17	SC/RP+AB	19.333	18.166
18	SC/RP+FL+AB	26.057	22
19	SC/RP+AB	27.088	17.666
20	SC/RP	14	15.833
21	SC/RP	20	18.166
22	SC/RP	20	18.166
23	SC/RP+FL+AB	31.116	30.561
24	SC/RP+FL	28.051	18.833
25	SC/RP+AB	14	17.166

Table 5.3.1 : Mean (CAL) of every patient for both teeth.

In *Figure 5.3.1* below we can easily discern the profile of the measurements for every patient over the time for both teeth per treatment group. We can observe certain differences between treatment groups. In the group which followed the treatment SC/RP it is evident that the clinical CAL index has dropped over time. On the other hand, in patients who followed the other treatments we see that approximately in the middle of treatment some increases appeared in the CAL index in both teeth, which then decreased towards the end of treatment.

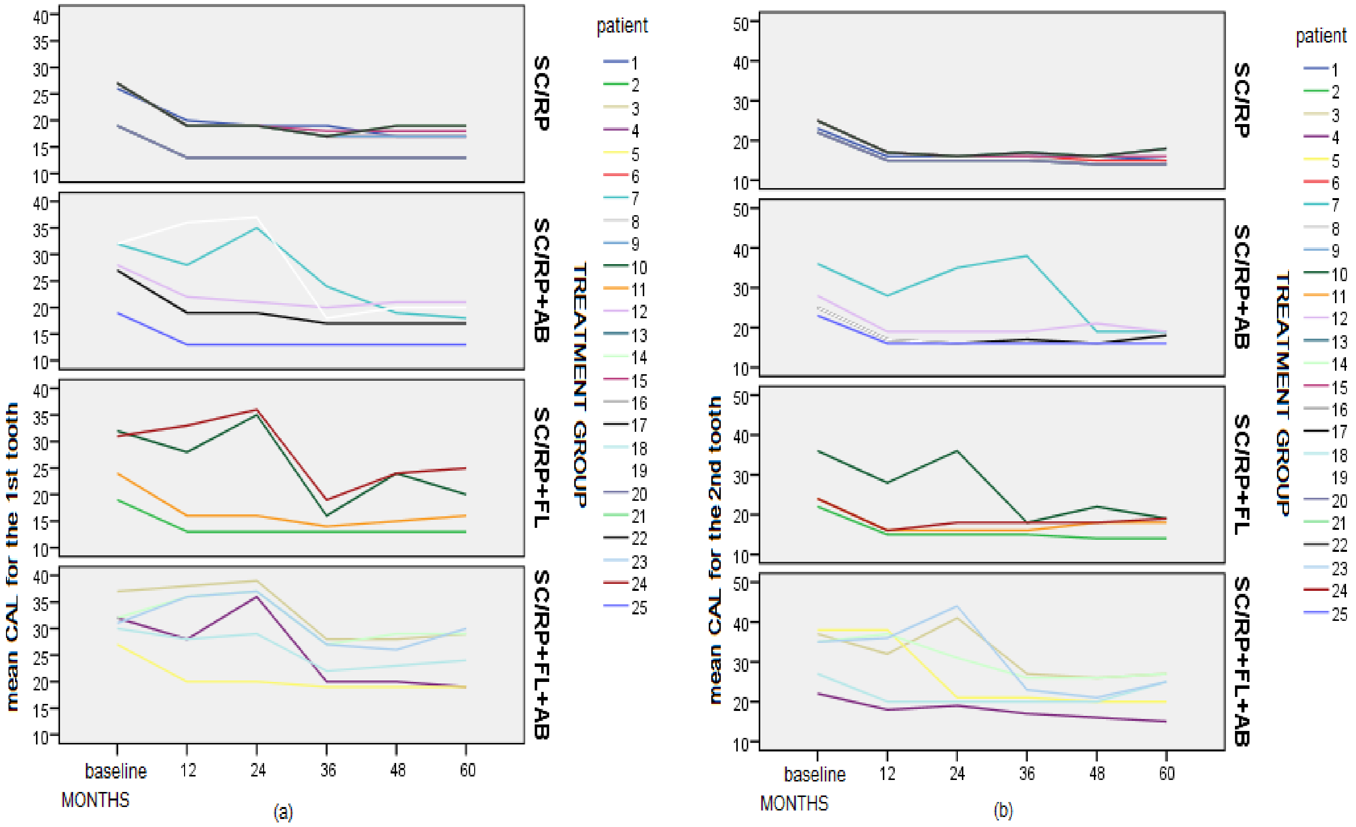

Figure 5.3.1 : Profile of the measurements of imputed values of the variable CAL over time, for the 25 patients per treatment group (a) for the first tooth and (b) for the second tooth.

Differences between treatment groups can be easily represented in a chart with the error bars in following *Figure 5.3.2*. We can observe that, regarding the measurement of the CAL index in both the 1st and the 2nd tooth, the only difference observed between treatment groups regards the patients taking the treatments SC/RP and SC/RP+FL+AB. These differences are in fact present in every measurement, as is shown in the figure, as these two treatment groups do not present any common point in their corresponding confidence intervals. In contrast, we observe that for the treatment groups SC/RP+AB, SC/RP+FL the respective confidence intervals are very close and almost identical.

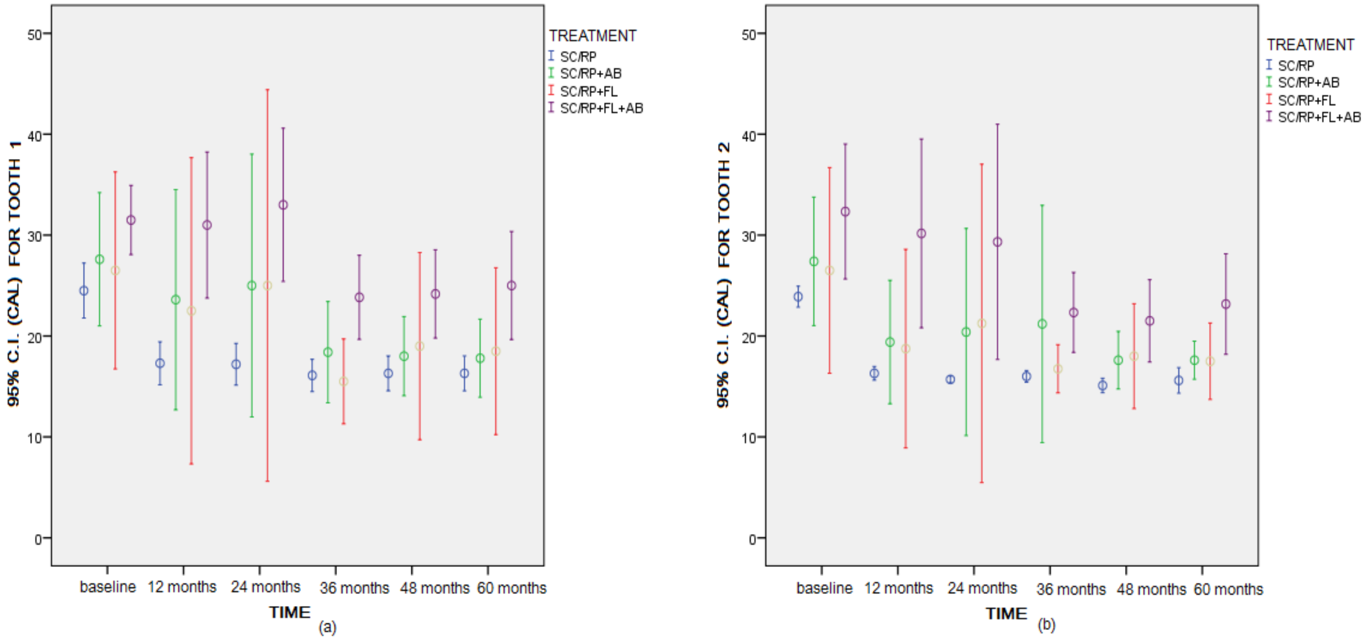

Figure 5.3.2 : (a) Error bars for the confidence intervals of the clinical index CAL for the first tooth.
(b) Error bars for the confidence intervals of the clinical index CAL for the second tooth.

Subsequently, in the following *Table 5.3.2* and *Table 5.3.3* are given the mean and the standard deviation of the clinical index CAL, for every treatment group, at each measurement separately for the two teeth respectively. It is obvious that for all treatment groups, the mean value of the index is reduced over time. Also we observe the difference between the patients receiving the treatments SC/RP and SC/RP+FL+AB as we have seen and graphics from the above error bars. In all measurements, the mean value of the index CAL is quite less for the treatment method SC/RP compared with the SC/RP+FL+AB. Further observation shows that the other two treatment groups have slight variations between their mean values for all the measurements, for both teeth.

		Baseline	12 months	24 months	36 months	48 months	60 months
SC/RP	mean	24.50	17.30	17.20	16.10	16.30	16.30
	sd	3.80	2.98	2.89	2.23	2.40	2.40
SC/RP+AB	mean	27.60	23.60	25.0	18.45	17.86	17.71
	sd	5.31	8.79	10.48	4.10	3.09	3.06
SC/RP+FL	mean	26.50	22.59	25.03	15.56	18.81	18.38
	sd	6.13	9.45	12.15	2.89	5.80	5.26
SC/RP+FL+AB	mean	31.51	30.92	33.03	23.67	24.28	25.09
	sd	3.34	6.93	7.26	3.87	4.12	5.21

Table 5.3.2 : Mean and standard deviation of clinical indicator CAL for the 1st tooth with the imputed values.

		Baseline	12 months	24 months	36 months	48 months	60 months
SC/RP	mean	23.90	16.30	15.70	16.00	15.10	15.60
	sd	1.44	0.94	0.48	0.81	0.99	1.77
SC/RP+AB	mean	27.40	19.40	20.40	21.20	17.66	17.63
	sd	5.12	4.92	8.26	9.47	2.35	1.55
SC/RP+FL	mean	26.50	18.75	21.25	16.75	17.94	17.58
	sd	6.40	6.18	6.91	1.45	3.17	2.46
SC/RP+FL+AB	mean	32.33	30.16	29.25	22.29	21.43	23.06
	sd	6.37	8.90	11.11	3.77	3.94	4.62

Table 5.3.3 : Mean and standard deviation of clinical indicator CAL for the 2nd tooth with the imputed values.

5.4 Model choice for the imputed dataset

In this section we will fit the models we created in the previous chapter in the new imputed dataset, with the purpose to find out the most suitable model, always following the BIC criterion. In this case, too, analysis will be done using the statistical package SPSS and assessment will be done following the method ML.

First we will apply the model (4.3) which has only random intercept for every patient, with the covariance matrix D presenting a variance components structure. The value that BIC criterion gives for the first tooth is 926.665, while for the second tooth it is 957.558.

Then we will apply the model (4.4) which includes random intercept and random slope over the time, where the covariance matrix D is unstructured. The value given by BIC for the first tooth is 904.907, while for the second tooth it is 934.916. Continuing with the same model, but changing the structure of the D matrix to diagonal, the BIC criterion gives the value 902.774 for the first tooth and 932.038 for the second tooth. And finally, applying the same model, with table D presenting a scaled identity structure, the result we obtain from the BIC criterion has the value 927.253 for the 1st tooth and 958.662 for the 2nd tooth.

Finally, we try out in the imputed dataset the model (4.5) which does not include random part. Initially we suppose that the covariance matrix has diagonal structure. The BIC criterion for the first tooth has value 917.145 and for the second tooth the value is 930.609. Then we suppose that the covariance table has structure compound symmetry. The value of the BIC criterion for the first tooth is 897.764 and for the second tooth 927.028. Finally, in the same model (4.5), we select the structure of the covariance matrix R as AR(1). In this case, the BIC is 905.152 for the first tooth and 924.569 for the second.

After applying the total of the models we created in the Chapter 3 in the imputed data, we conclude that the suitable model for the 1st tooth is model (4.5) with a compound symmetry covariance matrix. The most suitable model for the second tooth is again (4.5), but with a structure of AR(1).

5.4.1 Explanation of parameters of the final model for the first tooth

The appropriate model for the imputed data relating to the first tooth as mentioned above is the model (4.5) with compound symmetry covariance matrix. Subsequently, follows the *Table 5.4.1* that contains the coefficients of the final model of the first tooth.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	18.762427	13.261213	25.177	1.415	.169	-8.539811	46.064665
[sex=0]	2.862232	1.629180	25.000	1.757	.091	-.493128	6.217591
[sex=1]	0[b]	0	.	.	.	.	.
[treatment=0]	-8.899911	2.570222	34.401	-3.463	.001	-14.120990	-3.678831
[treatment=1]	-4.345528	2.943655	36.730	-1.476	.148	-10.311418	1.620363
[treatment=2]	-3.584266	2.918427	35.187	-1.228	.228	-9.507860	2.339328
[treatment=3]	0[b]	0	.	.	.	.	.
time	-1.582496	.314962	125.000	-5.024	.000	-2.205846	-.959147
age	.130352	.336466	25.000	.387	.702	-.562613	.823317
cigar	.016604	.049803	25.000	.333	.742	-.085967	.119175
boneloss_beg	.145928	.151545	25.000	.963	.345	-.166185	.458040
[treatment=0] * time	.293925	.398399	125.000	.738	.462	-.494557	1.082407
[treatment=1] * time	-.354667	.497999	125.000	-.712	.478	-1.340270	.630935
[treatment=2] * time	-.385298	.467165	125.000	-.825	.411	-1.309875	.539279
[treatment=3] * time	0[b]	0	.	.	.	.	.

Table 5.4.1 : Coefficients of the final model for the 1st tooth, of the imputed dataset.

According to the above *Table 5.4.1* about the 1st tooth, the value 18.762 shows the expected value of the CAL variable for female patients who received the therapy SC/RP+FL+AB, have the same time measurement, the same age, the same number smoked cigarettes and the same rate of bone loss at the begin of the therapy. The average difference of the dependent CAL variable between the two sexes is 2.86, with the other explanatory variables held constant. The average difference of the CAL variable between patients received the therapy SC/RP+FL+AB and the therapy SC/RP and the other explanatory variables held constant, is -8.89. Keeping constant the explanatory variables relating to age, gender, number of cigarettes and the bone loss in the begin of therapy, apply the following. The average difference between patients taking SC/RP+FL+AB and SC/RP+FL is -4.34 and the average difference between patients taking SC/RP+FL+AB and SC/RP+AB is -3.58. The rate of change of the CAL variable at the reference level, which is the patients taking SC/RP+FL+AB, is -1.582. The value 0.29 indicates the difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP. Specifically, the rate of change for the patients taking SC/RP is -0.582 + 0.29 = -0.209, i.e. the CAL variable drops faster in patients taking SC/RP+FL+AB. The difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP+FL is given by value -0.35. Therefore, the rate of change for patients taking SC/RP+FL is -0.932, which means that it drops at a faster rate in patients taking SC/RP+FL+AB. Also, value -0.38 indicates the difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP+AB. The rate of change for patients taking SC/RP+AB is given by value -0.962 and in this case the CAL variable is shown to be dropping faster in patients forming the reference level who take SC/RP+FL+AB.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Repeated Measures	CS diagonal offset	10.416140	1.317549	7.906	.000	8.129005	13.346772
	CS covariance	13.063507	4.191695	3.117	.002	4.847935	21.279078

Table 5.4.2 : Estimates of variances for the 1st tooth, for the imputed dataset.

The above *Table 5.4.2* contains the estimates of variances. We note that the variance between subjects is greater than that within subjects. This leads us to conclude that there is a moderate correlation between our observations. This can be calculated using the following formula:

$$Cor\left(Y_{ij}, Y_{ij'}\right) = \frac{{\sigma_{n_0}}^2}{{\sigma_{n_0}}^2+{\sigma_{n_\varepsilon}}^2} = \frac{13.063507}{13.063507+10.416140} = 0.55645 \approx 56\%$$

As we can therefore see, the correlation between the observations is 56%, so we conclude that it is a moderate correlation.

5.4.2 Explanation of parameters of the final model for the second tooth

The appropriate model for the imputed data relating to the second tooth is again the model (5.5) but here the covariance matrix has AR(1) structure. Below, the *Table 5.4.3* contains the coefficients of the final model for the second tooth.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	17.237793	10.253964	40.094	1.681	.101	-3.484724	37.960310
[sex=0]	-.584958	1.251269	39.461	-.467	.643	-3.114944	1.945028
[sex=1]	0[b]	0	.	.	.	.	.
[treatment=0]	-8.515510	2.479742	62.689	-3.434	.001	-13.471359	-3.559662
[treatment=1]	-7.602748	2.937799	66.256	-2.588	.012	-13.467830	-1.737665
[treatment=2]	-4.264278	2.849756	63.959	-1.496	.139	-9.957392	1.428835
[treatment=3]	0[b]	0	.	.	.	.	.
time	-1.959765	.532684	109.891	-3.679	.000	-3.015431	-.904100
age	.135645	.258418	39.461	.525	.603	-.386860	.658150
cigar	-.041764	.038250	39.461	-1.092	.282	-.119104	.035576
boneloss_beg	.234495	.116392	39.461	2.015	.051	-.000842	.469833
[treatment=0] * time	.459549	.673797	109.891	.682	.497	-.875774	1.794871
[treatment=1] * time	.245960	.842247	109.891	.292	.771	-1.423194	1.915113
[treatment=2] * time	.206644	.790097	109.891	.262	.794	-1.359161	1.772449
[treatment=3] * time	0[b]	0	.	.	.	.	.

Table 5.4.3 : Coefficients of the final model for the 2nd tooth, of the imputed dataset.

The value 17.237 indicates the expected value of the CAL variable for female patients taking SC/RP+FL+AB, have the same time measurement, the same age, the same number smoked cigarettes and the same rate of bone loss at the begin of the therapy. The average difference of the dependent CAL variable between the two sexes is -0.58, with the other explanatory variables held constant. The average difference of the CAL variable between patients received the therapy SC/RP+FL+AB and the therapy SC/RP and the other explanatory variables held constant, is -8.51. Keeping constant the explanatory variables relating to age, gender, number of cigarettes and the bone loss in the begin of therapy, apply the following. The average difference between patients taking SC/RP+FL+AB and SC/RP+FL is -7.60 and the average difference between patients taking SC/RP+FL+AB and SC/RP+AB is -4.26. The rate of change of the CAL variable in the reference level which consists of patients taking SC/RP+FL+AB, is -1.959. Value 0.45 indicates the difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP. Specifically, the rate of change in patients taking SC/RP is -1.959 + 0.45 = -1.509, which means that the CAL variable drops faster in patients taking SC/RP+FL+AB. The difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP+FL is given by value 0.24. That means that the rate of change in patients taking SC/RP+FL is -0.719, which means that it drops faster in patients taking SC/RP+FL+AB. Also, value 0.20 indicates the difference in the rate of change between patients taking SC/RP+FL+AB and SC/RP+AB. The rate of change in patients taking SC/RP+AB is given by value -1.759. In this case we also see that the CAL variable drops faster in patients forming the reference level and taking SC/RP+FL+AB.

Finally, *Table 5.4.4* contains the estimates of covariances for the 2nd tooth. We note that the variance between subjects is greater than the variance within subjects in this case, too.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Repeated Measures	AR1 diagonal	22.735926	3.184513	7.140	.000	17.277818	29.918266
	AR1 rho	.514151	.068058	7.555	.000	.368797	.634959

Table 5.4.4 : Estimates of variances for the 2nd tooth, for the imputed dataset.

In this case, because the covariance matrix has AR(1) structure, the correlation between the observations differ in a time measurement, is given by the formula:

$$Cor(Y_{ij}, Y_{i,j+1}) = \rho = 0.514151 \approx 51\%$$

while the correlation between observations differ in two time measurements, is given by the formula:

$$Cor(Y_{ij}, Y_{i,j+2}) = \rho^2 = (0.514151)^2 = 0.26435 \approx 26\%.$$

Finally, we observe that as the observations draw away annals, their correlation is reduced.

5.5 An application of Rubin's rules

This section presents an implementation of Rubin's rules, mentioned in the Chapter 2. We choose $m = 10$ imputations. This ensures 10 complete imputed datasets. We will fit the appropriate chosen mixed models of the section 5.4, for each of the two examined teeth, for the 10 complete imputed datasets, in order to obtain the estimates of the models. The estimates $Q^{(t)}$ are presented in the following tables for each tooth individually.

	$Q^{(1)}$	$Q^{(2)}$	$Q^{(3)}$	$Q^{(4)}$	$Q^{(5)}$	$Q^{(6)}$	$Q^{(7)}$	$Q^{(8)}$	$Q^{(9)}$	$Q^{(10)}$
Intercept	19.29	17.02	22.49	18.11	19.19	17.24	16.96	16.84	19.8	18.76
Sex=0	2.73	2.44	2.66	2.81	2.66	2.82	2.63	2.88	2.5	2.86
Treatment=0	-9.03	-8.7	-8.85	-8.86	-9.09	-8.63	-8.65	-8.38	-8.84	-8.89
Treatment=1	-4.47	-4.3	-3.75	-4.29	-4.21	-3.9	-4.42	-3.64	-3.84	-4.34
Treatment=2	-3.68	-3.51	-3.66	-3.63	-3.9	-3.08	-3.54	-3.15	-3.81	-3.58
Time	-1.76	-1.57	-1.7	-1.8	-1.73	-1.51	-1.59	-1.56	-1.88	-1.58
Age	0.12	0.17	0.06	0.14	0.1	0.15	0.17	0.13	0.17	0.13
Cigar	0.24	0.02	0.03	0.02	0.02	0.02	0.01	0.02	0.03	0.01
Boneloss_beg	0.13	0.14	0.1	0.15	0.15	0.14	0.14	0.17	0.07	0.14
Treatment=0*time	0.47	0.28	0.41	0.51	0.44	0.22	0.3	0.27	0.59	0.29
Treatment=1*time	-0.12	-0.19	-0.26	-0.2	-0.27	-0.13	-0.23	-0.56	0.17	-0.35
Treatment=2*time	-0.32	0.47	-0.3	-0.21	-0.23	-0.49	-0.31	-0.42	-0.08	-0.38
CS diagonal offset	11.48	10.88	10.4	11.37	10.24	10.38	11.52	10.43	10.9	10.41
CS covariance	12.14	12.43	11.62	12.17	12.69	13.61	12.59	12.81	12.08	13.06

Table 5.5.1 : Estimates $Q^{(t)}$ for the 1st tooth.

	$Q^{(1)}$	$Q^{(2)}$	$Q^{(3)}$	$Q^{(4)}$	$Q^{(5)}$	$Q^{(6)}$	$Q^{(7)}$	$Q^{(8)}$	$Q^{(9)}$	$Q^{(10)}$
Intercept	19.32	16.8	19.15	18.83	17.66	16.12	17.62	16.59	17.91	17.23
Sex=0	-0.68	-0.73	-0.64	-0.7	-0.62	-0.64	-0.7	-0.56	-0.82	-0.58
Treatment=0	-8.91	-8.7	-8.89	-8.99	-8.58	-8.44	-8.81	-8.5	-8.89	-8.51
Treatment=1	-7.65	-7.67	-7.76	-7.87	-7.55	-7.71	-7.81	-7.66	-7.91	-7.6
Treatment=2	-4.76	-4.61	-4.78	-4.88	-4.39	-4.27	-4.51	-4.23	-4.8	-0.26
Time	-2.14	-2.2	-2.08	-2.15	-2.03	-1.88	-2.2	-1.99	-2.2	-1.95
Age	0.1	0.17	0.1	0.11	0.13	0.16	0.14	0.14	0.15	0.13
Cigar	-0.03	-0.03	-0.03	-0.03	-0.04	-0.03	-0.04	-0.04	-0.03	-0.04
Boneloss_beg	0.2	0.21	0.21	0.21	0.23	0.23	0.22	0.24	0.2	0.23
Treatment=0*time	0.67	0.7	0.57	0.66	0.52	0.37	0.7	0.48	0.7	0.45
Treatment=1*time	0.48	0.54	0.38	0.4	0.28	0.39	0.41	0.23	0.72	0.24
Treatment=2*time	0.37	0.49	0.35	0.45	0.29	0.09	0.43	0.17	0.42	0.2
AR(1) diagonal	22.38	22.35	22.08	22.11	22.51	23.01	22.26	22.36	22.27	22.73
AR(1) rho	0.45	0.51	0.52	0.48	0.51	0.52	0.51	0.54	0.51	0.51

Table 5.5.2 : Estimates $Q^{(t)}$ for the 2nd tooth.

As seen from the above tables, the model's estimates are sufficiently close for both teeth. Then, according to the rules of Rubin, we compute $m = 10$ different estimates $\widehat{Q^{(t)}}$ with $t = 1, \dots, 14$ of the quantity Q and the estimated variance $^{(t)}$ of Q$^{(t)}$ with $t = 1, \dots, 14$. The results are shown in the following tables and we observe that the values of the estimates are also quite close.

	$\widehat{Q^{(1)}}$	$\widehat{Q^{(2)}}$	$\widehat{Q^{(3)}}$	$\widehat{Q^{(4)}}$	$\widehat{Q^{(5)}}$	$\widehat{Q^{(6)}}$	$\widehat{Q^{(7)}}$	$\widehat{Q^{(8)}}$	$\widehat{Q^{(9)}}$	$\widehat{Q^{(10)}}$	$\widehat{Q^{(11)}}$	$\widehat{Q^{(12)}}$	$\widehat{Q^{(13)}}$	$\widehat{Q^{(14)}}$
1st tooth	18.57	2.69	-8.79	-4.15	-3.55	-1.67	0.13	0.04	0.13	0.38	-0.21	-0.23	10.80	12.52
2nd tooth	17.72	-0.67	-8.72	-7.72	-4.15	-2.10	0.13	-0.03	-0.22	0.58	0.41	0.33	22.41	0.51

Table 5.5.3 : Estimates $\widehat{Q^{(t)}}$ for both teeth.

	$U^{(1)}$	$U^{(2)}$	$U^{(3)}$	$U^{(4)}$	$U^{(5)}$	$U^{(6)}$	$U^{(7)}$	$U^{(8)}$	$U^{(9)}$	$U^{(10)}$	$U^{(11)}$	$U^{(12)}$	$U^{(13)}$	$U^{(14)}$
1st tooth	3.08	0.02	0.04	0.09	0.06	0.001	0.001	0.001	0.001	0.001	0.03	0.07	0.25	0.31
2nd tooth	1.20	0.005	0.04	0.01	1.91	0.01	0.006	0.00	0.001	0.01	0.02	0.02	0.08	0.006

Table 5.5.4 : Estimated variances U$^{(t)}$ for both teeth.

And finally, we calculate the average $\bar{Q}$ of the estimates $\widehat{Q^{(t)}}$, the within imputation variance $\bar{U}$, the between imputation variance B, the total variance T and the quantities γ and r. Both γ and r are used as diagnostic statistics to examine the effect of missing data in estimates of $\bar{Q}$. The calculations of these estimates were made using R and the code is available in the appendix. From the following *Table 5.5.5*, we conclude that for both teeth the imputation made to our data does not affect so much the result.

	1st tooth	***2nd tooth***
$\bar{Q}$	1.907571	1.352000
$\bar{U}$	0.2876794	0.2378605
B	76.05882	105.4584
T	69.43206	96.10918
r	290.8262	487.6989
γ	0.9970627	0.998246

Table 5.5.5 : Results of Rubin's Rules.

Chapter 6

6.1 Conclusions

Out under this Chapter we will discuss the results of the models between the 4th and 5th Chapter as well as the similarities and differences. We recall that in the 4th Chapter we use the data with missing values while the 5th Chapter we use the imputed dataset.

The main difference between the initial and the imputed dataset is that we reached in different models. In every case the appropriate model for each dataset is chosen according to the Bayesian Information Criterion (BIC). Preferable model considered this with the smallest BIC value.

In particular in the 4th Chapter, when we use the dataset with the missing values, the most appropriate model was the model (4.3) which has only random intercept, for both teeth. In both cases the model (4.3) had the lowest BIC value. The correlation between observations for the first tooth was 65%. This indicating a moderate correlation. While, for the second tooth was 75%, which shows a strong correlation between observations.

On the other hand, in the 5th Chapter, which dealt with the imputed dataset, the most appropriate model for the first tooth was the model (4.5) with a compound symmetry covariance matrix and for the second tooth the appropriate model was again the model (4.5) but with a structure of AR(1). The correlation between observations pertaining to the first tooth, in this case, is 56% and this indicates a moderate correlation between observations. Concerning the second tooth we noticed that the correlation between observations is also weak and even decreases as the observations fend off annals. This leads us to the conclusion that if we had not apply the multiple imputation in the dataset we will lose a quite important information of our data.

Finally, it is worth mentioning that as regards the inference, for both models of the 4th and 5th Chapter, the interpretation of the parameters between the models, at the same tooth, concerned no significant differences.

Appendix

##Descriptives for the variables “age”, “boneloss_beg”, “boneloss_fin” and “cigar”

```
hist(age,main = paste("Age of patients"),xlab="years",col="coral2")
hist(Boneloss_start,main = paste("The percent of bone loss at the beggining of the therapy"),xlab="years",col="coral2")
hist(Boneloss_fin,main = paste("The percent of bone loss in the end of the therapy"),xlab="years",col="coral2")
hist(cigar,main = paste("Number of cigarettes per day"),xlab="cigarettes",col="coral2")
```

##3-D pie charts for the variables “treatment”, “sex”

```
library(plotrix)
slices <- c(73,38,33,42)
lbls <- c("SC/PR", "SC/RP+AB", "SC/RP+FL","SC/RP+FL+AB")
pie3D(slices, labels=lbls, explode=0.1, main="Method of treatment", col = c("blue","red","yellow","green") )

slices <- c(11,14)
lbls <- c("male", "female")
pie3D(slices,labels=lbls,explode=0.1,main="Patients Sex",col=c("wheat","thristle"))
```

##boxplots for the variables “boneloss_beg” and “boneloss_fin”

```
par(mfrow=c(1,2))
```

```
boxplot(Boneloss_start~sex,data=olokliro,col=c("wheat","thistle"), main="bone loss in the beginning of the therapy per sex")
boxplot(Boneloss_fin~sex,data=olokliro,col=c("wheat","thistle"), main="bone loss in the end of the therapy per sex")
par(mfrow=c(1,1))
##for Figure 5.1  page 39
library(VIM)
aggr_plot <- aggr(imptim, col=c('palevioletred2','black'), numbers=TRUE, sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))

##Rubin Rules
Qhsm3<-read.delim("C://Users//ni12__000//Desktop//SM3HAT.txt",header=T)
Qhsm4<-read.delim("C://Users//ni12__000//Desktop//SM4HAT.txt",header=T)

Qhat3<-apply(Qhsm3,1,mean)
Qhat3
Qhat4<-apply(Qhsm4,1,mean)
Qhat4

U3<-apply(Qhsm3,1,var)
U3
U4<-apply(Qhsm4,1,var)
U4

Qbar3<-mean(Qhat3)
Qbar3
Qbar4<-mean(Qhat4)
Qbar4

Ubar3<-mean(U3)
Ubar3
```

```
Ubar4<-mean(U4)
Ubar4

m<-10
B3<-sum( ((Qhat3-Qbar3)^2) )/ (m-1)
B3
B4<-sum( ((Qhat4-Qbar4)^2) )/ (m-1)
B4

###alliws ta B3~B4
var(Qhat3);var(Qhat4)

m<-10
T3<-Ubar3+B3*m/(m+1)
T3
T4<-Ubar4+B4*m/(m+1)
T4

m<-10
df3<-(m+1)*(1+(Ubar3/B3)*1/(m+1))
df4<-(m+1)*(1+(Ubar4/B4)*1/(m+1))
df3;df4

r3<-(1+1/m)*B3/Ubar3
r4<-(1+1/m)*B4/Ubar4
r3;r4

g3<-(r3+2/(df3+3))/(r3+1)
g4<-(r4+2/(df4+3))/(r4+1)
g3;g4
```

```
ci3<-c(Qbar3-qnorm(1-0.05/2)*sqrt(T3),Qbar3+qnorm(1-0.05/2)*sqrt(T3))
ci4<-c(Qbar4-qnorm(1-0.05/2)*sqrt(T4),Qbar4+qnorm(1-0.05/2)*sqrt(T4))
ci3 ; ci4
```

References

Jose C. Pinheiro, Douglas M. Bates (2000) Mixed Effect Models in S and S-PLUS, Springer-Verlag, New York

Geert Verbeke, Geert Molenberghs (2000) Linear Mixed Models for Longitudinal Data, Springer-Verlag, New York

Charles E. McCulloch, Shayle R. Searle (2001) Generalized, Linear and Mixed Models, John Wiley & Sons, New York

Clayton, D. (1992) Generalized Linear Mixed Models in Biostatistics, The Statistician, pp.327-328

Donald B. Rubin (1996) Multiple Imputation After 18+ Years, Journal of the American Statistical Association, pp. 473-489

Donald B. Rubin (1987) Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons, New York

Stef van Buuren, Karin Groothis-Oudshoorn (2011) Multivariate Imputation in R, Journal of Statistical Software

Nickolas J. Horton, Ken P. Kleinman (2007) Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models

Joseph L. Schafer, John W. Graham (2002) Missing Data: Our View of the State of the Art, American Psychological Association

Andrew Gelman, Iven Van Mechelen, Geert Verbeke, Daniel F. Heitjan, Michel Meulders (2005) Multiple Imputation for Model Checking: Completed-data Plots with Missing and Latent Data