LONGITUDINAL MODELS OF IRON STATUS IN A POPULATION-BASED COHORT OF MOTHERS AND CHILDERN IN SOUTHWEST ENGLAND

Sayed Mohsen Hosseini

A Dissertation Submitted to the

University of Glasgow

For the degree of

Doctor of Philosophy

UNIVERSITY
of
GLASGOW

Department of Statistics

July 2004

Abstract

Iron deficiency is the most common delineated form of nutritional deficiency. Its prevalence is high in childhood, and in women of childbearing age (particularly during pregnancy). Infants and young children are at high risk of iron deficiency because their need for iron during the period of growth is high and also their diet is often low in iron and iron supplements.

In almost all countries, measures to control iron-deficiency anaemia are implemented. These measures usually focus on women during pregnancy and young children and consist of distributing iron supplements and, to a lesser degree, iron fortification by an appropriate food method.

This thesis reports the longitudinal changes in haemoglobin concentration of pregnant women and their children in the Southwest of England, using data collected as part of The Avon Longitudinal Study of Parents and Children (ALSPAC).

Longitudinal data are used in the study of growth, or improvement, and consist of the same measurements made on the same subjects repeatedly over time. This longitudinal study is used to examine changes in the haemoglobin level of children and of mothers during pregnancy and the effect on haemoglobin of some time stationary and time varying covariates.

Longitudinal data requires special statistical methods because the observations on one subject tend to be correlated. (Although subjects can usually be assumed to be independent). When subjects are individually observed at varying sets of times with or without missing data, as is the case for ALSPAC data during pregnancy, then the resulting data is referred to as unbalanced data. This can cause further complications for the analysis.

The aim of this thesis is to contribute to longitudinal research of this topic by using mixed-effects models, which provide a powerful and flexible tool for the analysis of balanced and unbalanced data.

Chapter one consists of an introduction to the thesis. Subsequently, Chapter 2 details previous relevant research in the area of this thesis according to methods, and these methods are applied in Chapters 3, 4 and 5.

In Chapter 3 the inferential focus is on modelling the change in mean haemoglobin levels during pregnancy in three groups of mothers as defined by their use of iron supplements. Group 1 did not use iron supplements during pregnancy; Group 2 began to use supplements before Week 18 and Group 3 between Weeks 18 and 32.

This chapter fits a polynomial function to the data set. This function is fitted to the data using the method of General Linear Mixed Models. These Mixed Effect Models are fitted with and without interaction terms. Polynomial regressions are tested in order of complexity, when the highest order term is chosen then all lower order terms are included as well.

Attention is focussed on curve fitting methods that are able to provide explicit functions for these data. In this chapter the cubic spline algorithm is used to find an adequate model. We can see that the cubic spline reconstruction generally performs better than the polynomial fitting procedure to obtain a meaningful model. Therefore, the spline approach is preferred for the haemoglobin levels during pregnancy. The spline model is calculated with knots at 12,18, 24, 32 and 36 weeks of pregnancy for the three groups of mothers separately. The exception is that one parameter is not significant for the Group 2 model, but the full set of knots was retained even in this case.

Test are carried out to investigate the association between the iron supplement status of mothers during pregnancy and each of following variables; mother's education level, age, and ethnicity, number of cigarettes smoked per day in 1st 3

months, number of cigarettes smoked in 2 last weeks of pregnancy, number of cigarettes smoked by 32 weeks of pregnancy, singleton/multiple pregnancy, mother's parity and vegetarianism of mothers .The spline models are then extended in order to investigate the influence of these variables on haemoglobin levels during pregnancy in each group of mother separately.

In brief, in-Group 1 there are significant effects of the following covariates on haemoglobin levels: vegetarianism of mothers (without interaction between this covariate and time); number of cigarettes smoked per day in first 3 months, number of cigarettes smoked per day in last 2 weeks of pregnancy and parity (with interactions between these covariates and time).

In-Group 2, the model includes a significant effect of ethnicity of mother, multiple pregnancy and number of cigarettes smoked per day in first 3 months of pregnancy on haemoglobin levels.

There was also a significant interaction between multiple pregnancy and time. In-Group 3, haemoglobin levels were significantly associated with mother's parity, education level of mothers and vegetarianism of mothers but there was just a significant interaction between mothers who were vegetarian in the past and time at knots (18,24).

In Chapter 4, an identical statistical method to Chapter 3 is used to model haemoglobin levels in children up to 7 years of age.

The age of each child (to the nearest week) is recorded on each occasion, along with the haemoglobin level.

Explore the effect of several time stationary covariates, including maternal and children covariates, on haemoglobin levels, and to determine the effects of nutritional intake, which is represented by a set of time varying covariates. In chapter 4, covariates of interest in our investigation included: maternal age: vegetarianism of mother; education; number of cigarettes smoked per day in 1st 3

months, the last 2 weeks, by 32 weeks of pregnancy; parity; and singleton /multiple pregnancy; birth weight, sex and ethnicity of children.

Furthermore, fitted values of the intercepts and slopes from the model, for maternal data, were included as time stationary covariates.

Time-varying covariates included the weight of the children and nutritional intake such as: vitamin C, Haem iron, Fat, Saturated fat, Monounsaturated fat, Energy, Calcium, Iron and Non-Starch Polysaccharide (NSP). These measurements were only made at 18 and 43 months.

The best fitting model is produced by a cubic spline with three knots at 69, 228 and 324 weeks.

The mean haemoglobin level is significantly affected by education level of mother, sex of child, parity and birth weight (as time stationary covariates) and by Vitamin C, NSP, Calcium intake and weight of child, (as time varying covariates). There are no significant interactions between these covariates except in the case of education level of mothers. The effect of a mother's haemoglobin level during pregnancy on a child's subsequent haemoglobin levels is assessed. This had to be done separately for the three groups of mothers according to iron supplement use. The results show that random slopes are significantly associated with children's haemoglobin level just in Group 1.

The aim of Chapter 5 is to introduce a longitudinal reference curve for haemoglobin levels in children, from age 8 months to 7 years.

This study describes the extension of the reference curves approach of Cole and Green to create conditional reference curves (Cole 1994).

Firstly, cross-sectional reference curves were fitted to the data using the LMS method (Cole & Green 1992) ,and then a longitudinal approach to conditional reference curves is developed where haemoglobin levels at time t are predicted from haemoglobin levels one time previously ($t-1$). Subsequently, this work is

extended to give a new conditional reference curve by including haemoglobin levels at all previous times rather than just one time.

In all cases, reference curves representing the 3^{rd}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} and 97^{th} centiles are obtained for the age range 8 to 84 months.

In the cross-sectional reference curves, the centiles are relatively far from the median in early childhood, but thereafter they come closer to the median.

The conditional reference curves fitted using data from one previous time are roughly parallel up to 300 weeks (75 months) and then diverge for older children.

All conditional reference curves are much closer to the conditional median than cross-sectional reference curves to their median. The conditional reference curves that are based on data from all previous time perform particularly well.

Chapter 6 begins with an outline of the main findings of the thesis and a discussion of possible future work in this area. Although progress has been made in the study reported in this thesis, further extensions are required. As the longitudinal data typically need some structured covariance models, the overall findings indicate that when the number of occasions is large with some missing values, the use of polynomial functions is inadequate to describe the model. This study highlights an approach that applies cubic spline in longitudinal modelling, including an emphasis on the use of graphical representation for exploratory analysis and the assessment of model fit.

Cubic splines provide a flexible tool for longitudinal data. The main objective of this study is to investigate a methodology to incorporate cubic spline with linear mixed models in modelling longitudinal data with number of time points and missing values.

Acknowledgements

First, I would like to give thanks to God, who has made the completion of this project possible. There are many people who have helped me during my study. Moreover, I have found an amazing density of gifted and kind people here in this department who have helped me and I feel unjustifiably fortunate.

I would like to acknowledge my supervisor Mr. John McColl; none of this work would have been possible without John, to whom I am incredibly indebted for his time, advice and guidance over the past four years. Time will tell whether I have absorbed his ability think creatively.

My deepest appreciation also goes to the Ministry of Health and Medical Education in Iran and Esfahan University of Medical Sciences for giving me the graduate scholarship.

I would like to extend a special thank you to Professor Adrian Bowman. His guidance has been very helpful throughout my study and his assistance is appreciated.

My sincere gratitude is also extended to Dr Andrea Sherriff for helping me out by solving problems and for her helpful ideas and advice for my research.

My heartfelt appreciation to my father and mother whose encouragement and wisdom helped me overcome any challenges I faced, be they big or small.

Words cannot adequately describe the appreciation I feel for my wife and my sons who shared in both the despair and the elation I felt these past four years. My family's complete and constant faith in my endeavours allowed me to pursue my career goals without looking back.

I am also extremely grateful to all the mothers and children who took part and to the midwives for their cooperation and help in recruitment. The whole ALSPAC study team comprises interviewers, computer technicians, laboratory technicians; clerical workers, research scientists, volunteers and managers who continue to

make the study possible. The ALSPAC study is part of the WHO initiated European Longitudinal Study of Parents and Children. Core funding for the long term follow up of the cohort came from the Medical Research Council, the Welcome Trust, the UK Department of Health, the Department of the Environment, DfEE, the National Institutes of Health, and a variety of medical research charities.

Contents

List of Tables

List of Figures

CHAPTER 1

Background

1.1. Introduction:

1.1.1. Iron Deficiency:

The most common nutritional disorder in the world is Iron deficiency, the highest number of sufferers being found amongst children in early life and women, particularly during pregnancy. Due to Iron deficiency, 1.3 billion people, around 30% of the world's population, are anaemic(Cook, Skikne, & Baynes 1994).

Interestingly, Iron deficiency is not an exclusive health problem of the developing world, as it also affects the population of the industrialised world for example Europe (Hallberg 1995).

According to a report by the World Health Organization (WHO) around 43% of the children in the world are anaemic (Hb $\leq$ 11 g/dl).

Early detection of anaemia can help to prevent complications, related to pregnancy and delivery as well as child development problems.

Information on the prevalence of anemia is useful for the development of health intervention programmes designed to prevent it, such as Iron fortification programmes.

Iron levels in the body, are influenced by a range of factors which have been categorised into three main groups; dietary, host-related co-physiological and environmental factors.

Iron levels in the newly born appear to be indirectly influenced by environmental factors such as socio-economical background and birth order, through the diet type and food eaten amount (Wharf et al. 1997).

Haemoglobin levels are a widely used measure for assessing Iron deficiencies, many things can cause low haemoglobin levels and anaemia, and a common cause of anaemia is people's diet. Inadequate dietary intake and bioavailability are important factors contributing to poor Iron status, such as Iron deficiency anaemia (International Food Policy Research Institute 2000).

The aim of this study is to use statistical methods for the analysis of longitudinal data. That is, data in the form of repeated measurements of the same experimental unit over time, to produce a model for Iron status during pregnancy and in the early years of childhood.

1.1.1.1. Iron Deficiency in Pregnant Women:

The risk of anaemia is increased by pregnancy. Of the 8684 pregnant women who delivered in Oxford between January 1987 and January 1989, thirty-nine percent had haemoglobin levels lower than 11g/dl and ten percent of the mothers had haemoglobin levels lower than 10g/dl (Godfrey et al. 1991).

According to Tapiero's report, over 45% of women during pregnancy are anaemic in the world, whereas in developing countries around 60% of pregnant women are anaemic(Tapiero, Gate, & Tew 2001).

In order to evaluate anaemic women during pregnancy, it is essential to know their complete health history, also a physical and a blood smear examination must be carried through, plus a complete blood count with indices. More tests for measurement of red cell folate levels and ferritin may also be necessary according to the test findings. As the hematocrit and some other variables change the normal physiology during pregnancy, diagnosing between true anaemia and anaemia etiology is challenging (Jong, Romano, & Gibson 2002).

Since nutritional requirements increase during pregnancy, Iron deficiency anaemia is one of the most common forms of anaemia.

Therefore, this kind of anaemia is prevalent in women who have inadequate diets and who are not receiving antenatal Iron and folic acid (folate) supplements (Williams & Wheby 1992).

There are schemes to control Iron deficiency anaemia in almost countries. These schemes usually focus on women during pregnancy and their children. In the case of the former, consisting of distribution of oral Iron supplements and in the case of the latter, Iron fortification via a suitable food method.

Despite the worldwide efforts by an array of public health organizations in trying to keep down anaemia and Iron deficiency anemia, it is still an endemic disease in many areas of the world, and its annihilation is a serious health problem.

Moreover, prevalence of Iron deficiency anaemia is high in developing countries, being especially difficult to control and solve amongst pregnant women(Hercberg, Galan, & Preziosi 2000).

In the world, about half of women during pregnancy are anaemic and in most cases this is due to Iron deficiency.

So far, the introduction of Iron fortified formula coupled with cross-sectional haemoglobin measures, has clearly helped to reduce anaemia levels as reported by the Walravens study (Walravens 1989).

Worldwide, Iron deficiency is the most common nutrient deficiency, pregnant women being an especially high-risk group due to their low Iron reserves.
The worldwide estimation is that 60 million women during pregnancy are anaemic, of which 4 million of them live in developed countries.
The prevalence of Iron deficiency anaemia in pregnant woman varies amongst countries. Generally, it has been found that there is a tendency for levels of anaemia to be low in the first trimester while rising during the second trimester. Around half of Iron deficiency anaemia cases are found after the 25th week of pregnancy (Hercberg, Galan, & Preziosi 2000;Milman, Agger, & Niesn 1994).The effects of Iron deficiency anaemia include a premature delivery and delivering a low-birth weight child.

1.1.1.2. Iron Deficiency in children:

Iron deficiency can be a serious problem in childhood. In children, the most likely causes are low levels of available Iron in the diet. This together with an increased Iron demand for growth exacerbates the problem.
In newborns, Iron reserves are usually adequate up to the first 4 to 6 months of age. However there after as Iron is needed for producing energy, children become dependent on extra available Iron. Making the under two years age group, the most vulnerable one according to(Aggett, Barclay, & Whitley 1989), when this demand is not met.
The development of Iron deficiency which can lead to the development of Iron deficiency anaemia can have serious consequences, as it has been observed to delay both mental and psychomotor development, as well as physical growth (Lozoff, Jimenez & Wolf 1991).

1.1.1.3. Iron supplementation:

Around 30% of the people in the world are affected by Iron deficiency. Iron deficiency is most common in the developing world due to vegetarian diets, which reduces the availability of Iron in the diet.

Additionally, requirement for Iron is increased by rapid growth and blood losses, the former increasing the risk of anaemia in infants and the latter increasing the risk of anaemia in women who undergo menstruation (Ma et al. 2002).

Moreover, several studies have shown that there appears to be a correlation between Iron supplementation and a lower frequency of problems during pregnancy as well as after birth (Allen 1997). For instance in a study performed on Finnish women during pregnancy, randomized into mothers with non-routine and mothers with routine Iron supplement, newborns had shorter mean length in mothers with non-routine Iron supplement than those born to mothers who routinely took Iron supplement. Also, in this latter group, significantly longer gestation took place(Hemminki & Rimpela 1991).

For several reasons, infants are particularly sensitive to nutritional deficiencies, partly due to the fact that often their range of food as well as its amount is inadequately low, in relation to their growing nutritional requirements.

Nestel 's 1993 study, suggested that this problem could be solved by providing the infants with a greater variety of highly nutritious food, these ones being absent in traditional diets(Nestel 1993).

Haemoglobin level is one of the most important clinical measurements used to diagnose and treat anaemia, via public health interventions.

In the main, a data analysis for cross-sectional study is of little help when wanting to evaluate the disease history, especially when interesting variables change over a period of time(Ware 1985). Thus, a better option is longitudinal haemoglobin measurements, which must be obtained from the same individual over a period of time at suitable intervals. This method is a reliable way of evaluating the risk factors in anaemia.

A repeated measures design, with suitable analysis, allows study of haemoglobin change in a time period within individuals and therefore provides a way of examining the effects of covariates at the individual subject level(Zeger & Liang 1992). This is precisely the type of data that can be found in the ALSPAC study.

1.2. The ELSPAC and ALSPAC studies:

The Avon Longitudinal Study of Parents and Children (ALSPAC) which was previously known as The Avon Longitudinal Study of Pregnancy and Childhood, forms part of the European Longitudinal Study of Pregnancy and Childhood (ELSPAC) cohort study(Golding 1989;Sherriff et al. 1999). The aims of the ELSPAC study are to identify ways in which to prevent illnesses happening and to maximise the health of the child.

The study centres involved with ELSPAC include Yarsoslavl in Russia, Brno and Znojmo in the Czech Republic, Bratislava in Slovakia, five centres in the Ukraine, as well as the Isle of Man and Avon in the UK.

All of the ELSPAC study centres share the following study pattern:

- The study covers all women during pregnancy resident in a determined geographical area with an expected date of delivery between specific dates.

- Self-completion questionnaires posted to the mother and her partner, this information can then be linked to information from health records.

- The questions should be the same in all study centres and asked in the same way to obtain a core data set. However, questions such as diet and educational level, which were culture-specific, were excluded.

- As much additional detail as desired locally could be added to that required by the overall study strategy, as long as none of the core data was dropped.

The branch of ELSPAC known as ALSPAC, The Avon Longitudinal Study of Parents and Children is a population based study which investigates which factors appear to have an influence on the health and the development of children (North, Emmett, & The ALSPAC Study Team 2000). All women during pregnancy resident within a certain area of Avon, in Southwest England, with an expected date of delivery between 1 April 1991 and 31 December 1992 inclusive, were invited to take part in this study and eligible for inclusion(Emmett, North, & Noble 2000).

In order to persuade women during pregnancy to enrol in this study, several plans were used.

Displaying posters in many places including pharmacies, general practitioner waiting rooms, antenatal clinics and inviting mothers to ask for further details of the study.

All mothers who were living within the area being studied were sent information. The total number of mothers enrolling in this study was over 14000, which was over 80% of eligible pregnancies in this area(Emmett, North, & Noble 2000).

Their children were as known *Children of the nineties* and provided a sample that appears to be a reasonable representation of British children as a whole(Sherriff et al. 2001).

1.2.1. The ALSPAC study area:

The Avon area under observation was within the Southwest Regional Health Authority but excluded Bath and district.

This area is 120 miles west of London and situated on the Seven estuary, its population is one million, including Bristol whose population is 0.5 million.
The data comprises a mixture of rural areas, inner city deprivation, leafy suburbs as well as moderate sized towns (Sidebotham & Golding 2001) .

1.2.2. Preparatory work:

A pilot study was done on approximately one hundred subjects; one of its aims was to make the questionnaire reader friendly. With this in mind, parents were asked to make suggestions on the wording of the first questionnaire drafts. So that it could be adequately modified.
Additionally, the ALSPAC Ethics and Law sub-committee also made suggestions on the phrasing of questions and other matters.

1.2.3. The questionnaires (During pregnancy):

Supposing that the mother wanted to take part in the study, the first questionnaire was posted to her, around one week after the brochure for joining in *children of the nineties* had been sent out. The gestation at enrolment of the woman determined, which questionnaire was sent to her. Table 1.1 shows the schedule for sending the questionnaires. Four questionnaires were sent to pregnant women, two of them were sent out at a fixed time, at Week 18 of pregnancy (B =‘ Having a baby ’) and at Week 32 of pregnancy (C =‘ Your Pregnancy ’).
If the woman had enrolled before 14th weeks of pregnancy, the questionnaire labelled ‘Your Environment’ (A) was posted to her forthwith.
The designers of this questionnaire were interested in finding out to which extent early environment could have an influence on the fetus.
The other questionnaire labelled *about yourself* (D) was related to the mother’s medical record, as well as her social and environment history and the time during

pregnancy. Then if these were necessary, after birth of baby the questionnaire was posted.

The questionnaire called *having a baby* (B) was not useful or valid for the mothers, which enrolled at week 18th of pregnancy or later, because of questions, which dealt with earlier stages of pregnancy.

However, in the mother's late enrolment questionnaire, some relevant information could be found, in relation to their environment and lifestyle.

Consequently, this useful information was joined together in a single questionnaire called *your home & lifestyle* (E).

After 7 days, if a mother had not replied to the questionnaire, a remainder was sent to her and if necessary a second remained was also sent, this time after waiting for a reply for a period of time of anything in excess of 10 days.

In the event that the mother had still not reply after one month, a member of staff would call her and pay her a home visit, offering her assistance in filling the questionnaire.

A short questionnaire called *Filling the Gaps* was posted to mothers after the birth of their children. These mothers did not receive *questionnaire C* that dealt with the ethniticity of both mother and child, as well as her social, educational and occupational background. All of these questions were not related to pregnant women in their third trimester (Golding et al. 2001).

1.2.4. Eligibility criteria of study:

All women resident in the Avon Health Authority area with an expected date of delivery between 1 April 1991 and 31 December 1992 were eligible for this study.

Resident women in this area, who had left Avon following a short time after enrolment, were dropped from follow-up.

Nevertheless, the questionnaires that had been filled by all of the mothers during the third trimester of pregnancy were used in the study, irrespective of the fact whether the mother had left the area being studied before or after delivery.

1.2.5. Children in Focus:

The total sample of children in the ALSPAC study was called *Children of the nineties.*

A 10% sample of the children born in the last 6 months of the study were selected at random to take part in this study known as *Children in Focus (CIF).* Parents of these babies were invited to bring their children to a research clinic at 4, 8 and 12 months of age and at six monthly intervals thereafter, where a number of clinical, physiological, and developmental assessments were carried out(Sherriff, Emond, Hawkins, Golding, & the ALSPAC Children in Focus Study Team 1999). The aims of this study (*Children in Focus*) were to validate some phases of the self-completion questionnaire, as well as finding information that might not be determined by the questioners.

Important questions investigated with the Children in Focus. Children were correlated with the effect that childhood diet, growth, anaemia, otitis media with effusion, visual defects, parenting skills and early cognition can have on the development of intellectual competence, speech, language and motor development(Golding, Pembrey, Jones, & The ALSPAC Study Team 2001).

Of the 1509 children's mothers who were invited to the 4-month clinic, 1023 of them attended.

With an invitation at 8 months to all of the mothers who had or had not attended the 4 months clinic visit, 390 subjects were added to the total size of the cohort study. Then 1413 mothers attended the 8-month clinic, of the 1509 who were originally invited to attend.

1.2.6. Ethics approval:

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Sub-Committee and from the three Medical Research Ethics Committees in The

Avon study area including the local ethics committee of United Bristol, Southmead and Frenchay Health Care Trusts(Sherriff, Emond, Hawkins, Golding, & the ALSPAC Children in Focus Study Team 1999).

1.3. The Data Availability:

1.3.1. Data Preparation:

There are a number of self-coding in the questionnaire responses. Arrange for data from the completed questionnaires to be keyed by the ticked box available in the questionnaire as print number. In the questionnaire, some questions ask for response as text, some ask for response as tick and comments.

A group of undergraduate students coded the complete questionnaires under the supervision of the staff, throughout the summer months.

The questionnaires were checked very carefully to be sure that there were no more than one tick per answer for each question and that any comments do not materially affect the meaning of the response.

Sometimes, converting dates, description of some problems included multiple ticking, or rounding of ages is necessary. In order to avoid error, a second group double-checked all coding.

The major problem with textual responses is the great variety of possible questions and responses to each of them.

Hence, keying all written responses, and then dividing each answer by each question type, solved this problem and one file was created to put all responses to one question.

The data are ready for analysis when those are accomplished for an especial questionnaire and also are edited (Golding, Pembrey, Jones, & The ALSPAC Study Team 2001).

1.3.2. The built data files:

Editing was done so as to exclude all information, which might help to identify an individual person, on receipt of all the coded and keyed data from any one questionnaire and then a file name *built file* was made.

The detail of questions asked, the coding outlines, the details of the derivation of scores and summary variables included in each built file were made into a document.

The questions, the table of response frequencies and the labels used can be found in The ALSPAC study Team document.

1.3.3. Data collection:

A myriad of sources were used to collect information from the beginning of pregnancy these being:

- Self-completion questionnaires, which were sent to mothers, their partners and their children (from age 5);

- Medical and educational records;

- Environmental measurements of sub samples of home such as air pollutants levels and noise;

- The assessment of a 10% sample of the data which had been randomly selected from the study population, from age 4 months (Children in Focus sample);

- A thorough interview, including an examination of the especial sub-group and their controls;

- An assessment of the total study using a standardised environment, each year from seven years old of the children onwards.

- The mother, her partner and the child biological samples(Golding, Pembrey, & Jones 2001).

Three sets of measurements are of particular interest in this thesis:

1. The Mothers' haemoglobin levels were under observation during pregnancy.

Demographic data for this study, obtained by postal questionnaire, included; the ethnic background, health, and lifestyle of the mother as well as the environment and the development of the child.

Blood samples of the mother during pregnancy were taken in routine antenatal testing, as well as DNA umbilical cord blood samples (Mumford 1999).

As any use of Iron supplements by the mother during pregnancy would have a very important impact on her haemoglobin level, the category of the mother in Iron supplement groups was determined from the questionnaires that were sent to the mother during pregnancy.

At two fixed time points, 18-20 weeks gestation and 32 weeks gestation, the relevant questionnaires were sent out. These weeks reflect the second and third trimesters of the pregnancy (Sidebotham & Golding 2001).

This means that we only know whether or not a mother was taking iron supplements by the 18th week of pregnancy and by the 32nd week.

2. The children's haemoglobin levels were determined at each research clinic visit.

A sample capillary of the children's blood from a heel prick was taken in an EDAT capillary tube. The HEMOCUE B-Hb photometer was used to check on the haemoglobin levels (Sherriff, Emond, Bell, Golding, & Team 2001).
Emond A in 1996 reported on the quality confidence strategy operating within the laboratory and the measures taken to test the stability of the samples once the 8 months blood samples were analysed (Emond et al. 1996) .It notes that there was no recording of the children's iron supplementation status.

3. Dietary intake data were assessed in children for three days in a particular week; two day during the week and one day during the weekend. The data was collected using household measures (unweighed) and records at ages 18 and 43 months old.
One week before of the scheduled clinic visit, the mother was sent a three-day dietary diary, in which she could record all the eating and drinking done by her child using household measures. All nutrition intakes were accounted by these records (Cowin et al. 2001).

1.4. Thesis outline:

The question, which we are going to answer in this study, is how mean haemoglobin level changes over time and other issues concerning the relationship between response and time.
It is necessary to represent the situation in terms of a statistical model that acknowledges the way in which the data were collected in order to address this question. Complementing the models, specialized methods of analysis are required.

The aims and objectives of this thesis are:

- To model haemoglobin concentration during pregnancy and investigate the effects of various covariates and Iron supplementation on haemoglobin levels in pregnant women.

- To model haemoglobin concentration during childhood from 8 months to 7 years old ages in *Children in Focus* and investigate the effects of the covariates on Iron status in these children.

- To present and compare reference curves for haemoglobin concentration in children aged between 8 months old and 7 years old, for detecting abnormal haemoglobin levels.

In this study, we will explore some recent approaches to analysing our data set, as methods for dealing with this type of longitudinal data have developed tremendously in recent years.

Chapter 2 discusses and reviews statistical modelling for dealing with longitudinal data and introduces, General Linear Model for longitudinal data, cubic spline, the Generalized Estimating Equation approach, software for analysis of longitudinal data, dealing with missing data and reference curves.

Chapter 3 presents results of modelling the mean haemoglobin concentration during pregnancy with particular emphasis on how Iron supplementation affects haemoglobin levels. Also in Chapter 3 the influences of a number of important covariates on the development of haemoglobin concentration during pregnancy are examined and discussed.

Chapter 4 applies the same methods used in Chapter 3 in order to model haemoglobin levels in the children between 8 months and 7 years in Children in Focus.

The aim of this Chapter is to describe the Iron status, and the effects of the covariates and Iron intakes on Iron status, in children followed longitudinally from 8 months old to 7 years old.

In Chapter 5, this study is looking for appropriate statistical methods for estimating reference curves.

For childhood haemoglobin levels, Chapter 5 describes an extension of the reference curve of Cole and Green to generate a longitudinal reference curve for haemoglobin concentration in Children in Focus.

The main aim of this part of work is to demonstrate some methods for producing and comparing reference curves for both cross-sectional and longitudinal data for children haemoglobin concentrations by covering ages between 8 months old and 7 years old.

Finally, Chapter 6 provides a discussion of the work done so far and prospects for the future of this study.

Gestation at enrolment (week)	Administration of questionnaires (week)				
	A	B	C	D	E
≤10	≤10	18	32	14	-
11-14	11-14	18	32	23	-
15-18	22	18	32	26	-
19-21	24	19-21	32	28	-
22-23	28	22-23	32	36	-
24-30	-	-	29-33	33-36	24-30
31-40	-	-	31-40	PD**	34-41

Table.1.1: Timing of antenatal questionnaires

* This table is taken from Golding J, 2001.

**Post delivery

CHAPTER 2

Literature review

2.1. Introduction:

It is interesting to researchers to model data that arise from a longitudinal or repeated measures study where there exists a correlation between observations on a given subject.

Normal Linear Models (NLMs) and Generalised Linear Models (GLMs) are standard regression models for independent data that are assumed to follow an exponential family distribution, but these must be generalised for correlated data structures(McCullagh & Nelder 1989). If the outcomes are approximately multivariate normal, then there are acceptable methods of analysis, namely General Linear Mixed-effects Models (GLMMs), which are generalisations of the NLM(Laird & Ware 1982).

The Generalised Estimating Equations (GEE) approach re-discovered by(Liang & Zeger 1986), which is based on multivariate quasi-likelihood theory, is an extension of the GLM approach to deal with longitudinal data. Both GLMMs and GEEs are described and discussed in this Chapter.

Some data that can be analysed using the General Linear Mixed Model include longitudinal data, repeated measures data and correlated data.

The General Linear Mixed Model can be represented as a two-stage model, where random effects are first sampled from a prior distribution and measurement data are then sampled from independent normal distributions with linear functions of these random effects and additional population parameters as fixed effects.

It is important to mention that choice of a multivariate normal model for the outcome implies that the full likelihood can be written down and parameter estimation can proceed in the usual maximum likelihood manner.

Several software tools for estimating this class of models are available, such as SAS Proc Mixed (SAS Institute 1992) , BMDP5V (Dixon & Chief 1990), HLM (Bryk, Raudenbush, & Congdon 1994), and MLn(Woodhouse 1995).

On the other hand, if the multivariate normal assumption breaks down, because the individual outcomes are binary or count data for example, general likelihood approaches are less easy to use, for the reason that it is difficult to specify a full multivariate distribution for such data. Some modelling approaches have been proposed for these kinds of data (Fitzmaurice, Laird, & Rotnitzky 1993). Possibly the most productive approach is Generalized Estimating Equations (GEEs), presented by (Liang & Zeger 1986;Zeger & Liang 1986) which takes account of the correlation between measurements in Generalized Linear regression Models without specifying the full multivariate distribution. The full distribution of the outcomes is not specified, but only the correlations among the outcomes from an individual or cluster. In this sense, Generalized Estimating Equations are generalisations of General Linear Models to the case of correlated data.

Additionally, Generalized Estimating Equations permit a range of different correlation patterns within clusters. Finally, Generalized Estimating Equation are estimable with a lot of currently available software packages(Zorn 2001).

In this chapter, we review the available methods for dealing with correlated data; in particular, the techniques of General Liner Mixed Models and Generalized Estimating Equations.

In the following section, General Linear Mixed Models are introduced, the Generalized Estimating Equations approach is outlined and software for analysis of longitudinal data is discussed.

2.2. Longitudinal Studies:

The defining feature of a longitudinal study is that individuals are repeatedly measured during a period of time. Whereas in cross-sectional studies an outcome is measured at a single time point for each individual.

In comparison between cross-sectional and longitudinal study, the major advantage of the longitudinal study is its capability to separate out what in the population studies are called cohort and age effects.

The effect of age varies over a period of time within individuals; the cohort effect is the difference amongst population groups who are born at various times. As a single outcome is only available in cross-sectional data, therefore cross-sectional studies cannot distinguish between these two effects.

Longitudinal studies can differentiate changes during a time period within individuals from differences amongst cases at the start of the study.

Longitudinal studies are most suitable for the investigation of individual changes during a period of time and for the study of effects of some important factors to influence variation. The main aim is comparison of the effects of treatments on an outcome variable.

However, a comprehensive analysis of such longitudinal data requires an adjustment of between and within subject change, unbalanced and missing data.

In the last two decades, much progress in the modelling and analysis of longitudinal data has been made methodologically and computationally (Cnaan, Laird, & Slasor 1997;Diggle, Liang, & Zeger 1994;Molenberghs & Verbeke 2002).

Consider a longitudinal clinical trial, where cases are first randomly allocated to one of a set of possible treatments, and then followed for a time period. This is the simplest type of longitudinal study.

Effects of treatment are shown by differences in evolution during a period of time and by interactions of treatment with time. Randomisation ensures that the treatment groups are certainly comparable at baseline according to factors that potentially influence change later. Therefore, a statistical model for this data does not need a cross-sectional model component (Verbeke , Spiessens, & Lesaffer 2001).

In observational studies, some groups may not be very similar at the beginning, therefore longitudinal changes require to be studied after correction for some confounders such as age and gender.

For future reference, it is important to mention here that ALSPAC is an observational longitudinal study.

Suitable models for data from observational studies of this type cannot assume that all cases are measured on the same occasions nor that the same number of observations is available for all cases. This is very important since individuals may enter and drop out of the study at any time.

Particular statistical methods are demanded by longitudinal data because the observations on one case tend to be correlated. This correlation must be taken into account in order to draw valid inferences.

2.2.1. The aim of longitudinal research:

1. To characterise outcomes patterns (e.g. growth, haemoglobin level, blood pressure) over time.

2. To investigate the effects of important covariates on outcome, the covariates in longitudinal studies are two groups on the following;

   - Time-independent or time stationary (between-subjects) covariates (e.g. sex, ethnicity),

   - Time-dependent or time varying (within-subjects) covariates (e.g. age, weight, income, nutrient intake).

When each outcome is measured at the same times, then the resulting data is described as balanced data. Once measurements are made on different subjects at different times and/or there are missing data, and then the resulting data is described as unbalanced data.

The main longitudinal studies characteristics are as follow:

- Correlated outcomes,
- Unbalanced data set,
- Missing observations.

Therefore, the analysis of longitudinal data should deal with subject correlation, the observations that are at unequal time intervals and the missing observations.

Repeated measures analysis is used to analyse longitudinal or repeated measures data for a balanced study, once all subjects are measured at equal time points and

there are no missing values in data set. It is infrequent to obtain balanced data in longitudinal studies therefore it has been found necessary to develop other techniques which can deal with unbalanced data. The main techniques in use are General Linear Mixed Model and General Estimating Equations.

2.2.2. Linear model for longitudinal data:

For a longitudinal study in which the outcome is a continuous measurement, we consider three general models:

1- population average or Marginal Models,

2- Transition models,

3- Subject-specific models (random effect),

2.2.2.1. Marginal Models:

When a population is of interest, the most suitable model to fit is a Marginal Model. In Marginal Models, the population-averaged outcome is modelled as a function of the covariates. These are known as population-averaged models because the regression coefficients are interpreted for the population rather than for individuals.

In the Marginal Models, the regression of the outcome on explanatory variables is modelled separately from within-subject correlation.

2.2.2.2. Transitional Linear Models:

Once the time varying is important, models for the conditional distribution of Y_{ij} given $\sum_{k=l}^{j-l} Y_{i(j-k)}$ may be more suitable. These are also called conditional models.

A linear model for the conditional mean of Y_{ij} given the observed value $Y_{i(j-1)}$ of the outcome instantly preceding $\mathbf{Y_{ij}}$ is commonly named a first-order autoregressive model. One can also have second-order, third order or higher order autoregressive models (Diggle, Liang, & Zeger 1994).

2.2.2.3. Random effects Models:

A simple model is the fixed effect model with independent errors. In a fixed-effects model the levels of any given factor are fixed.

The model can be written as:

$$\mathbf{Y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \quad , (i = 1,2,3,...n)$$

$\mathbf{Y}_i$ = Outcome for subject i,

X_i = Covariate value for subject i.

The errors are assumed to be independent with constant variance, i.e. $Var[\,\boldsymbol{\varepsilon}_i\,] = \sigma^2$. It is also assumed that the p-dimensional vector of fixed effect parameters $\boldsymbol{\beta}$ is constant. Interest centres on estimating $\boldsymbol{\beta}$ and σ^2.

Some times the effects are not fixed but random, in which case we would like to construct a random effects model. For example, if the observations are not obtained by simple random sampling, but come from a cluster or multi-level sampling design, then random effects models are often useful.

A factor must be treated as a random effect if the factor levels present in an experiment are a random sample from some population of possible levels. In the case of a single random factor the model might be:

$$\mathbf{Y_{ij}} = \mu + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij}$$

Where

$i = 1,2,...n$

$j = 1,2,...n_i$

Also model assumptions are:

- μ is an unknown constant representing an overall mean,
- α_i is the random effect of unit i and $\alpha_i \sim N(0, \sigma_\alpha^2)$.
- The α_i's are random and independent of each other.
- The ε_{ij}'s are random, independent of each other and $\varepsilon_{ij} \sim N(0, \sigma^2)$.
- α_i and ε_{ij} are independent of each other.
- $Var(Y_{ij}) = \sigma_\alpha^2 + \sigma^2$
- $E(Y_{ij}) = \mu$

In this model, it is important to estimate both σ_a^2, σ^2 and α_i. A model for a data set with both Fixed effects and Random effects is called a Mixed effect model. Random-effects and Mixed-effects models are also suitable for the study of variation in an individual, so these models are sometimes called subject-specific models.

2.2.3. General Linear Mixed Model:

General Linear Models with fixed and random effects are known as General Linear Mixed Models and these have been discussed widely in recent literature. Some useful references are (Breslow & Clayton 1993;Diggle, Liang, & Zeger 1994;Searle, Casella, & McCulloch 1992).

Now $Y_1, Y_2, ..., Y_n$ are n variables each is dependent on the values for each of p independent variables $(x_{11}, x_{12}, ..., x_{1p})$, $(x_{21}, x_{22}, ..., x_{2p})$, ...,$(x_{n1}, x_{n2}, ..., x_{np})$.

The normal linear model with fixed effects is written as follows:

$$y_i = \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i$$

$$i = 1,2,...n$$

$$\varepsilon_i \text{ is } N(0, \sigma^2)$$

The error term (ε_i) is error term in the model. The fixed effect parameters are $\beta_1, \beta_2, ..., \beta_p$ and the error variance is σ^2.

The above normal linear model can be written again, in matrix form, as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & . & . & . & x_{1p} \\ x_{21} & x_{22} & . & . & . & x_{2p} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ x_{n1} & x_{n2} & . & . & . & x_{np} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

This formula is written again as follow:

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$$

Where

- The outcome vector is $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n)'$,
- X is the matrix of X_{ij}'s,
- Further, the vector of fixed effect parameter is $\beta = (\beta_1, \beta_2, ..., \beta_p)'$,
- The vector of errors is $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$
- and where I_n is the $n \times n$ identity matrix .

Although the General Linear Model is useful, often the assumptions about the ε distribution are too restrictive (Searle 1971).

Mixed-effect models are extensions of general linear models that include additional random-effect terms, and are suitable for longitudinal data, when measurements are correlated over time on the same individuals.

2.2.3.1. The Linear Mixed Models:

Mixed effect models are used to account for correlation between a response and covariates in data that are grouped according to one or more factors(Pinheiro & Bates 2000).

Mixed-effects Models supply a powerful and flexible tool for the analysis of balanced and unbalanced data. These data arise in different field of investigation and are characterized by the presence of correlation between measurements within the same group(Cnaan, Laird, & Slasor 1997).

The term longitudinal data is to mean that each subject is measured repeatedly on the same response at several different times. The main aim is in characterising the way the response changes over time.

An example of a longitudinal study is looking for the effects of starting to drink and giving up drinking on health.

The advantage of this study over many studies in this area is that the modelling enabled the authors to account for both, the age at starting and giving up drinking, in addition to other important covariates, and thus give a more comprehensive picture than previous studies which approached only one aspect or the other of the problem.

One more example is a study of stress and immune outcome in mothers with and without normal weight infants(Gennaro et al. 1997).

Two groups of mothers were measured for stress and various immune function markers at delivery of the infants and then at one, two and four months after delivery.

The data for each response were analysed separately, using the balanced and complete repeated measure data for four measurements.

The mothers without normal birth weight infants had increased anxiety and decreased lymphocyte proliferation as well as decreased in percentage of some immunology cell subsets.

Longitudinal models of these markers as both linear or a quadratic function of time showed that resolution of immune-suppression of pregnancy was substantially faster in mothers without normal birth weight infants than mothers with normal birth weight children, although neither group archived normal levels by four months.

2.2.3.2. Theory of Mixed Models:

An overview of a likelihood approach is provided to General Linear Mixed Models. Some statistical analyses, such as repeated measures and random effects, are unified by General Liner Mixed Model. A linear relation for unobserved multivariate normal random variables is a basic assumption in this model. Further information is in(Diggle, Liang, & Zeger 1994;Pinheiro & Bates 2000).

The General Linear Mixed Model can be written as follows:

$$\mathbf{Y}_i = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

$$\mathbf{b}_i \sim N(0, \psi)$$

$$\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2 I_n)$$

$$i = 1,2,\ldots n$$

The above formula is the same as the General Linear Model formula except for Z (the known design matrix) and $\mathbf{b}$ (unknown random effects parameters vector). The Mixed Model includes $\boldsymbol{\beta}$, as a vector of fixed-effects parameters and $\mathbf{b}_i$ as random-effects parameters, where:

- $\mathbf{Y}_i$ is the $n_i \times 1$ outcome vector for observations in the i^{th} group.
- X_i is the $n_i \times p$ model matrix for the fixed effects for observations in group i.
- $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed-effect coefficients for the population.
- Z_i is the $n_i \times q$ model matrix for the random effects for observations in group i.
- $\mathbf{b}_i$ is the $q \times 1$ vector of random-effect coefficients for group i.
- $\boldsymbol{\varepsilon}_i$ is the $n_i \times 1$ vector of errors for observations in group i.
- ψ is the $q \times q$ covariance matrix for the random effects.
- $\sigma^2 I_n$ is the $n_i \times n_i$ covariance matrix for the errors in group i .

For further information about the Mixed Model, see (Searle, Casella, & McCulloch 1992) .

As one assumption is normal distribution for $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$ with:

$$E\begin{bmatrix}\mathbf{b}_i\\ \boldsymbol{\varepsilon}_i\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}$$

$$Var\begin{bmatrix}\mathbf{b}_i\\ \boldsymbol{\varepsilon}_i\end{bmatrix}=\begin{bmatrix}K & 0\\ 0 & R\end{bmatrix}$$

Hence,

$$V(Y) = ZKZ' + R$$

This is a general description of the mixed model. A random effects model is a particular case when Z consists of dummy variables, K are variance components in a diagonal structure and R is $\sigma^2 I_n$, where I_n illustrates the identity matrix. With $Z=0$ and R is $\sigma^2 I_n$, on the other hand we recover the General linear model as a further particular case.

2.2.4. Generalized Estimating Equation (GEE):

The method of Generalised Estimating Equation (GEEs) was improved to extend the GLM to accommodate correlated data and has been used successfully by researchers in several fields.

The GEE approach is a common method for fitting models to data involving repeated measurements on the same subject. The outcome may be discrete or continuous (Zeger & Liang 1986).

This method permits the researchers to describe for within-subject correlations, among repeated measurements on the same subject. Different subjects can have different numbers of repeated measurements.

In GEE, the correlation between measurements on an individual subject is summarised in the Working Correlation Matrix $R(\alpha)$.

The matrix is determined by the vector of parameters α, which can take various forms of structures. The method estimates parameters of model by repetitively solving a system of equations based on quasi-likelihood distributional assumptions. The user can choose from several model forms by describing a link function; therefore the model form can be logistic, log-linear, or linear.

While modelling longitudinal data, the primary objective of regression analysis is to describe the relationship between the expected value $E(Y)$ of the outcome variable Y and the covariates $X_1, X_2, ..., X_p$. Modelling the correlation structure is of less importance, nevertheless it is necessary to take into account any within-subject outcome correlation when making statistical inferences about the regression coefficient(Nicholas & Stuart 1999;Zorn 2001).

Some of the most used within-subject correlation matrices are as follows:

- **Independence:** repeated observations are independent,

$$R = I = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

- **Unstructured:** All the measurements are correlated but the correlations are unknown;

$$\text{Corr}(Y_{ki}, Y_{kj}) = \alpha_{i,j}$$

$$R(\alpha)=\begin{bmatrix} 1 & \alpha_{1.2} & \dots & \alpha_{1,n} \\ \alpha_{1.2} & 1 & \dots & \alpha_{2,n} \\ . & . & \dots & . \\ \alpha_{1.n-1} & \alpha_{2,n-1} & \dots & \alpha_{n-1,n} \\ \alpha_{1,n} & \alpha_{2,n} & \dots & 1 \end{bmatrix}$$

- **Exchangeable:** correlation between any two observation of an individual is the same.

Where

$$Corr(Y_{ki}, Y_{kj}) = \alpha$$

and

$$R(\alpha)=\begin{bmatrix} 1 & \alpha & \dots & \alpha \\ \alpha & 1 & \dots & \alpha \\ . & . & \dots & . \\ . & . & \dots & . \\ \alpha & \alpha & \dots & 1 \end{bmatrix}$$

Auto regression of first order [AR (1)]: The measurements are taken in possibly unequal intervals and the assumption is that measurements near in time are more highly correlated.

Where

$$Corr(Y_{k,i}, Y_{k,j}) = \alpha^{|i-j|}$$

and

$$R(\alpha)=\begin{bmatrix} 1 & \alpha^{1} & \dots & \alpha^{n-1} \\ \alpha^{1} & 1 & \dots & \alpha^{n-2} \\ . & . & \dots & . \\ . & . & \dots & . \\ \alpha^{n-1} & \alpha^{n-2} & \dots & 1 \end{bmatrix}$$

2.3. Missing Data:

Not long ago, methods that were available to analyse longitudinal data had a major drawback. If one of the repeated measurements was missing, then all other available data for that subject also were excluded from the analysis.

To solve this problem imputation methods for missing data have been improved(Little & Rubin 2002).

Now, with modern methods to analyse longitudinal data, such as Generalized Estimating Equation (GEE), subjects with incomplete data are not excluded from the analyses. If a subject is missing one or more repeated measurements, the remaining available data from the other measurement for that subject are used in the analyses. That means that, once more these type of methods for the analysis of longitudinal data are used, it is probably less necessary to estimate the missing data (Twisk & Vente 2002).

Missing data is a common happening for longitudinal studies because not all outcomes are observed due to study dropout, lost measurements, subjects cannot come to the clinic to be measured, and so forth.

As longitudinal studies are balanced and complete very rarely, then a subject usually can have a missing response at one point time and be measured at the next point time. Usually, missing observation data in the studies is non-ignorable because the reason for missingness depends on the missing values themselves.

As an example, medication side effects may make the patients worse and thereby affect patient participation.

According to Diggle et al (1994), in Chapter 11, gives a good discussion of missing values in longitudinal data.

You suppose that you want to take measurements $Y_{i1},...,Y_{in}$ for the i^{th} subject. Whenever Y_{ij} is missing for all $j \geq k$, these missing values for Y_{ij} are called dropouts. In other way, missing values that happen intermixed with non-missing values are intermittent missing values.

2.3.1. Types of Missing Data:

2.3.1.1. Missing Completely at Random (MCAR):

Missing Completely at Random concerns to data where there is no dependence between the type of missing data and the outcome, or other measurements variable in the data. Missing Completely at Random is both missing and observed at random. Thus, missing data is very rarely Missing Completely at Random. For example, income in a family would not be dealt as a Missing Completely at Random when the family with low incomes were less likely to report their income than a family with higher incomes.

2.3.1.2. Missing at Random *(MAR):*

Data are Missing at Random if the probability of missing data on outcome is unrelated to the outcome value, after controlling for other variables in the analysis such as education, geographic location.

2.3.1.3. Not Missing at Random (NMAR):

Data not Missing at Random, or informatively missing arises if the probability of missing data on the outcome is related to the outcome value even if other variables are controlled in the analysis, this is the most difficult condition to model for adequately (Schafer 1997;Scheffer 2002).

2.3.2. Dealing with Missing Data:

As we are going to deal with missing data in a longitudinal study some of the available imputation methods to replace missing data will be discussed.

2.3.2.1. Imputation methods:

Imputation methods can be discussed in cross-sectional and longitudinal imputation methods(Twisk & de Vente 2002).

The cross-sectional methods include;

- Mean of series method,
- The hot-deck method,
- The cross-sectional linear regression method.

Longitudinal imputation methods include;

- Last value carried forward method,
- Linear interpolation methods,
- The longitudinal linear regression methods.

2.3.2.1.1. Cross-sectional imputation methods:

The mean of the available data for a variable at a time point is calculated. This mean value is imputed for the missing values.

The hot-deck imputation method is another approach. In this method, the average value or a random draw from similar cases is imputed.

A linear regression of the complete subjects is used in cross-sectional regression methods with all available predictor variables at the time that the response variable Y was missing. This linear regression predicts a value for the response variable Y at that time point and the predicted value is used for the imputation of the missing results.

It could be that this approach is suitable in situations where the response variable is missing but not the predictor variables.

2.3.2.1.2. Longitudinal imputation methods:

Last Value Carried Forwarded (LVCF) is the simplest longitudinal imputation methods. In this method the value of a variable at time t is imputed for a missing value at time $t+1$, our assumption being that the variable is almost constant over time.

The linear interpolation imputation method is another longitudinal imputation method.

In this method a missing value at time t is imputed by the average of the values at time $t-1$ and $t+1$.

The population longitudinal regression imputation method is based on a regression of Y on the previous measurement of outcome, on the all predictor variables and on time.

2.3.2.1.3. Multiple imputation method:

Multiple imputations are a method to deal with data sets with missing values. Multiple imputation methods fill each missing value in several simulated values. Subsequently multiple imputations create several believable complete data sets. Afterward, the standard methods can analyse these complete data sets and the results of the whole analysis are combined to produce a single inference. The variance between the sets of imputations represents dubiety about the 'true' values of the missing data (Schafer 1997).

2.3.2.2. Advantages of Imputation:

Bias is minimised by imputation. Imputation permits for analysis using a complete data set, hence using usual software and techniques, therefore that standard analysis can then proceed.

2.3.2.3. Disadvantages of Imputation:

Imputed data is not real data; this uncertainty must be reflected in the variance estimates. Single imputation often gives reduced variance estimates; Hence imputation is not reflecting the true dubiety.

2.4. Software for the analysis of longitudinal data:

Longitudinal data analysis can be more complicated than for cross-sectional data, hence software design for the modelling of longitudinal data can be more complicated. In this section, some of the important software programs and packages that can be used for longitudinal analysis are reviewed but the detail is not discussed here.

2.4.1. TERRACE:

James Hilden Minton wrote this software for his PhD thesis, which is multilevel research software. It can be found at this URL address:

http://www.stat.ucla.edu/theses/

2.4.2. NLME:

The S Language has been developing during several years for longitudinal data modelling with linear or non-linear Mixed Models. The NLME library is compatible with S-Plus version 3.4 and higher for the analysis of both linear and non-linear Mixed effect Models.

It is available on the following webpage, written by (Pinheiro & Bates 2000).

http://cm.bell-labs.com/cm/ms/departments/sia/NLME/

2.4.3. BUGS:

BUGS is computer software which analyses complex statistical models using Markov Chain Monte Carlo methods with emphasis on the Monte Carlo methods, and ability to analyse a great variety of multilevel models.

For further details can be seen:

http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

2.4.4. OSWALD:

A statistics team in Lancaster University developed Oswald, which is a suitable function for S-Plus in analysing longitudinal data, the details are in follow URL address:

http://www.maths.lancs.ac.uk/Software/Oswald/

Mixed effects Models, and many other possible options are included in OSWALD.

2.4.5. The packages available for the GEEs:

2.4.5.1. SAS:

SAS/STAT release 6.12 is the SAS version that used for the evaluation of GEEs. More information about SAS is available in the SAS URL address;

http://www.sas.com

2.4.5.2. STATA:

Stata version 5.0 and higher evaluates GEEs. The *xtgee* command can fit GEE models in Stata, which is part of the *xt* cross-sectional time-series analysis. More information is available in the web page;

http://www.stata.com

2.4.5.3. SUDAAN:

The SUDAAN version 7.5 and higher evaluates GEEs. More information is available on the following web page;

http://www.rti.org/sudaan/home.cfm

GEE models for binary and continue variables are fitted by PROC LOGISTIC, PROC MULTILOG, and PROC REGRESS in this package (Shah, Barnwell, & Bieler 1997).

2.4.5.4. S-PLUS:

S-Plus versions 3.4 and higher evaluate GEEs. More information is available on the following web page;

http://www.mathsoft.com/splus,

and

http://www.insightful.com/downloads/libraries/default.asp

Also some packages such as YAGS or YET are available to implement GEEs and it can be added as a library to S-Plus.

The library is available on the web:

http://www.biostat.harvard.edu/~carey

Those are also available on web page of Brian Ripley:

http://www.stats.ox.ac.uk/pub/SWin/

The Gaussian (normal), Bernoulli/binomial, Poisson, and Gamma families of distributions are supported by SAS, Stata and S-Plus.

The Gaussian and Bernoulli/ binomial distributions are supported by SUDAAN. Many of the packages are being extended to support more distributions in later releases (Nicholas & Stuart 1999) .

2.5. Software available for Missing Data:

2.5.1. SOLAS (Version 3):

Import Data from various file including SAS, SPSS, S-Plus and many others is possible. Some imputation methods such as the group means, last value carried forward and hot deck are available in this software.

SOLAS is designed specifically for the analysis of datasets with missing observations and is available on following web page:

www.statsol.ie/solas/solas.htm

2.5.2. SPSS:

In SPSS, List wise analysis, all value analysis; regression imputation and EM imputation are available.

2.5.3. S-PLUS:

New version (6.0) of S-Plus supports missing data models, using different imputation methods such as EM.

2.5.4. SAS:

SAS version 8.2, for multiple imputation procedure introduces the experimental MI and MIANALYZE and as these are bundled with SAS/STAT, we do not need to more installation (Scheffer 2002) .

2.5.5. Statistical Model:

ALSPAC is an observational longitudinal study, which started in 1991. Although ALSPAC is still going on, this thesis deals with data collected over a period of seven years, covering children up to seven years of age and their mothers during pregnancy.

In this longitudinal observational study several repeated measurements were carried out in a period of times on subjects.

The purpose of this thesis is to model the changes in mean haemoglobin levels during pregnancy and childhood and find the effects of some important covariates on those haemoglobin levels. This relationship will be analysed using all available longitudinal data.

Previously, one of the limitations in methods was that all available longitudinal data are not used in the analysis.

With the development of techniques in statistics, such as Generalised Estimating Equation (GEE) (Liang & Zeger 1986;Zeger & Liang 1986;Zeger & Liang 1992) and General Linear Mixed Models (GLMM), the analysis of longitudinal data using all available data is possible.

In this study the following statistical model is used to analyse the data:

$$Y_{it} = \beta_0 + \sum_{j=1}^{J} \beta_{1j} t^j + \sum_{k=1}^{K} \beta_{2k} Z_{ikt} + \sum_{m=1}^{M} \beta_{3m} G_{im} + \varepsilon_{it}$$

Where:

Y_{it} = Observations of Subject i at time t where $1 \leq t <$ number of measurement

β_0 = Intercept

β_{1j} = Regression coefficient of time

t = Time

J = Order of polynomial of time

Z_{ikt} = Time-dependent covariate k of subject i at time t

β_{2k} = Regression coefficient of time dependent covariate k

K = Number of time dependent covariates

G_{im} = Time-independent covariate m of subject i

β_{3m} = Regression coefficient of time-independent covariate m

M = Number of time-independent covariates

ε_{it} = Measurement error of subject i at time t.

In this model one of the interesting coefficients are β_{1j} because the regression coefficients illustrate the relationship between the longitudinal change of the haemoglobin level Y_{it} and the time. The covariates such as dietary intakes are included in the model as time-dependent-covariates, whereas gender and ethnic are time-independent covariates in the model.

2.6. Interpolation method:

In this section, methods for using polynomials for interpolation will be presented. An interpolation problem happens when a function's value is required at a point where data is not available. Finding the best function to fit to the data is needed. Interpolation is the method applied to the problem of finding this function when unavailable data is inside the range of the given data.

Interpolation methods are divided into two main categories (Press et al. 1992):

- **Global interpolation:**

A single equation to fit all the data points is constructed by these methods, which is usually a high degree polynomial. Although these methods result in smooth curves, as they suffer from oscillation and overshoot at intermediate points, they are not always suitable for applications.

- **Piecewise interpolation:**

A lower degree polynomial between each pair of known data points is constructed by these methods. First-degree polynomial is called linear interpolation. Also second and third degree polynomials are called quadratic and cubic splines respectively. The curves become smoother the higher the degree of the spline.

For obtaining a smoother curve, cubic splines are usually recommended. It is supposed to be well behaved and continuous up to the second order derivative at the data points. Cubic splines one less prone to oscillation or overshoot than polynomial equations.

2.6.1. Linear spline interpolation:

Firstly a linear spline is fitted to the data as follow;

$$\{(x_0, y_0),(x_1, y_1),.....,(x_{n-1}, y_{n-1}),(x_n, y_n)\}$$

It is supposed that data is in ascending order and then the linear splines are given by $y_i = f(x_i)$.

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0), \qquad x_0 \le x \le x_1$$

$$= f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1), \qquad x_1 \le x \le x_2$$

.

.

.

$$= f(x_{n-1}) + \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x - x_{n-1}), \quad x_{n-1} \le x \le x_n$$

The formula (2.1) is the slope between x_{i-1} and x_i;

$$\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \tag{2.1}$$

Now using quadratic splines is an improvement.

2.6.2. Quadratic Splines:

The data between each consecutive pair of points is approximated by a quadratic polynomial.

For fitting a quadratic spline through the data, the splines are given by;

$$f(x) = a_1x^2 + b_1x + c_1, \qquad x_0 \le x \le x_1$$

$$= a_2x^2 + b_2x + c_2, \qquad x_1 \le x \le x_2$$

.

$$= a_n x^2 + b_n x + c_n, \qquad x_{n-1} \le x \le x_n$$

Then we can find the coefficients of these quadratic splines.

2.6.3. Cubic Splines:

As linear splines are not smooth and cannot be fitted to highly curved functions well, it is more common for cubic splines to be used in practice.

Generally, a function S is a spline with k degree on $t_1 < t_2 < \cdots < t_n$ if

(1) $S \in [t_1, t_n]$;

(2) $S^{(j)}$ is a continuous function on $[t_1, t_n]$ where $s^{(j)}$ is the jth derivative and $j = 0,1,2,\ldots k-1$;

(3) S^j is a polynomial of degree $\le k$ on each interval $[t_i, t_{i+1}]$.

If $k = 3$, the spline is called a cubic spline. The function S is as follow:

$$S(x) = \begin{cases} S_1(x) & t_1 \le x \le t_2 \\ S_2(x) & t_2 \le x \le t_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ S_{n-1}(x) & t_{n-1} \le x \le t_n \end{cases}$$

Where S_i is a cubic spline that $t_i \in [t_i, t_{i+1}]$ and $S(t_i) = y_i \;\; 1 \le i \le n_i$.

Other conditions are as follow;

There are another $n-2$ equations for slopes at the interior points $(t_2, t_3, ..., t_{n-1})$.

$$\lim_{x \to t_i^-} s^k(t_i) = \lim_{x \to t_i^+} s^k(t_i) \qquad k = 0,1,2$$

Once the second order derivatives of the splines at the end points are zero the spline is called a natural cubic spline (Cheney & Kincaid 1985).

$$s''(t_1) = s''(t_n) = 0$$

2.7. Maximum likelihood estimation (ML):

The maximum likelihood principle is a very popular approach used to obtain practical estimators .

When the observations in a random sample, $y_1, y_2, ..., y_n$ are drawn independently from a distribution, influenced by an unknown parameter θ, then $f(y_1, y_2, ..., y_n \mid \theta)$ is proportional to the probability of obtaining the data given the value of θ. This expression is called the likelihood of the data .

It can be viewed as the chance of obtaining the sample data, given θ. Since θ is usually unknown, it must be estimated from the data.

As an estimate of θ, the value $\hat{\theta}$ is selected such that when evaluated at $\hat{\theta}$, the expression for the likelihood of the sample reaches a maximum. The process of finding estimated values of unknown parameters is maximum likelihood estimation. Estimates obtained in this way are known as maximum likelihood estimates.

2.8. Restricted maximum likelihood estimation (REML):

One of the preferable methods for estimating variances and covariances is the Restricted Maximum Likelihood (REML) method. Using this method, the

variance-covariance components are estimated by Maximum Likelihood averaged over all possible values of the fixed effects. The REML estimated variance components are not affected by the fixed effects, this means variance estimates are invariant to the parameterisation of the fixed effects while ML estimates would be affected by the parameterisation. For example, for estimating σ^2 from a normal distribution, where $y_i \sim N(\mu,\sigma^2)$ for $i=1,2,...n$ and $\bar{y}=\sum_{i=1}^{n} y_i/n$ and $S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$, then $\sigma^2=S_{yy}/(n-1)$ by REML and $\sigma^2=S_{yy}/n$ by ML (Diggle, Liang, & Zeger 1994;McCulloch & Searle 2001).

2.9. Reference Curves:

A useful screening tool in medical problems is reference curves. Reference curves can identify unusual subjects, when the value of a particular measurements, lies in one tail of the reference distribution (Cole & Green 1992).
A reference curve is a graph that illustrates how the distribution of some interesting variable changes with time.
For constructing the reference curve, the data are assumed to be representative of the reference population, which are normal. Also observations are supposed to be abnormal or unusual if they lie outside the reference curves.
A range of measurements symmetric around the median and bounded by the $(100-\alpha)/2$ and $(100+\alpha)/2$ centiles is a definition of an $\alpha\%$ reference interval $(0<\alpha<100)$.
Wright and Royston describe and compare several methods that are available to calculate age-specific curves using real data sets (Wright & Royston 1997).
Reference curves are usually chosen from a symmetric subset of the 3^{rd}, 5^{th},10^{th}, 25^{th} 50^{th},75^{th} ,90^{th} 95^{th} and 97^{th} centile.
Once a reference curve is constructed using values from a normal or healthy population, it can be used to assess the health of future subjects.

An individual's result is classified as abnormal if it lies below a particular centile (e.g. 5^{th}, 10^{th}) or above a particular centile (e.g. 90^{th}, 95^{th}).

It is very common for the distribution of a variable to be affected by some factors such as sex, age and ethnicity. For many medical studies, major influence on the variables is the age at which an observation is made. The best example of age related reference curves might be growth curves.

According to the reviews, the LMS method that was proposed by (Cole 1988) is widely used to obtain reference curves, further details are as follows.

2.9.1. The LMS method:

The varying distribution of a variable with age is summarized by the LMS method according to three time dependent natural spline curves, for the median (M), the coefficient of variation (S) and the skewness expressed as a Box-Cox power (L).

The data must be grouped by age and interpolation between age groups is required.

The LMS method has been used with the maximum penalized likelihood (Cole & Green 1992). The use of the maximum penalized likelihood approach makes it possible to provide smooth estimates of the L, M and S curves directly.

These three curves (L, M and S) are fitted as cubic splines by nonlinear regression, using the maximum penalized likelihood, and the extent of smoothing required is controlled by equivalent degrees of freedom parameters (Cole & Green 1992).

The fitting process ensures that the L, M and S values change smoothly with age, therefore they can be smooth curves plotted against age. The L, M and S values then provide the reference curve using the following formula:

$$C_{100\alpha}(t) = M(t)\left[1 + L(t)S(t)z_{\alpha}\right]^{\frac{1}{L(t)}}$$

Where

$C_{100\alpha}(t)$ = the reference curve plotted against t (t is age),

z_α = the normal equivalent deviate for the centile (α=0.97, z_α=1.88),

$L(t)$, $M(t)$, and $S(t)$ are the fitted smooth curves plotted against age.

The data can be converted to standard deviation (SD) scores directly by the LMS method, using the following formula (Cole & Green 1992) ;

$$Z = \frac{\left[Measurement(t)/M(t)\right]^{L(t)} - 1}{S(t)L(t)}$$

where measurement is the individual's value at age (t), and L (t), M (t), and S (t) are values of the smooth curves at the particular age.

Reference curves are constructed for variable measurements and curves for the 3rd , 10th , 25th , 50th , 75th , 90th and 97th centiles are generated.

Reference curves were obtained using Cole and Green's LMS computer program Copyright 1998, Institute of Child Health (Cole & Green 1992).

2.9.2. Conditional Reference Curve:

This study draws a clear distinction between reference curves used for cross-sectional data and longitudinal data. While a cross-sectional reference curve can show whether a measurement at a time point is normal compared to a comparable population at the same time, a longitudinal reference curve can report whether an individual's measurement at one time point is consistent with that other same person's results.

The typical example is that when a patient returns for a new measurement (blood pressure) at the hospital, according to the new measurement and the record of earlier measurements, the physician wishes to know that variation is normal. The cross-sectional reference curve may be very misleading when considering the consistency of an individual's data.

Therefore, using other methods to consider this curve rather than the cross-sectional curve should be beneficial.

Constructing a reference curve using the previous measurements to predict the current measurement is one of them (Scheike, Zhang, & Juul 1999) .

Cross sectional reference curves have been developed into conditional reference curves to use for monitoring longitudinal data. A conditional reference curve is constructed from outcomes at time t, which are predicted from outcomes one previous time $(t-1)$;

$$outcome_t = \alpha_t + \beta_t outcome_{t-1} + error$$

Where

α_t = The intercept

β_t = The slope coefficient

Error term ~ $N(0, \sigma_t^2)$

Furthermore, this conditional reference curve can be extended to give an other conditional reference curve to include all previous times point (Cole 1994) .

The above formula is extended, as follow;

$$outcome_t = \alpha_t + \alpha_{t-1} outcome_{t-1} + \alpha_{n-2} outcome_{t-2} + \ldots + \alpha_1 outcome_1 + \varepsilon_t$$

Where

$\alpha_{t-1}, \alpha_{n-2}, \ldots, \alpha_1$ =The slope coefficients

α_t = The intercept

Error term ~ $N(0, \sigma_t^2)$

CHAPTER 3

Factors associated with mother's iron status during pregnancy

3.1. Background:

As discussed in the previous chapter, The Avon Longitudinal Study of Parents and Children (ALSPAC) is a cohort study investigating a huge range of factors influencing the heath and development of infant and children.

All women during pregnancy residing within the Avon area of Southwest England with an expected date of delivery between April 1991 and December 1992 inclusive were eligible to take part in this study.

The participation of approximately 85% of the eligible population, resulted in a cohort study of around 15000 pregnancies (Rogers, Emmett, & The ALSPAC Study Team 1998).

From this population, haemoglobin levels were obtained from a 10% random sample of mothers during pregnancy whose babies were born during the last six months of the survey, Children in Focus (Sherriff, Emond, Hawkins, Golding, & the ALSPAC Children in Focus Study Team 1999) .

The data were obtained from the records that were prepared by sending a series of self-completion postal questionnaires to the mothers. Haemoglobin measurements were obtained from mothers at different times between first week and the fortieth week. A small number of haemoglobin measurements are available for each mother during pregnancy (Rogers, Emmett, & The ALSPAC Study Team 1998).

As the preparation of iron supplements to mother during pregnancy is one of the most important measurements in public health approach, the history of iron supplement in mothers was observed from the questionnaires that were distributed to the mothers during pregnancy.

Two of the questionnaires were sent to mothers at fixed time points, one between 18 and 20 weeks of pregnancy and another at 32 weeks of pregnancy, relating to the second and third trimesters of the pregnancy (Sidebotham, Golding, & The ALSPAC Study Team 2001).

The aim of this part of study is to model the haemoglobin levels during pregnancy in the three groups of mothers (as defined by their use of iron supplements) and the effect of some important covariates on haemoglobin levels.

3.2. Function Fit to the Data:

As you can see in Figure 3.1, a scatter diagram is used to show changes in haemoglobin levels in mothers during pregnancy against time of measurement (week). The diagram shows that the data set suffered from the problem of data sparsity before 5 weeks, between 17 and 22 weeks, and after 40 weeks of pregnancy.

If the data set is sparse in some weeks, consisting only of a few samples, then a judgment of how haemoglobin levels change during pregnancy in these weeks may become more difficult and inaccurate.

The first step in this data analysis is to plot the data with a plot of the means at each time point superimposed, which gives a clearer picture of the data as appear in Figure 3.2.

The next step in this chapter is to fit a function, which is represented by a polynomial function. Polynomials of increasing degree were fitted to the data using the method of General Linear Mixed Models and are illustrated in Table 3.2.

Describing such trends with a suitable polynomial is complicated because there are so many possible parameters. The polynomial degree, and the number of coefficients, can be as large as we need. In fact, we can find many such polynomials, of higher and higher degree.

It is clear that the most important decision in polynomial regression is the choice of polynomial degree; this choice specifies the form of the curve fit. We have fitted a quadratic, cubic and higher degree, and then we attempted to see if we could reduce the model by a few terms.

For the global polynomial fitting technique in this data set, the fitted models become more complicated with an increase in degree.

Polynomial regressions were fitted successively starting with the quadratic degree in Group 1. These were tested in degree, when the highest degree term was determined then others lower degree terms were included.

Then criteria such as log-likelihood and AIC can be used to decide which model is the best model. The model preference is in favour of small AIC and big log-likelihood. As you can see in Table 3.2, both these criteria show a strong preference for cubic polynomial over quadratic, 4^{th} and 5^{th} order, whereas the significance of some or all of the terms in the models indicate that these should be a better fit.

In this case, the polynomial did not give any meaningful approximation of the relationship between haemoglobin levels and time. Also, it was difficult to determine whether a curvilinear relationship with an exact quadratic, cubic or higher order is a suitable model statistically.

Therefore, a meaningful fit should be provided by a low order polynomial and hopefully no deficiency of fit will be found.

We focus our attention on curve fitting methods, which were able to provide clear functions from the data. In this study to find an adequate model, we have used the cubic spline algorithm, which applies a piece-wise polynomial fitting technique.

Polynomial regression undertakes to fit a single equation to the data set, while spline interpolation fits a sequence of curves to segments of the data. This technique is rather simple minded, but it does ease the problem generated by polynomial regression.

Figure 3.3 shows a scatter plot of haemoglobin levels against time; the curves have been fit by cubic spline, cubic polynomial and 4^{th} order polynomial on mother's haemoglobin levels during pregnancy in Group 1, which are comparable with each other.

One reason why the spline reconstruction generally performs better than the polynomial fitting procedure is that fitting the data to a more complex function with more parameters will almost certainly improve the fit. Therefore, the spline approach, which has more parameters than a polynomial with the same order, dose appear to give better description of the haemoglobin levels in mothers during pregnancy.

The log-likelihood test can be used to test the difference between two or more fitted mixed effects models by using an ANOVA procedure. As in the *lme* (function for fitting linear mixed-effects models), the default method of estimation is restricted maximum likelihood (REML). However, likelihood comparisons between REML fits with different fixed effects model structures are not meaningful, therefore using maximum likelihood (ML) to refit the objects before using ANOVA is necessary.

These models could have been fitted to the whole data set, but instead it was decided to fit separate models to the data for mothers separated according to their use of Iron supplements.

3.3. Statistical Methods:

The statistical model used to analyse these data incorporates a cubic spline to represent the complex mean function. The parameters of this model are estimated using General Linear Mixed Model (GLMM).

A spline is a piecewise polynomial function, where different sections of polynomials are fitted together smoothly. The locations of the breaks are called knots. We need to choose the required number of knots and their positions. In this case all the pieces of curves between the knots are produced from polynomials of the same order.

For choosing the number and position of knots there are a number of possible options. We would like to place knots at points in the data where we expect significant changes in the relationship between the predictor and to avoid position where data are sparse.

As the log-likelihood and also AIC (Akaike's Information Criterion) can be used for a data-based selection of number of knots then the value of number of knots that gives the largest log-likelihood ratio and lowest AIC can then be selected.

For finding a satisfactory fit, several knots in different positions in the period measurements of the study were tested. The following turned out to be the three best choices for fitting the model to the data.

Firstly, a cubic spline model was built that included 11 (interior) knots within the period from 6 weeks to 36 weeks with the knots all equally spaced (each 3 weeks).

Then the second model included 7 knots from 6 to 34 weeks, with the knots all equally spaced (each 4 weeks).

Finally, knots were chosen at 12,18,24,32 and 36 weeks because these were the most important points in pregnancy for the purposes of this study. The second and fourth knots were weeks, at which mothers started taking iron supplement and the rest of them, are weeks in which blood samples were taken from mothers at first, second and thirds trimesters of pregnancy. These three models were compared in all cases using likelihood ratio tests.

As Table 3.1 has shown, because the P-values for comparing these three models are more than 0.05, the models fitted with 11 and 7 knots are not significantly better than models with 5 knots therefore the number of knots of regression coefficients is fixed at five.

Some covariates were added to the basic spline model. These included mothers who were/were not vegetarian during pregnancy, who were /were not white, highest educational level of mothers, number of cigarettes smoked per day in 1st 3 months of pregnancy, number of cigarettes smoked per day in last 2 weeks of pregnancy, number of cigarettes smoked per day in 32 weeks of pregnancy, mother's parity (number of births up to and including the children in the study that we moved to a binary variable for parity, either the mother had had children previously or she had not), singleton/ multiple pregnancy and maternal age.

These mixed effect models were fitted with and without additive and interaction terms. These covariates were then included in the model and the statistical significance of the interaction term between the two variables (each covariate and time) was tested.

Suppose that we now want to know one of the questions of interest for haemoglobin level data is whether categorical covariates have different changing patterns.

We illustrate the effect of the interaction term in the model in Tables 3.9, 3.10, 3.11, 3.14, 3.15 and Figures 3.6 to 3.10.

As Figures 3.11, 3.13 and 3.15 present, around 95% of the standardized residuals are between –2 and +2, therefore the normality assumption is valid. Normal plots of these residuals in Figures 3.12, 3.14 and 3.16 confirm the normality as well.

3.4. Results:

The total number of mothers in this study was 1287. Fifty-seven percent of them never took iron supplement (Group 1), 19% had started to take iron supplements by 18 weeks (Group 2) and a further 24% started on iron supplements between 18 and 32 weeks (Group 3). Table 3.3 shows the distribution of covariates for the three iron supplement groups.

For a variety of reasons some mothers did not complete questionnaires, which included questions on ethnic origin, education, smoking and vegetarian situation, social and occupation levels. Therefore, as you can see in Table 3.3, there are some missing values for each covariate.

Association between iron supplement status and each covariates (education level, maternal age, ethnicity of mother, number of cigarettes smoked per day in 1st 3 months per day, number of cigarettes smoked per day in 2 last weeks of pregnancy, number of cigarettes smoked per day in 32 weeks of pregnancy, singleton/multiple pregnancy, mother's parity and vegetarianism of mothers) was assessed using a χ^2 test.

Of the 731 mothers in the first group 178 (25%) were in 'CSE' or 'VOC' level (low level of formal education) and 427 (59%) were in 'O' or 'A' level (middle level) and 118 (16%) were in degree level.

Of the women who took iron supplement by week 18, fifty-five (23%) completed the study in 'CSE' or 'VOC' level, a hundred fifty (64%) of them were in 'A' level or 'O' level and thirty-three (13%) also were in degree level of education. There was no information for ten mothers in this group.

In-group 3, majority of mothers (66%) were in 'O' and 'A' level of education, and mothers with 'CSE','VOC' and Degree level of education were 14%, 9% and 11% respectively.

Chi-square test showed that education level of mothers was significantly associated with iron supplement status of mothers.

Mothers with higher education were marginally more likely than other to take iron supplements (P=0.036).

Ninety eight percent, ninety six percent and ninety six percent of mothers' in-group one, two and three respectively were with white background, the remaining mothers were non-white in these groups. No significant association was found when tested by Chi-square.

It can be seen that the percentage of mothers who were not vegetarian is highest in non-iron supplement group (89%) and lowest percentage of cases were in-group three (79%).

Chi-square test showed that the vegetarian status of mother was associates with iron supplement group.

It would seem that vegetarianism increases using iron supplement in mothers during pregnancy.

There was no significant association between taking iron supplement and number of cigarette smoked per day in 1st 3 months of pregnancy.

In relation to number cigarette smoked per day in last 2 weeks of pregnancy, non-smokers were more than 80% in all groups.

Also there was no significant association between the use of supplement and the number of cigarette smoked in this group.

Number of cigarettes smoked per day in 32 weeks of pregnancy was not significantly different in three groups of mothers.

However, there was a strongly significant association between singleton/multiple pregnancy and taking iron supplement. Highest percentage of multiple pregnancy were in mothers who took iron supplements by 18 weeks.

Number and percentage of mothers in all three groups according to mother's parity also are shown in Table 3.3. There was no association between mother's parity and iron supplement groups. Percentage of mothers was lower in women with first-born children in all groups.

As can seen in Figure 3.2, variation in means of haemoglobin concentration are observed during pregnancy and also Figure 3.5 shows different in haemoglobin concentration in the three groups of mothers.

The means have shown in Figure 3.2 that, as time increased to the middle of pregnancy, haemoglobin concentration decreased and after that there was gradual increase in haemoglobin concentration up to the 40^{th} week.

Figure 3.5 illustrates that at the start of pregnancy, haemoglobin concentrations were generally highest in the non-iron group than the other groups but they were nearly the same in last weeks.

At the beginning of pregnancy, the lowest mean haemoglobin concentrations were in mothers who chose to take iron supplements early (by week 18).

The fitted mean haemoglobin concentration for group 2 begins to increase from about week 18.

Similarly, the average haemoglobin concentration for mothers' in Group 3 increased from around week 32 but decreased again around week 40.

In the group that never took iron supplement, haemoglobin concentration decreased continuously from beginning of pregnancy and reached a minimum haemoglobin level after week 30, then increased up to end of pregnancy.

This emphasises that, in this observational study, mothers with the poorest haemoglobin levels tended to start on iron supplements earliest, an understandable and sensible decision, which does however lead to difficulties of interpretation later.

The group of mothers who chose to start taking supplements after week 18 are the group whose haemoglobin levels are most worrying, since their mean haemoglobin level dipped to a very low level for a sustained period in weeks 20-

30. Perhaps it would be wise for women to start iron supplements earlier in pregnancy.

The parameters of spline function calculated at knots 12,18, 24, 32 and 36 weeks of pregnancy for group 1 are shown in table 3.5.

Tables 3.6 and 3.7 illustrate the cubic spline functions for mothers with iron supplementation.

In all cases but one, the parameter is significant indicating that 5 interior knots are required in the model. Although one parameter is not significant for the Group 2 model, the full set of knots was retained for comparability of results for the three groups of mothers.

The model parameter estimates with their standard error for each selected model are summarised in tables 3.5, 3.6 and 3.7 as well.

Because our interest is in whether age, vegetarian diet, education level, number of cigarettes smoked per day in 1st 3 months of pregnancy, number of cigarettes smoked per day in last 2 weeks of pregnancy, number of cigarettes smoked per day in 32 week of pregnancy, singleton/multiple pregnancy, mother's parity and ethnic background affect the mean response profiles or not, these covariates are added to the model one at a time and these results are shown in Table 3.4.

As shown in Table 3.4, mean changes in haemoglobin concentration during pregnancy were not significantly associated with maternal age and number of cigarette smoked per day in 32 weeks of pregnancy in three groups of mothers.

Also in Group 1, there were no significant differences between haemoglobin concentrations in mothers who were/were not white, mother's parity, singleton or multiple pregnancy.

In addition haemoglobin concentration were not significantly associated with education level but as you can see in Table 3.8 and 3.4, a significant negative effect was found for vegetarian mothers.

That is, haemoglobin levels were generally lower in woman who had been vegetarian in the past.

There was no significant interaction between vegetarians and time in-group one. Table 3.9 illustrates the cubic spline functions when modelled separately by number of cigarette smoked per day in last 2 weeks of pregnancy in-group one and there was significant effect for this covariate (P=0.0206).

However, according to Figure 3.6 and Table 3.9, the interaction between this covariate and time appeared significant at knots (12,18) positively and at knot (18,24) negatively (P-Value= 0.0134,P-Value=0.0096), indicating that mothers who smoked more than one in last 2 week of pregnancy had a more positive and then a negative profound effect on the haemoglobin levels at knots (12,18) and (18,24) respectively.

The purpose of producing Figure 3.6 is to illustrate the significant interactions in the model. The plot shows an interaction between numbers of cigarettes smoked per day in last 2 weeks of pregnancy and time (week). This might be a real effect but it cannot be estimated reliably because of the sparsity of the data in the interaction area, which has already been explained.

Figures 3.7 to 3.10 are produced to illustrate interaction between the covariates, such as number of cigarettes smoked per day in first 3 months of pregnancy, mother parity in Group 1, multiple pregnancy in Group 2 and vegetarianism of mother in Group 3 and time. The locations (weeks) and pattern of interaction are similar to Figure 3.6. According to Figure 3.6, the interaction may be due to the sparsity of the data around this area.

To assess whether the number of cigarette smoke per day in first 3 months of pregnancy was a significant covariate on cubic relationship, this covariate added to model. The results indicated in Table 3.10 and Figure 3.7 that, it was a positively significant covariate in the model (P-Value = 0.0196).

Also the number of cigarette smoke per day in first 3 months of pregnancy as an interaction term was considered on the model.

Figure 3.7 and Table 3.10 demonstrate that in this model tested, interaction between the number of cigarette smoke per day in first 3 months of pregnancy and

cubic spline contributed significantly to the model fit at knots (12,18), (18,24) and (32,36).

However, it is clear from Figure 3.7 and Table 3.10 that, there was a significant positive interaction between the numbers of cigarette smoke per day in first 3 months of pregnancy and time at knot (12,18) and a significant negative interaction between this covariate and time at knots (12,18) and (32,36).

That means, as time increased, haemoglobin concentration became progressively smaller in mothers who were smoker in first 3 months of pregnancy at knots (18, 24) and (32,36) and also haemoglobin levels in this group of mothers have increase over knot (12,18) significantly.

As shown in Table 3.11 and Figure 3.8, the effect of the interaction between mother's parity and time reach statistical significance.

Statistical interaction between mother's parity and time at knot (24,32) indicated that the effect of parity on the haemoglobin level is negative in-group one.

In Group 2, No significant differences were observed in education level, vegetarian statues of mothers during pregnancy, mother's parity and number of cigarette smoked per day in last 2 week of pregnancy but ethnic status, number of cigarette smoked per day in 1st 3 months of pregnancy and singleton /multiple pregnancy reported correlate significantly with haemoglobin levels.

As you can see in Table 3.12, haemoglobin concentrations were significantly higher in mothers who were white than the non-white in-group 2. There was no significant interaction between this covariate and time.

Table 3.13 shows haemoglobin concentration was significantly higher in mothers who did smoker per day in 1st 3 months of pregnancy compared to the remaining mothers in-group 2 but the interactions between number smoked in 1st 3 months of pregnancy and time were no significant.

In Table 3.14, the model demonstrated a significant negative relationship between multiple pregnancy and haemoglobin concentration in-group 2.

Our findings in Table 3.14 and Figure 3.9 also showed an interaction between singleton/multiple pregnancy and time in-group 2, indicating that among women

who were multiple pregnancy, changing in time had a more positive effect on haemoglobin concentration at knots(,12) ,(24,32) and (32,36) ,otherwise a more negative effect on haemoglobin concentration at knots (12,18) and (18,24).

As Table 3.14 and Figure 3.9 have shown, the interactions between multiple pregnancy and time may be due to having a small number of mothers with multiple pregnancies.

It should be better to work in large sample of mothers to investigate effect of twin pregnancy on haemoglobin level in pregnant women as a further work.

In mothers who started taking iron supplement by week 32, No correlation was found between haemoglobin concentration and ethnicity of mothers, mothers who were/were not vegetarian, number of cigarette smoked per day in 1st 3 months of pregnancy and number of cigarette smoke per day in last 2 week of pregnancy.

There was a significant interaction between mothers who were vegetarian in the past at the present and time at knots (18,24) in-group 3, which is given in Table 3.15 and Figure 3.10.

This interaction meant that women who were vegetarian in the past and at the present were negatively associated with time at knot (18,24).

Dividing mothers according to their highest education level in group 3, Haemoglobin level in mothers with a Voc level and mothers with O level (middle education groups) were significantly higher then other groups of mothers is shown in Table 3.16.

Table 3.17 shows haemoglobin levels were significantly associated with mother's parity. Mean haemoglobin level were significantly lower in mothers with a parity of more than one born children compared with mothers with first-born child.

The analyses demonstrated that there are no enough data to find any association between singleton/multiple pregnancy and haemoglobin level in-group 3 of the study.

3.5. Final Model:

Initially, a multivariate model was obtained by including the main effects of all the significant covariates. Then if a covariate in model was not significant, it was removed. Non-significant covariates were removed one at a time, starting with the highest P-values.

The interaction between each of the significant covariates and time was then evaluated.

One of the significant covariates was included in an interaction term with time in each fitted model to study whether covariates modify the effect of time. The effects of other covariates were examined with a similar approach.

We fitted the combination of models, if models were with more than one significant interaction.

Hemoglobin levels were significantly higher in mothers who were not vegetarian compared with mothers who were vegetarian in the past as results are shown in Table 3.18.

In the study, the relationship between haemoglobin levels and parity is also depicted for group one in Table 3.18 and Figure 3.8. A statistically significant difference in mean haemoglobin was not found due to mother's parity alone, but the interactions between parity and time were significant for both knots at (24,32) and (36,). As you can see in Figure 3.8 in knots (24,32), mother parity was negatively correlated with time, suggesting that the variation in time of pregnancy may be effective in lowering haemoglobin levels in mother with parity more than one.

Whereas, there was positive significant interaction between mother parity and time at knot (36,). It means that in this period of time the haemoglobin levels increase in mothers with a parity of more than one.

In group 2, number of cigarettes smoke per day in first 3 months, maternal ethnicity and singleton or twin of children were associated with haemoglobin

levels in the main effect model, but in final model except of singleton or twin of children other covariates were not statistically significant.

There was a significant negative association between haemoglobin levels and twin pregnancy in final model in-group 2; the haemoglobin levels were significantly higher in mothers with a singleton pregnancy.

However, as Table 3.19 and Figure 3.9 show there are significantly positive interaction between singleton or multiple pregnancy and knots (,12), (24,32), (32,36) ,suggesting that mothers with multiple pregnancy relatively high haemoglobin levels gain then the singleton pregnancy whereas there are significantly negative interaction between this covariate and knots (12,18) and (18,24).

As you can see in Table 3.20, in Group 3, haemoglobin levels in mother with parity more than one is significantly lower than mother with parity one.

No significant haemoglobin means differences were observed for education levels of mothers during pregnancy. However the interactions between maternal education in degree levels in knots (,12),(18,24),(36,) and Voc levels in knot (36,) just reach significance and are positive at (,12) and (36,) in degree levels and negative at knots (18,24), (36 ,) in degree and in Voc levels respectively. This implies that education level in high level (Degree level) in first and last knots with increase of time behavior interacts to produce higher haemoglobin levels in mothers. The negative interaction at knots (18,24) and (36,) suggests that with increase of time behavior interacts to produce lower haemoglobin levels in mothers in these education levels.

Hemoglobin levels and vegetarianism of mothers was not significantly correlated in Groups 3, although a significant interaction between vegetarian and time were demonstrated in Table 3.20 and Figure 3.10. Significant negative interactions were found between vegetarianism of mother (in the past) and time at knots (18,24) and (36,) and at the present at knot (18,24).

Whereby there was a decrease in haemoglobin levels when mothers were vegetarianism in the past at knots (18,24) and (36,) and were vegetarianism at the present at knot (18,24).

Table 3.20 and Figure 3.10 also indicated that there was a significant positive interaction between vegetarianism of mother (in the past and at the present) and time at knots (32,36) and (36,) respectively .

It means that haemoglobin level increase with time at knots (32,36), and (36,) when mothers were vegetarian in the past and at the present respectively.

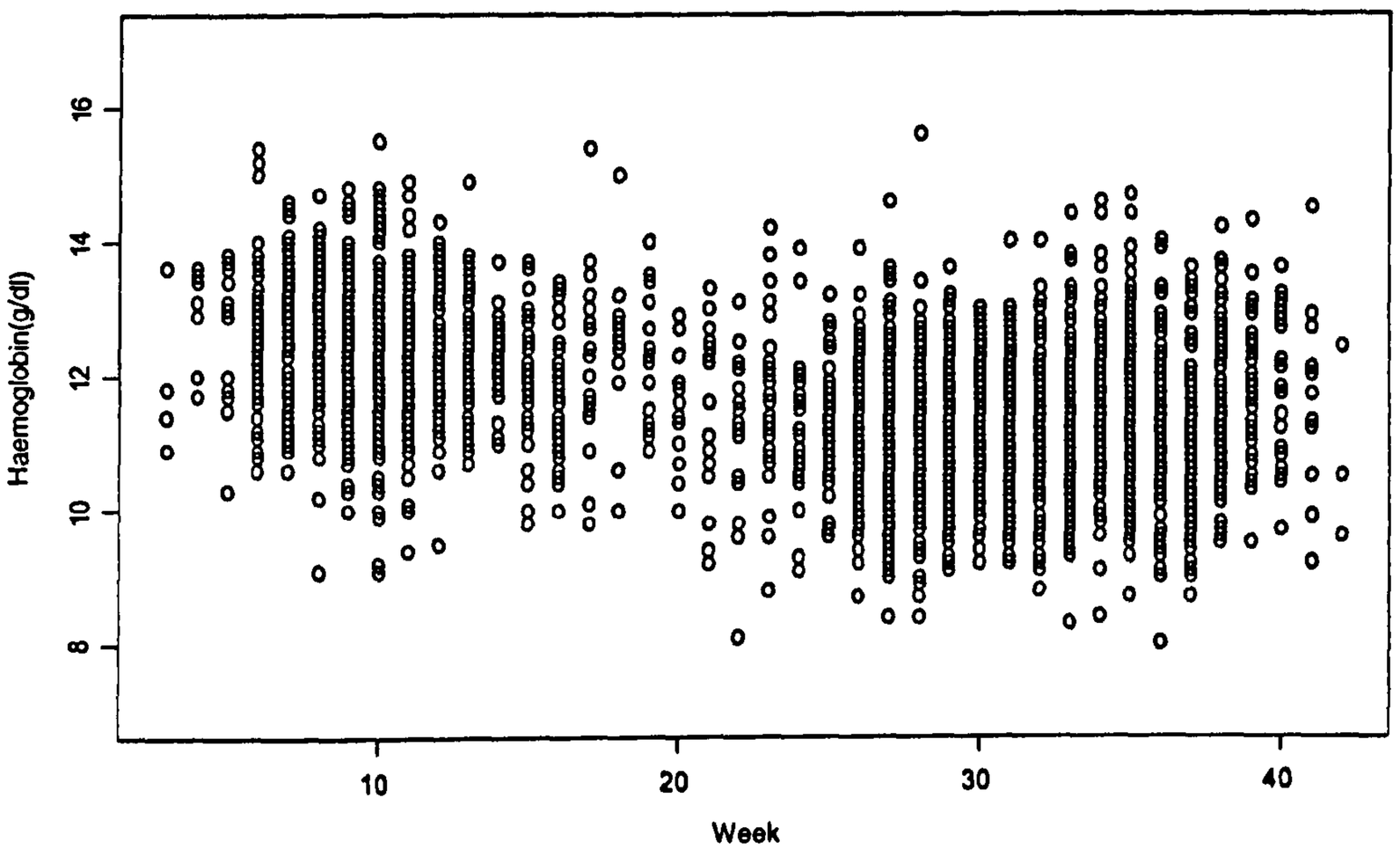


Figure.3.1: Scatter plot of mothers' haemoglobin concentration during pregnancy vs. time.

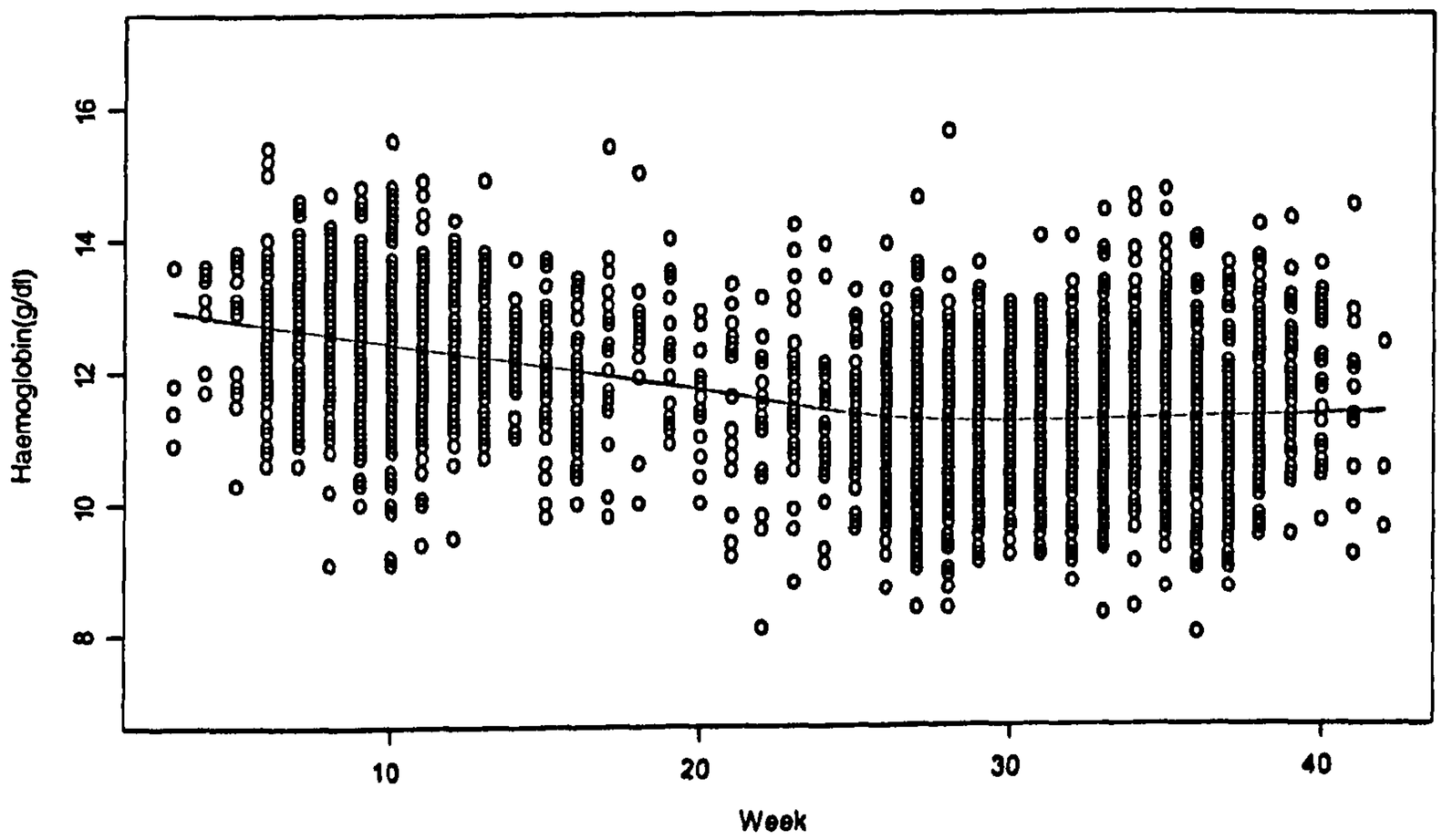


Figure.3.2: Variation in means haemoglobin concentration in mother during pregnancy.

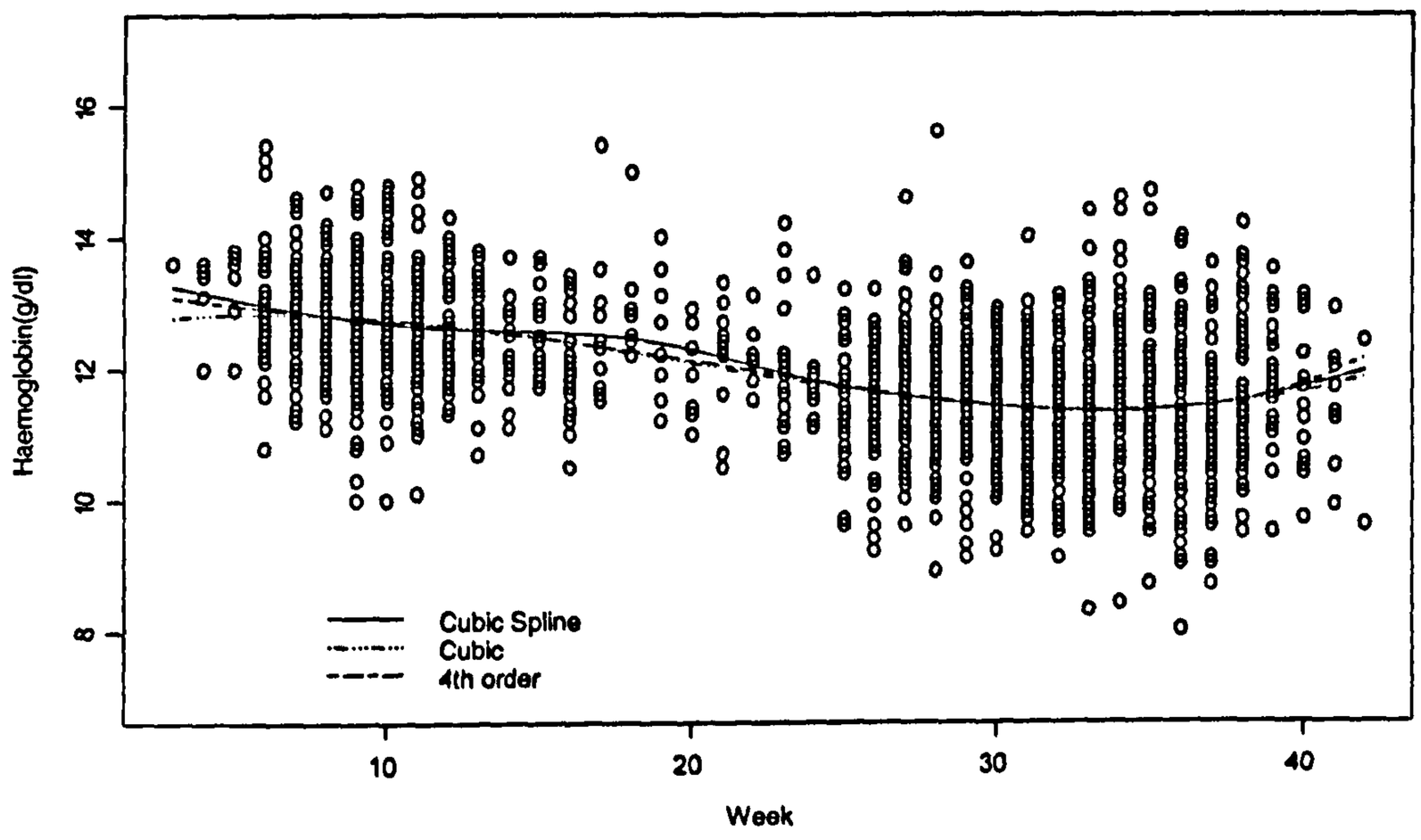


Figure.3.3: Comparison between Cubic, 4th order polynomial and Cubic Spline curves were fitted to haemoglobin concentration during pregnancy in Group 1.

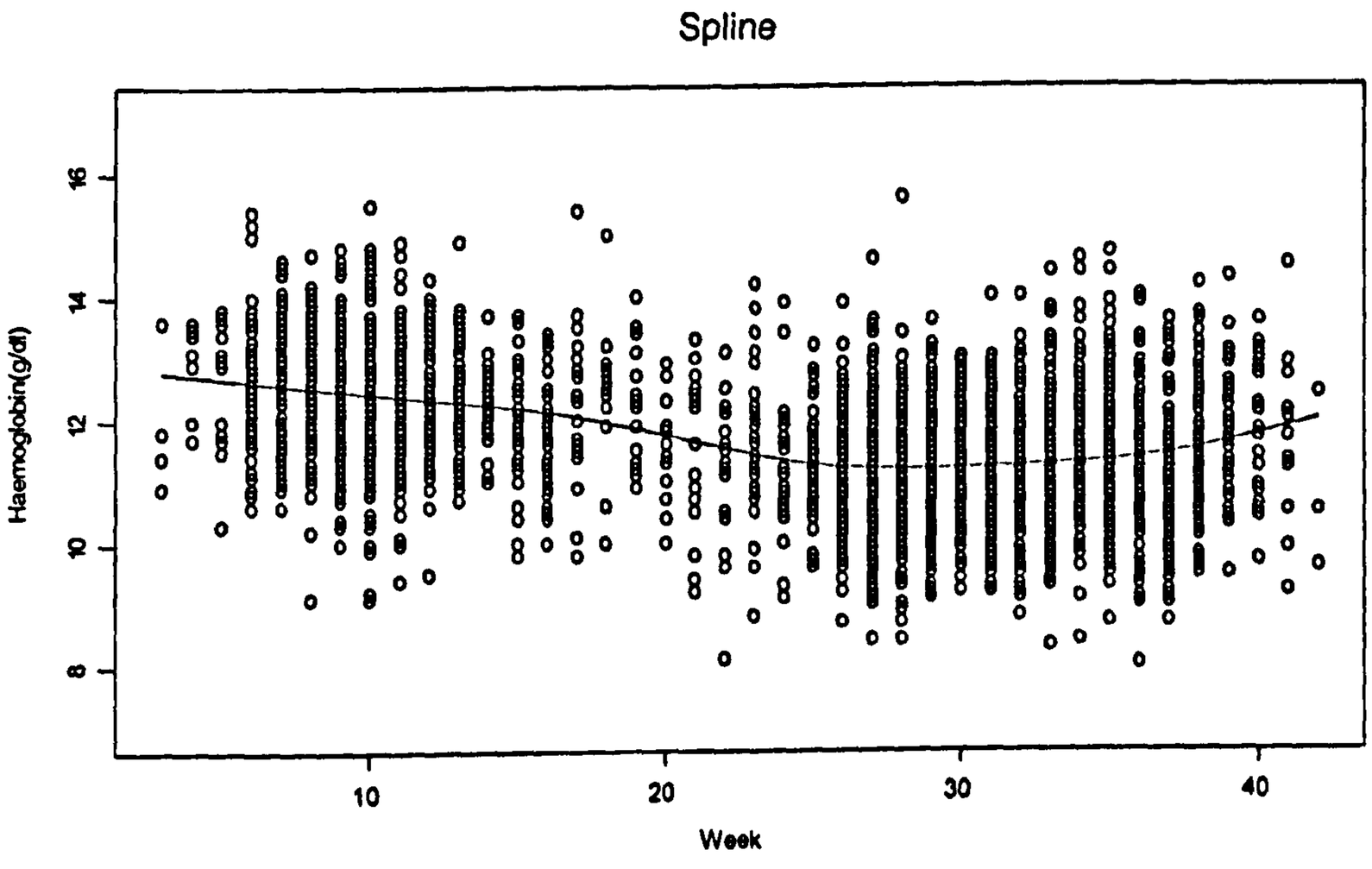


Figure.3.4: Cubic spline curve to haemoglobin concentration during pregnancy.

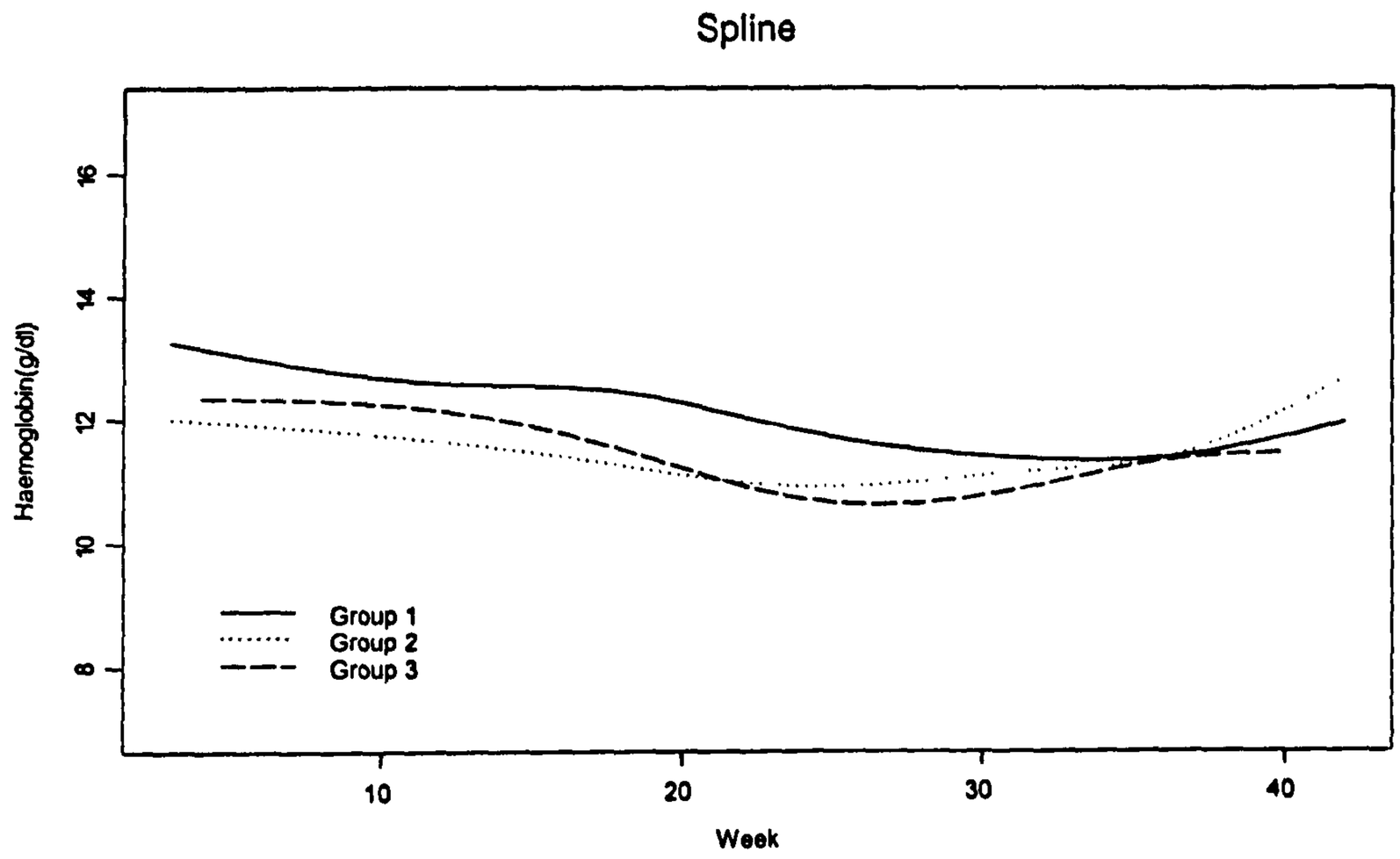


Figure.3.5: Cubic spline curves were fitted to haemoglobin concentration in mothers with and without iron supplementation.

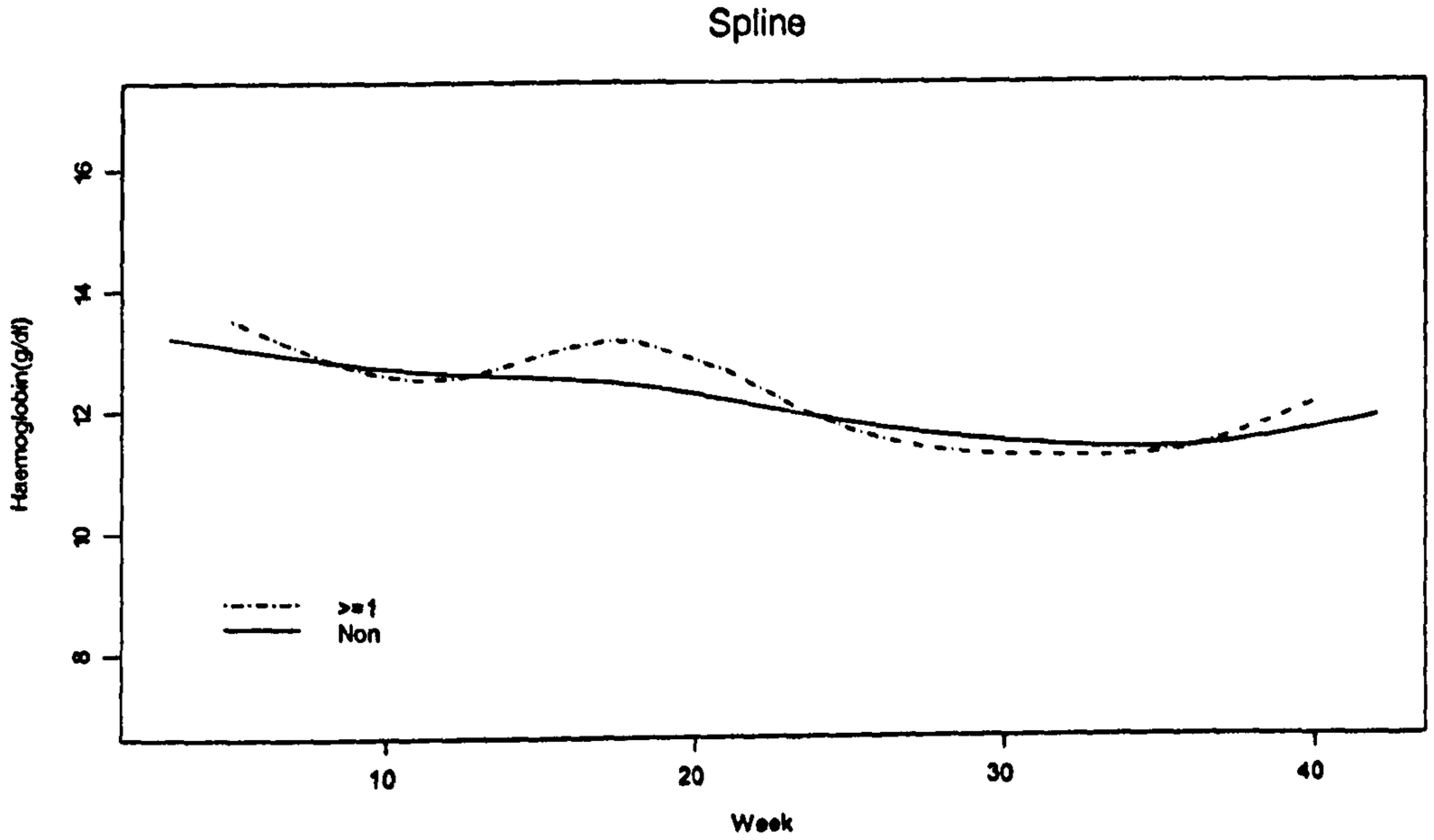


Figure.3.6: Cubic spine curves were fitted to mothers' haemoglobin concentration during pregnancy depending to the number of cigarettes smoked per day in last 2 weeks in Group 1.

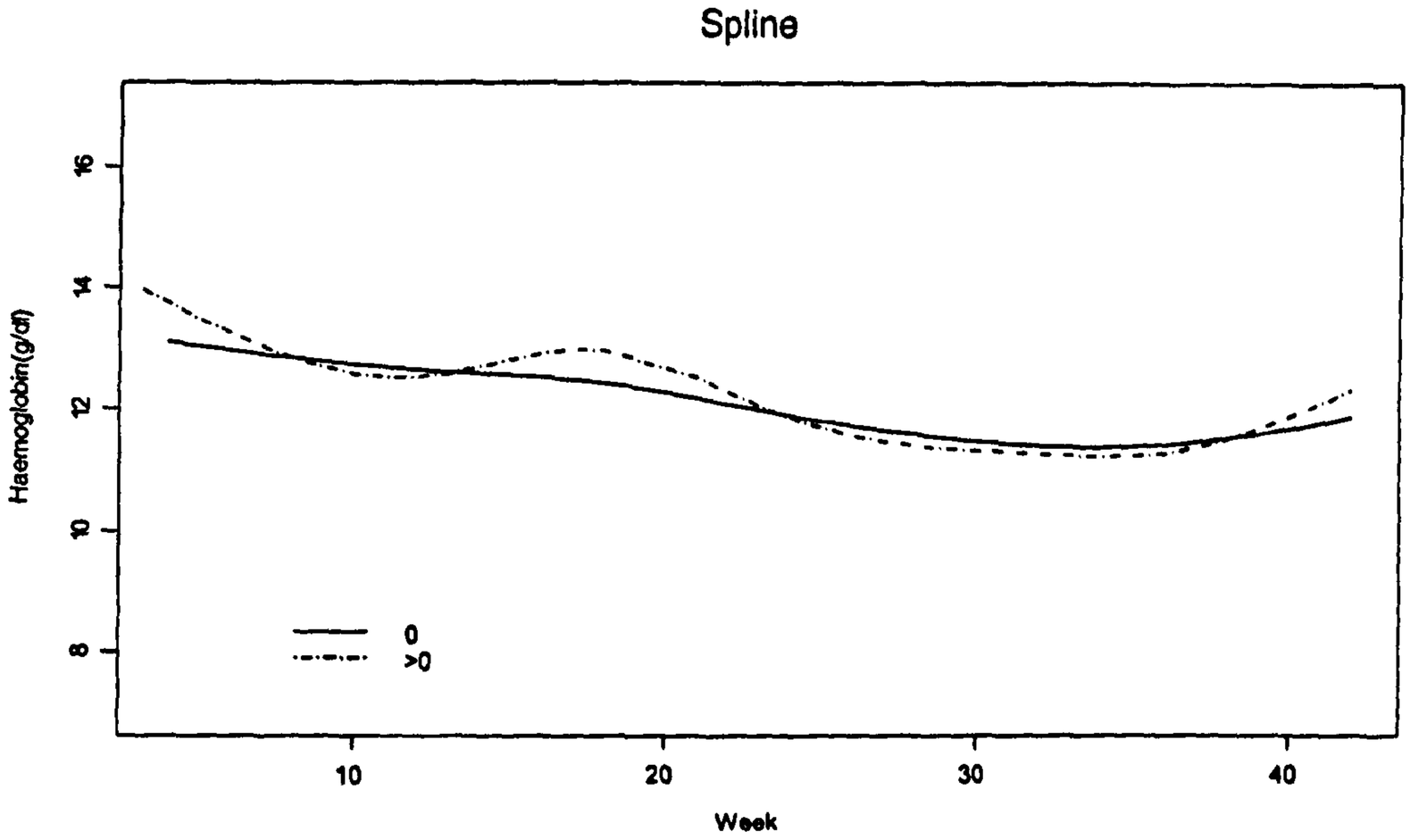


Figure.3.7: Cubic spine curves were fitted to mothers' haemoglobin concentration during pregnancy depending to the number of cigarettes smoked per day in first 3 months of pregnancy in Group 1.

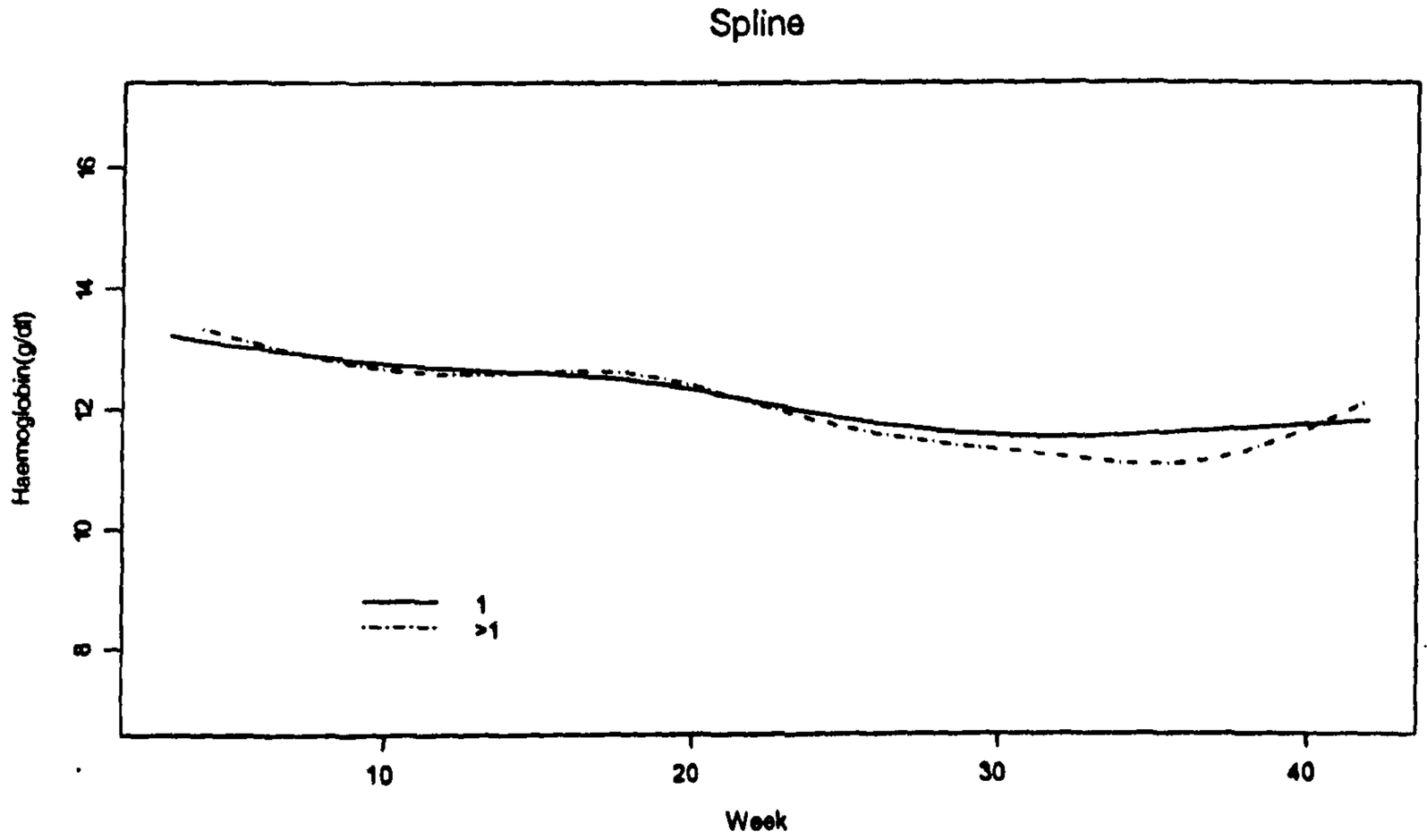


Figure.3.8: Cubic spine curves were fitted to mothers' haemoglobin concentration during pregnancy depending to the mother's parity in Group 1.

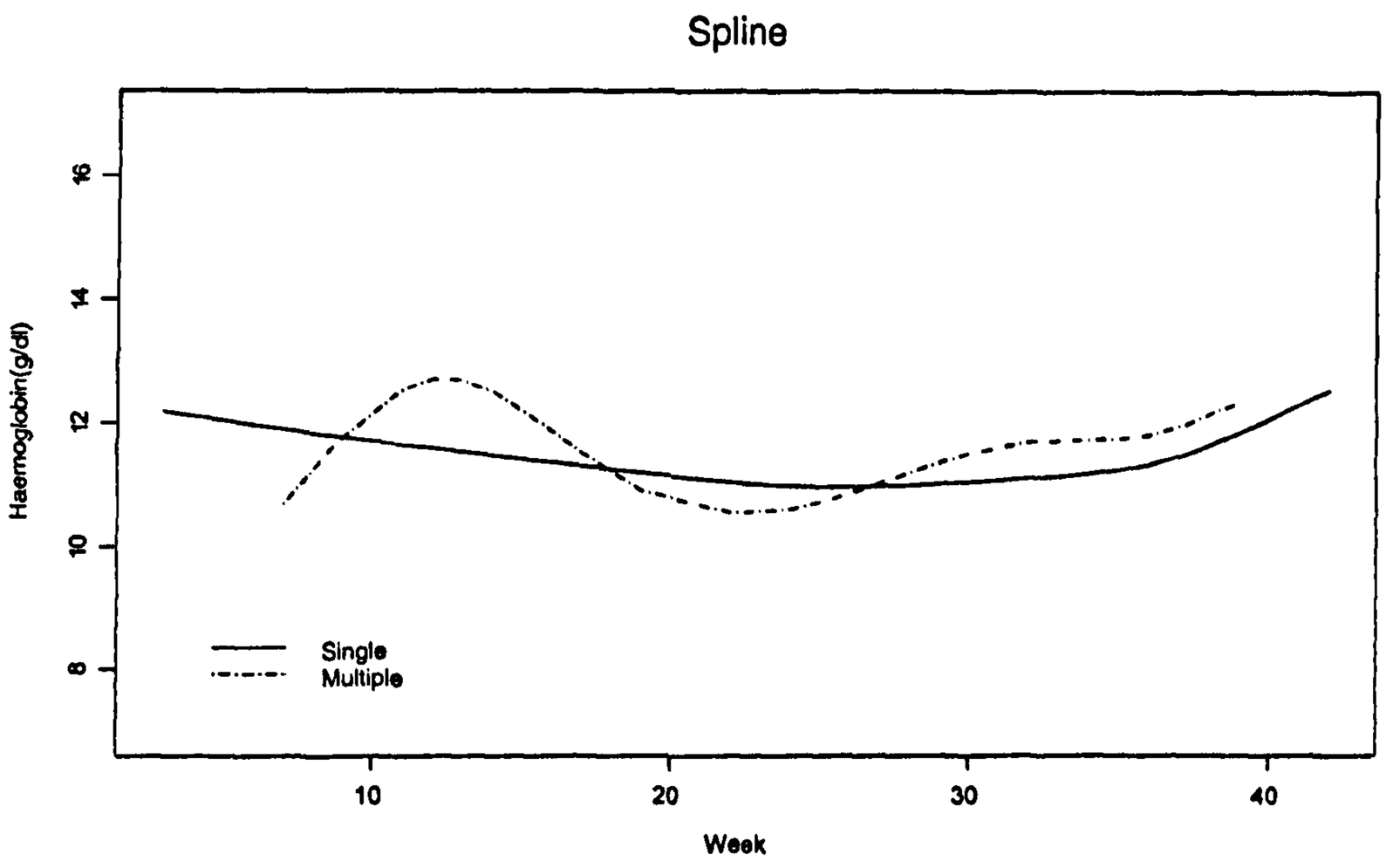


Figure.3.9: Cubic spine curves were fitted to mothers' haemoglobin concentration during pregnancy depending to the singleton/multiple pregnancy in Group 2.

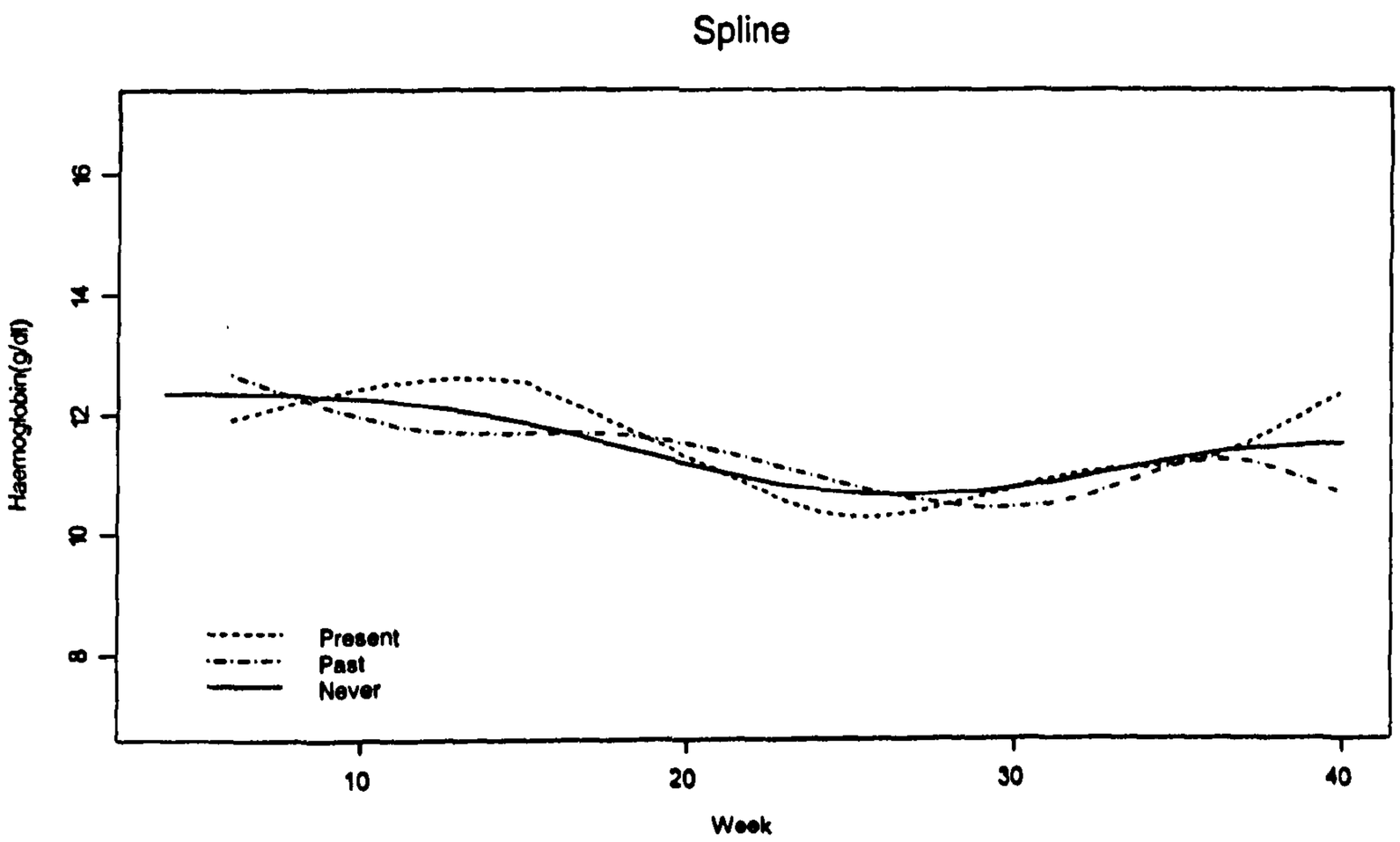


Figure.3.10: Cubic spine curves were fitted to mothers' haemoglobin concentration during pregnancy depending to the vegetarian statues of mother in Group 3.

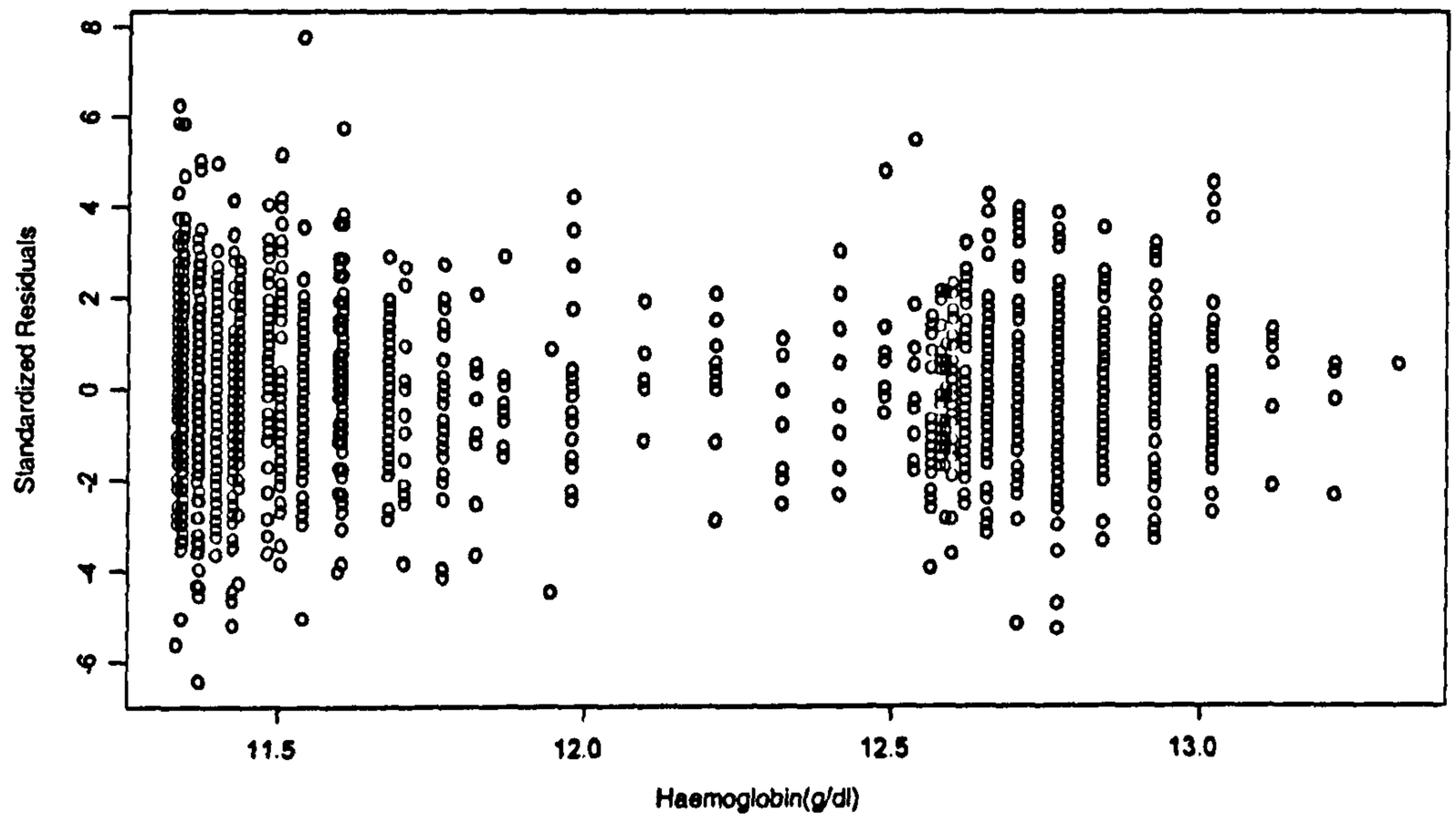


Figure.3.11: Plot of the Standardized Residuals against the fitted value of Haemoglobin level in Group 1.

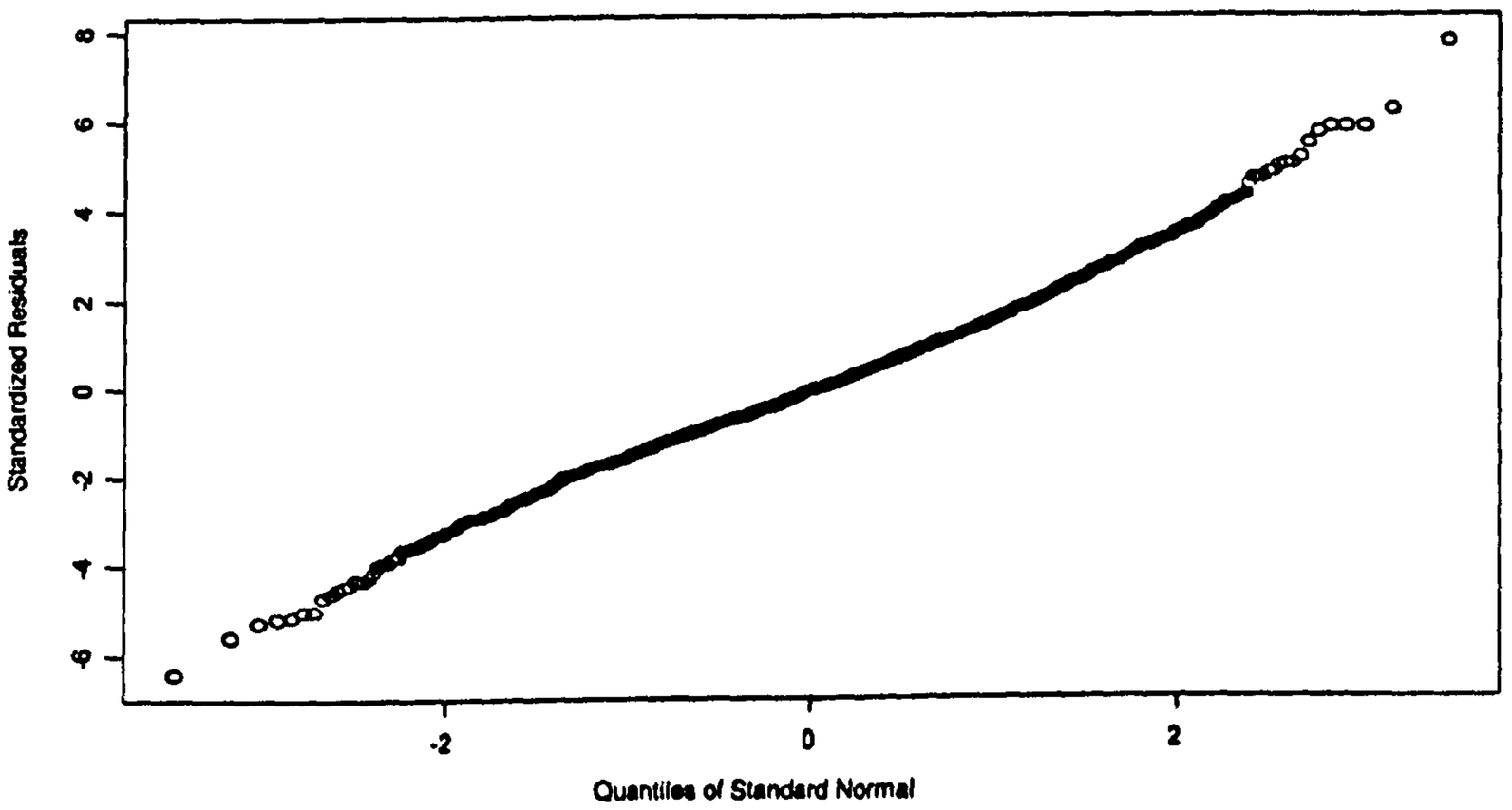


Figure 3.12: Normal Q-Q Plot of Residuals, Model for Group 1.

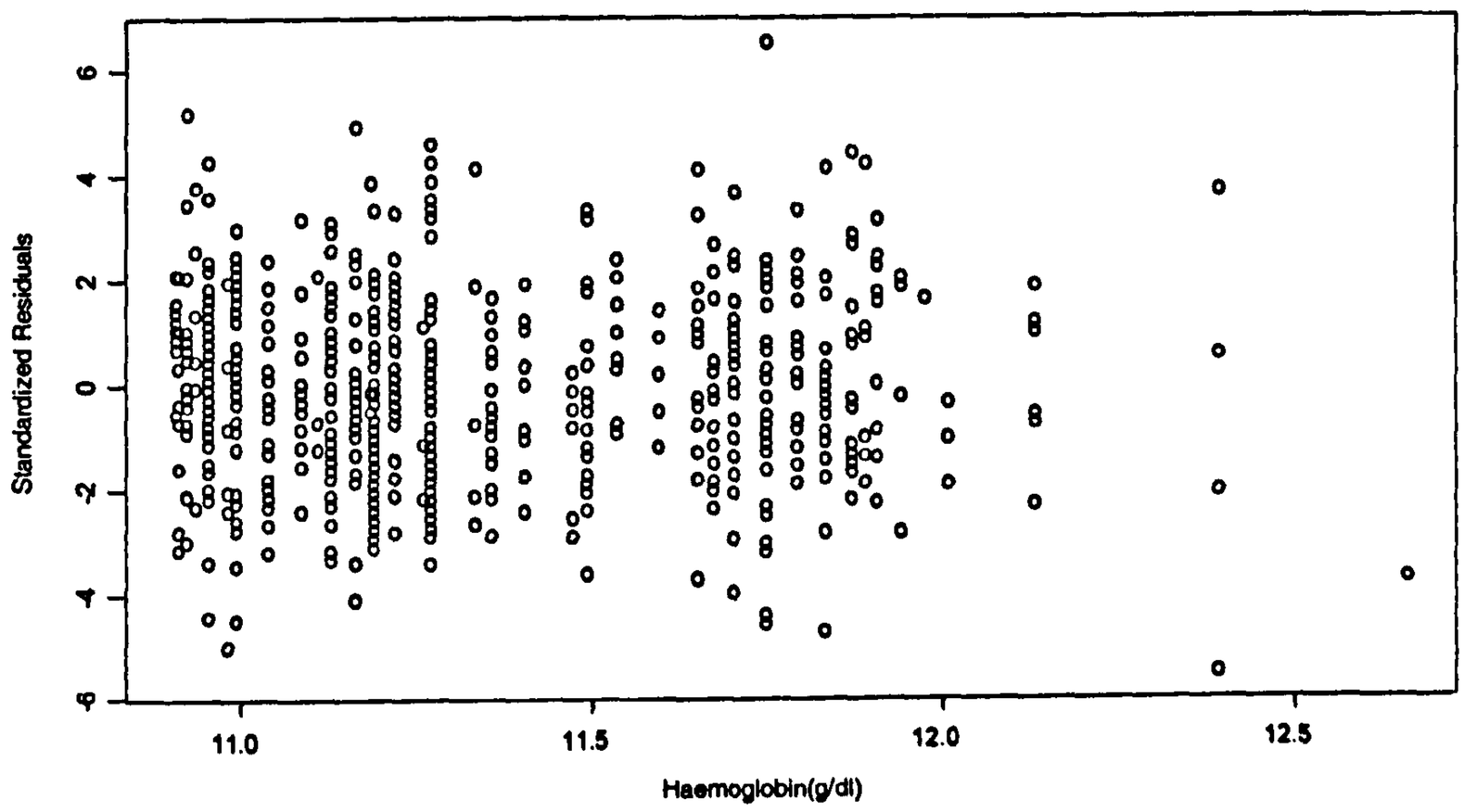


Figure.3.13: Plot of the Standardized Residuals against the fitted value of Haemoglobin level in Group 2.

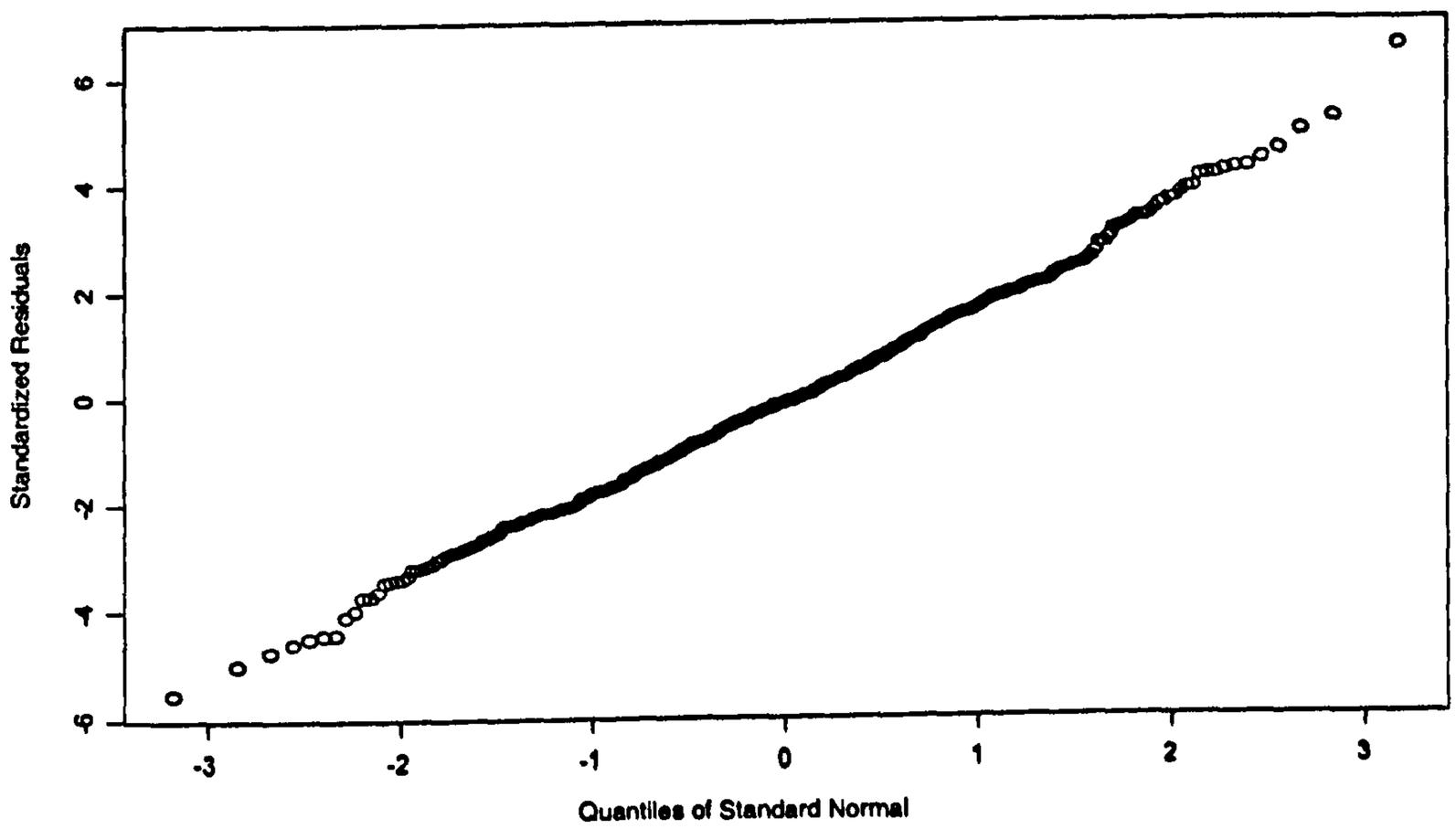


Figure 3.14: Normal Q-Q Plot of Residuals, Model for Group 2.

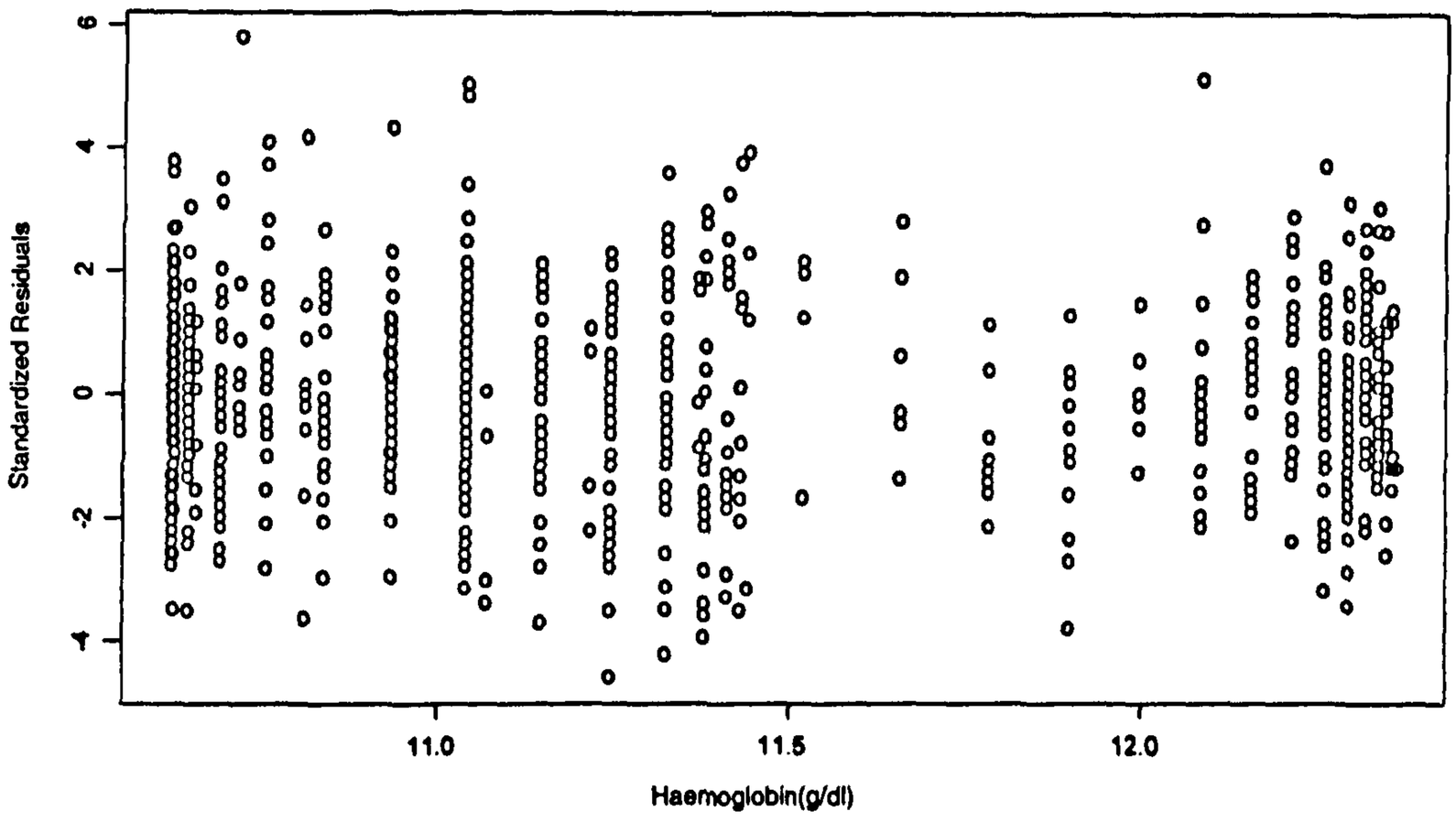


Figure.3.15: Plot of the Standardized Residuals against the fitted value of Haemoglobin level in Group 3.

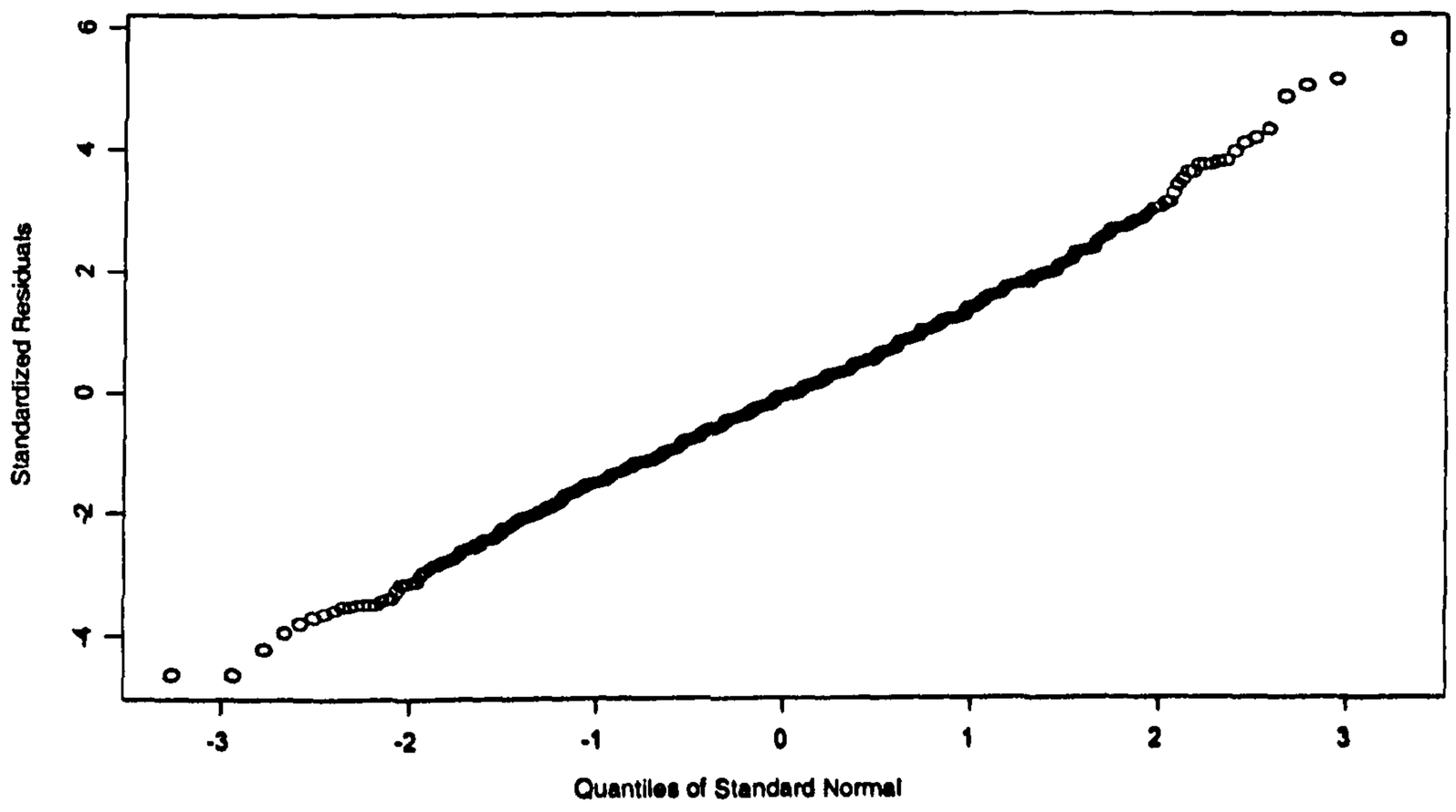


Figure 3.16: Normal Q-Q Plot of Residuals, Model for Group 3.

Knots	*11*	*7*	*5*
Loglik	-2430.497	-2434.488	-2434.685
AIC	4894.994	4894.977	4891.370
P-value		0.0922	0.8216

Table.3.1: Log likelihood and AIC values for different number of knots of the baseline.

Linear mixed-effects model fit by REML
Data: ni

AIC	BIC	logLik
4523.391	4562.481	-2254.696

Random effects:
Formula: ~ time | sub
Structure: General positive-definite

	Std Dev	Corr
(Intercept)	0.67783920	(Inter
time	0.02531379	-0.588
Residual	0.54197123	

Fixed effects: HAEM ~ time + (time^2)

	Value	Std.Error	DF	t-value	p-value
(Intercept)	13.76847	0.07726957	1237	178.1875	<.0001
time	-0.11763	0.00826173	1237	-14.2381	<.0001
I(time^2)	0.00140	0.00018905	1237	7.4292	<.0001

Linear mixed-effects model fit by REML
Data: ni

AIC	BIC	logLik
4507.245	4551.915	-2245.622

Random effects:
Formula: ~ time | sub
Structure: General positive-definite

	Std Dev	Corr
(Intercept)	0.69401169	(Inter
time	0.02583653	-0.594
Residual	0.52786401	

Fixed effects: HAEM ~ time + (time^2) + (time^3)

	Value	Std.Error	DF	t-value	p-value
(Intercept)	12.65061	0.1956244	1236	64.66786	<.0001
time	0.08062	0.0328972	1236	2.45072	0.014
I(time^2)	-0.00807	0.0015323	1236	-5.26607	<.0001
I(time^3)	0.00013	0.0000216	1236	6.23632	<.0001

Linear mixed-effects model fit by REML
Data: ni

AIC	BIC	logLik
4524.678	4574.927	-2253.339

Random effects:
Formula: ~ time | sub
Structure: General positive-definite

	Std Dev	Corr
(Intercept)	0.68708393	(Inter
time	0.02585501	-0.588
Residual	0.52628565	

Fixed effects: HAEM ~ time + (time^2) + (time^3) + (time^4)

	Value	Std.Error	DF	t-value	p-value
(Intercept)	13.52844	0.3518430	1235	38.45020	<.0001
time	-0.13694	0.0796964	1235	-1.71830	0.086
I(time^2)	0.00958	0.0060898	1235	1.57286	0.116
I(time^3)	-0.00043	0.0001907	1235	-2.26699	0.024
I(time^4)	0.00001	0.0000021	1235	2.99288	0.003

Linear mixed-effects model fit by REML
Data: ni

AIC	BIC	logLik
4549.454	4605.282	-2264.727

Random effects:
Formula: ~ time | sub
Structure: General positive-definite

	Std Dev	Corr
(Intercept)	0.6943596	(Inter
time	0.0260986	-0.596
Residual	0.5236155	

Fixed effects: HAEM ~ time + (time^2) + (time^3) + (time^4) + (time^5)

	Value	Std.Error	DF	t-value	p-value
Intercept	15.04292	0.7136062	1234	21.08014	<.0001
time	-0.62927	0.2168475	1234	-2.90192	0.0038
I(time^2)	0.06586	0.0238343	1234	2.76340	0.0058
I(time^3)	-0.00329	0.0011841	1234	-2.77666	0.0056
I(time^4)	0.00007	0.0000271	1234	2.66759	0.0077
I(time^5)	0.00000	0.0000002	1234	-2.44396	0.0147

Table.3.2: Results of fitting Mixed Effects Models with different order to mother's haemoglobin concentration during pregnancy in Group 1.

	Group 1 No	Group 1 V-per*	Group 2 No	Group 2 V-per	Group 3 No	Group 3 V-per	Total	Total V-pre	χ^2 Test P-Value
Education:									
'CSE'	105	15%	29	12%	43	14%	177	14%	
'VOC'	73	10%	26	11%	27	9%	126	10%	
O level	246	34%	75	32%	133	43%	454	36%	
A level	181	25%	75	32%	70	23%	326	26%	
Degree	118	16%	33	13%	34	11%	185	14%	
Missing	8		10		1		19		16.49
Total	731		248		308		1287		0.036
Ethnicity of mother									
White	712	98%	229	96%	301	96%	1242	96%	
Non-white	13	2%	9	3%	7	4%	29	4%	
Missing	6		10				16		3.178
Total	731		248		308	100	1287		0.204
Vegetarian:									
Current	30	4%	12	5%	31	10%	73	6%	
In past	53	7%	26	11%	34	11%	113	9%	
Never	634	89%	195	84%	237	79%	1066	85%	
Missing	14		15		5		34		21.21
Total	731		248	100	308		1287		0.0003
Age:									
Mean (SD)	29.02 (4.52)		29.0 (4.8)		29.2 (4.5)		28.9 (4.68)		
Total	731		248		308		1287		
No of cig smoked in 1st 3 months of preg.									
None	594	82%	189	76%	252	82%	1035	81%	
1+	131	18%	59	24%	56	18%	246	9%	
Missing	6						6		4.17
Total	731		248		308		1287		0.1242
No cig smoked in last 2- week of preg									
None	628	87%	200	81%	264	86%	1092	85%	
1+	98	13%	48	19%	44	14%	190	15%	
Missing	5						5		5.11
Total	731		248		308		1287		0.078
No cig smoked in week 32									
None	605	86%	177	82%	268	87%	1050	85%	
1+	102	14%	38	18%	40	13%	180	15%	
Missing	24		33				57		2.28
Total	731		248		308		1287		0.319
Sing/twin									
1	730	99%	235	94%	305	98%	1270	99%	
2	1	1%	13	6%	3	2%	17	1%	
Missing									37.4
Total	731		248		308		1287		0.000
Parity									
1	349	49%	108	44%	126	41%	583	46%	
≥2	367	51%	137	56%	181	59%	685	54%	
Missing	15		3		11		29		5.6
Total	731		248		308		1287		0.062

Table.3.3: Number and percentage of subjects in each group by covariates.

* The percentages are based only on the cases who actually answered the question.

Covariate	Group 1 P-Value	Group 2 P-Value	Group 3 P-Value
Education:			
'CSE'	0.36	0.45	0.65
'VOC'	0.71	0.56	0.08
O level	0.13	0.18	0.046
A level	0.83	0.62	0.96
Likelihood ratio test	0.54	0.45	0.044
Ethnic background:			
Non-white	0.2	0.04	0.93
Age:	0.67	0.33	0.43
Vegetarian:			
In Past	0.017	0.32	0.14
At Present	0.98	0.55	0.87
In past*ns(time, knots = c(18,24))3	0.67	0.12	0.038
At Present*ns(time, knots = c(18,24))3	0.09	0.19	0.03
Likelihood ratio test	0.042	0.57	0.044
Singleton /Multiple:	0.29	0.01	
SM*ns(time, knots = c(,12))1	0.58	0.01	
SM*ns(time, knots = c(12,18))2	0.66	0.037	
SM*ns(time, knots = c(18,24))3	0.45	0.002	
SM*ns(time, knots = c(24,32))4	0.35	0.033	
SM*ns(time, knots = c(32,36))5	0.48	0.003	
SM*ns(time, knots = c(36,))6	0.18	0.27	
Likelihood ratio test		0.003	
No smoked /day in 1st 3 mths of prog:			
≥1	0.02	0.041	0.30
N3M*ns(time, knots = c(12,18))2	0.01		
N3M*ns(time, knots = c(18,24))3	0.019		
N3M*ns(time, knots = c(32,36))5	0.023		
Likelihood ratio test	0.026		
No smoked /day in last 2 wks of prog:			
≥1	0.021	0.07	0.78
N2W*ns(time, knots = c(12, 18)2	0.013	0.83	0.59
N2W*ns(time, knots = c(18, 24)3	0.01	0.98	0.2
Likelihood ratio test	0.001	0.07	0.78
No smoked /day at 32 weeks of prog: ≥1	0.62	0.15	0.52
Parity:			
≥2	0.31	0.87	0.017
Par*ns(time, knots = c(24, 32)4	0.0002	0.46	0.35
Likelihood ratio test	<.0001	0.20	0.017

Table.3.4: Coefficient values for covariates in Linear Mixed Models.

```
Linear mixed-effects model fit by REML
 Data: mum
       AIC      BIC    logLik
  4465.522 4526.926 -2221.761

Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 0.6959215 (Inter
       time 0.0262359 -0.599
   Residual 0.5230956

Fixed effects: HAEM ~ ns(time, knots = c(12, 18, 24, 32, 36))
                                                Value Std.Error   DF   t-value  p-value
                                    (Intercept) 13.32120 0.1427025 1233  93.34950   <.0001
ns(time, knots = c(12, 18, 24, 32, 36))1 -0.66739 0.1461893 1233  -4.56527   <.0001
ns(time, knots = c(12, 18, 24, 32, 36))2 -1.62688 0.1847999 1233  -8.80350   <.0001
ns(time, knots = c(12, 18, 24, 32, 36))3 -1.95536 0.1563000 1233 -12.51029   <.0001
ns(time, knots = c(12, 18, 24, 32, 36))4 -1.76258 0.1152704 1233 -15.29084   <.0001
ns(time, knots = c(12, 18, 24, 32, 36))5 -2.33555 0.3496822 1233  -6.67907   <.0001
ns(time, knots = c(12, 18, 24, 32, 36))6 -0.93095 0.1779243 1233  -5.23230   <.0001
```

Table.3.5: REML estimates for the model fitted to haemoglobin concentration during pregnancy in Group 1.

```
Linear mixed-effects model fit by REML
 Data: mum18
       AIC      BIC   logLik
  1743.084 1792.515 -860.542

Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 1.10591582 (Inter
       time 0.03853903 -0.767
   Residual 0.57432760

Fixed effects: HAEM ~ ns(time, knots = c(12, 18, 24, 32, 36))

                                                Value  Std.Error  DF   t-value  p-value
                                    (Intercept) 12.00677 0.3002784 414  39.98545   <.0001
ns(time, knots = c(12, 18, 24, 32, 36))1 -0.72967 0.3048507 414  -2.39352   0.0171
ns(time, knots = c(12, 18, 24, 32, 36))2 -1.29327 0.3668720 414  -3.52511   0.0005
ns(time, knots = c(12, 18, 24, 32, 36))3 -0.81834 0.3194060 414  -2.56207   0.0108
ns(time, knots = c(12, 18, 24, 32, 36))4 -0.70399 0.2288808 414  -3.07577   0.0022
ns(time, knots = c(12, 18, 24, 32, 36))5 -0.11416 0.7181552 414  -0.15897   0.8738
ns(time, knots = c(12, 18, 24, 32, 36))6  0.79470 0.3326700 414   2.38887   0.0173
```

Table.3.6: REML estimates for the model fitted to haemoglobin concentration during pregnancy in Group 2.

```
Linear mixed-effects model fit by REML
 Data: mum32
       AIC     BIC   logLik
  2080.879 2133.73 -1029.44

Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 0.70119289 (Inter
       time 0.02570154 -0.648
   Residual 0.55069795

Fixed effects: HAEM ~ ns(time, knots = c(12, 18, 24, 32, 36))
                                                Value Std.Error  DF   t-value  p-value
                                    (Intercept) 12.35924 0.1909684 595  64.71880    <0001
ns(time, knots = c(12, 18, 24, 32, 36))1 -0.77207 0.2067783 595 -3.73380  0.0002
ns(time, knots = c(12, 18, 24, 32, 36))2 -1.96414 0.2529932 595 -7.76361  <.0001
ns(time, knots = c(12, 18, 24, 32, 36))3 -1.62359 0.2165084 595 -7.49897  <.0001
ns(time, knots = c(12, 18, 24, 32, 36))4 -0.94588 0.1597919 595 -5.91944  <.0001
ns(time, knots = c(12, 18, 24, 32, 36))5 -0.97104 0.4698162 595 -2.06685  0.0392
ns(time, knots = c(12, 18, 24, 32, 36))6 -0.89471 0.2034032 595 -4.39870  <.0001
```

Table.3.7: REML estimates for the model fitted to haemoglobin concentration during pregnancy in Group 3.

```
Linear mixed-effects model fit by REML

       AIC      BIC   logLik
  4388.519 4460.834 -2181.26

Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 0.69931018 (Inter
       time 0.02657072 -0.611
   Residual 0.52194146

Fixed effects: HAEM ~ Vg + ns(time, knots = c(12, 18, 24, 32, 36))
                                                Value Std.Error   DF   t-value p-value
                                    (Intercept) 13.25945 0.1524068 1211  87.00037  <.0001
                                       In past -0.11623 0.0487660  714  -2.38337  0.0174
                                    At present -0.00129 0.0440799  714  -0.02929  0.9766
ns(time, knots = c(12, 18, 24, 32, 36))1 -0.63185 0.1481690 1211  -4.26438  <.0001
ns(time, knots = c(12, 18, 24, 32, 36))2 -1.69441 0.1864420 1211  -9.08816  <.0001
ns(time, knots = c(12, 18, 24, 32, 36))3 -1.97072 0.1578168 1211 -12.48741  <.0001
ns(time, knots = c(12, 18, 24, 32, 36))4 -1.79322 0.1163469 1211 -15.41272  <.0001
ns(time, knots = c(12, 18, 24, 32, 36))5 -2.42966 0.3530025 1211  -6.88284  <.0001
ns(time, knots = c(12, 18, 24, 32, 36))6 -0.92607 0.1782382 1211  -5.19570  <.0001
```

Table.3.8: REML estimates for the model fitted to haemoglobin concentration during pregnancy stratified by vegetarianism of mothers in Group 1.

```
Linear mixed-effects model fit by REML

       AIC      BIC    logLik
  4424.373 4524.594 -2194.186
Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 0.68676725 (Inter
       time 0.02593954 -0.591
   Residual 0.52107389
Fixed effects: HAEM ~ N2W* ns(time, knots = c(12, 18, 24, 32, 36))
                                              Value Std.Error   DF  t-value   p-value
                                  (Intercept) 13.68214 0.1980438 1211  69.08642   <.0001
                                         >=1  0.45966 0.1980438  724   2.32100   0.0206
    ns(time, knots = c(12,18,24,32,36))1 -0.51392 0.2021564 1211  -2.54219   0.0111
    ns(time, knots = c(12,18,24,32,36))2 -2.15821 0.2670330 1211  -8.08219   <.0001
    ns(time, knots = c(12,18,24,32,36))3 -2.39222 0.2175481 1211 -10.99630   <.0001
    ns(time, knots = c(12,18,24,32,36))4 -1.94271 0.1813487 1211 -10.71259   <.0001
    ns(time, knots = c(12,18,24,32,36))5 -3.09165 0.5036590 1211  -6.13838   <.0001
    ns(time, knots = c(12,18,24,32,36))6 -0.59706 0.3225539 1211  -1.85105   0.0644
N2W*ns(time, knots = c(12,18,24,32,36))1  0.15239 0.2021564 1211   0.75382   0.4511
N2W*ns(time, knots = c(12,18,24,32,36))2 -0.66150 0.2670330 1211  -2.47721   0.0134
N2W*ns(time, knots = c(12,18,24,32,36))3 -0.56466 0.2175481 1211  -2.59556   0.0096
N2W*ns(time, knots = c(12,18,24,32,36))4 -0.22594 0.1813487 1211  -1.24587   0.2131
N2W*ns(time, knots = c(12,18,24,32,36))5 -0.94907 0.5036590 1211  -1.88435   0.0598
N2W*ns(time, knots = c(12,18,24,32,36))6  0.39128 0.3225539 1211   1.21308   0.2253
```

Table.3.9: REML estimates for the model fitted to haemoglobin concentration during pregnancy stratified by number of cigarettes smoked per day in last 2 weeks of pregnancy in Group 1.

```
 Linear mixed-effects model fit by REML
       AIC      BIC    logLik
  4432.018 4532.231 -2198.009
Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 0.68621452  (Inter
       time 0.02581754  -0.589
   Residual 0.52324284
Fixed effects: HAEM ~ N3M* ns(time, knots = c(12,18,24,32,36))
                                              Value Std.Error   DF  t-value p-value
                                  (Intercept) 13.58981 0.1740743 1211  78.06905 <.0001
                                         >=1   0.40704 0.1740743  723   2.33831 0.0196
    ns(time, knots = c(12,18,24,32,36))1 -0.60798 0.1785353 1211  -3.40537 0.0007
    ns(time, knots = c(12,18,24,32,36))2 -2.01139 0.2272888 1211  -8.84948 <.0001
    ns(time, knots = c(12,18,24,32,36))3 -2.26235 0.1909988 1211 -11.84485 <.0001
    ns(time, knots = c(12,18,24,32,36))4 -1.93690 0.1444880 1211 -13.40525 <.0001
    ns(time, knots = c(12,18,24,32,36))5 -3.01317 0.4312545 1211  -6.98699 <.0001
    ns(time, knots = c(12,18,24,32,36))6 -0.82509 0.2215716 1211  -3.72380 0.0002
N3M*ns(time, knots = c(12,18,24,32,36))1  0.04254 0.1785353 1211   0.23830 0.8117
N3M*ns(time, knots = c(12,18,24,32,36))2 -0.58653 0.2272888 1211  -2.58053 0.0100
N3M*ns(time, knots = c(12,18,24,32,36))3 -0.44708 0.1909988 1211  -2.34073 0.0194
N3M*ns(time, knots = c(12,18,24,32,36))4 -0.27946 0.1444880 1211  -1.93414 0.0533
N3M*ns(time, knots = c(12,18,24,32,36))5 -0.98044 0.4312545 1211  -2.27345 0.0232
N3M*ns(time, knots = c(12,18,24,32,36))6  0.19187 0.2215716 1211   0.86595 0.3867
```

Table.3.10: REML estimates for the model fitted to haemoglobin concentration during pregnancy stratified by number of smoke per day in first 3 months in Group 1.

```
Linear mixed-effects model fit by REML

    AIC      BIC    logLik
4342.73 4442.679 -2153.365

Random effects:
Formula:  ~ time | sub
Structure: General positive-definite
                StdDev   Corr
(Intercept) 0.68241740 (Inter
       time 0.02472366 -0.588
   Residual 0.52595698
 Fixed effects: HAEM ~ Parity* ns(time, knots = c(12, 18, 24, 32, 36))
                                                 Value Std.Error   DF   t-value p-value
                                           (Intercept)  12.90697  0.448141 1192  28.80113  <.0001
                                                  >=2   0.29299  0.289187  714   1.01314  0.3113
    ns(time, knots = c(12,18,24,32,36))1  -0.63240  0.476303 1192  -1.32773  0.1845
    ns(time, knots = c(12,18,24,32,36))2  -0.83691  0.584417 1192  -1.43205  0.1524
    ns(time, knots = c(12,18,24,32,36))3  -1.31867  0.493789 1192  -2.67052  0.0077
    ns(time, knots = c(12,18,24,32,36))4  -0.50202  0.359118 1192  -1.39792  0.1624
    ns(time, knots = c(12,18,24,32,36))5  -1.08490  1.100717 1192  -0.98563  0.3245
    ns(time, knots = c(12,18,24,32,36))6  -1.45231  0.548638 1192  -2.64713  0.0082
Par*ns(time, knots = c(12,18,24,32,36))1  -0.00729  0.297303 1192  -0.02453  0.9804
Par*ns(time, knots = c(12,18,24,32,36))2  -0.55622  0.375035 1192  -1.48313  0.1383
Par*ns(time, knots = c(12,18,24,32,36))3  -0.43325  0.315906 1192  -1.37144  0.1705
Par*ns(time, knots = c(12,18,24,32,36))4  -0.86073  0.232205 1192  -3.70677  0.0002
Par*ns(time, knots = c(12,18,24,32,36))5  -0.89520  0.709266 1192  -1.26215  0.2071
Par*ns(time, knots = c(12,18,24,32,36))6   0.35144  0.361569 1192   0.97197  0.3313
```

Table.3.11: REML estimates for the model fitted to haemoglobin concentration during pregnancy stratified by mother's parity in Group 1.

```
Linear mixed-effects model fit by REML
 Data: mumba
       AIC      BIC    logLik
  1679.854 1733.373 -827.9269

Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 1.10117359 (Inter
       time 0.03789715 -0.765
   Residual 0.56971540

Fixed effects: HAEM ~ MumEthn + ns(time, knots = c(12, 18, 24, 32, 36))
                                                 Value Std.Error  DF   t-value  p-value
                                           Intercept)  12.64954 0.4186260 403  30.21680   <.0001
                                            Non-white  -0.57341 0.2782141 236  -2.06103   0.0404
ns(time, knots = c(12, 18, 24, 32, 36))1  -0.74524 0.3067974 403  -2.42908   0.0156
ns(time, knots = c(12, 18, 24, 32, 36))2  -1.34804 0.3672751 403  -3.67038   0.0003
ns(time, knots = c(12, 18, 24, 32, 36))3  -0.86999 0.3201270 403  -2.71764   0.0069
ns(time, knots = c(12, 18, 24, 32, 36))4  -0.69225 0.2301984 403  -3.00718   0.0028
ns(time, knots = c(12, 18, 24, 32, 36))5  -0.19744 0.7185399 403  -0.27477   0.7836
ns(time, knots = c(12, 18, 24, 32, 36))6   0.74798 0.3301518 403   2.26558   0.0240
```

Table.3.12: REML estimates for the model fitted to haemoglobin concentration during pregnancy stratified by ethnicity of mothers in Group 2.

```
Linear mixed-effects model fit by REML
 Data: mumb670
       AIC      BIC    logLik
  1744.689 1798.596 -860.3445

Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 1.08569709 (Inter
       time 0.03845807 -0.762
   Residual 0.57427949

Fixed effects: HAEM ~ N3M + ns(time, knots = c(12, 18, 24, 32, 36))
                                                  Value Std.Error  DF   t-value  p-value
                                        (Intercept)  12.05203 0.2982238 414  40.41271   <.0001
                                                >=1   0.12626 0.0615027 246   2.05298   0.0411
ns(time, knots = c(12, 18, 24, 32, 36))1  -0.75294 0.3026674 414  -2.48769   0.0133
ns(time, knots = c(12, 18, 24, 32, 36))2  -1.25449 0.3649461 414  -3.43745   0.0006
ns(time, knots = c(12, 18, 24, 32, 36))3  -0.80395 0.3171346 414  -2.53504   0.0116
ns(time, knots = c(12, 18, 24, 32, 36))4  -0.69488 0.2280401 414  -3.04719   0.0025
ns(time, knots = c(12, 18, 24, 32, 36))5  -0.04423 0.7130829 414  -0.06203   0.9506
ns(time, knots = c(12, 18, 24, 32, 36))6   0.79476 0.3328918 414   2.38744   0.0174
```

Table.3.13: REML estimates for the model fitted to haemoglobin concentration during pregnancy stratified by number smoke per day in first 3 months of pregnancy in Group 2.

```
  Linear mixed-effects model fit by REML
 Data: mumS
       AIC      BIC    logLik
  1736.281 1816.977 -850.1404

Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 1.09546784 (Inter
       time 0.03704626 -0.754
   Residual 0.56582344

Fixed effects: HAEM ~ SM * ns(time, knots = c(12, 18, 24, 32, 36))
                                              Value Std.Error  DF   t-value  p-value
                                  (Intercept)  9.727579  0.942448 408  10.32161  <.0001
                                     Multiple -2.459348  0.942448 246  -2.60953  0.0096
   ns(time, knots = c(12,18,24,32,36))1  1.235635  0.835646 408   1.47866  0.1400
   ns(time, knots = c(12,18,24,32,36))2  0.852563  1.045317 408   0.81560  0.4152
   ns(time, knots = c(12,18,24,32,36))3  1.774431  0.928684 408   1.91069  0.0567
   ns(time, knots = c(12,18,24,32,36))4  0.586949  0.630110 408   0.93150  0.3521
   ns(time, knots = c(12,18,24,32,36))5  5.605659  2.090149 408   2.68194  0.0076
   ns(time, knots = c(12,18,24,32,36))6  1.668844  0.849722 408   1.96399  0.0502
SM*ns(time, knots = c(12,18,24,32,36))1  2.166383  0.835646 408   2.59246  0.0099
SM*ns(time, knots = c(12,18,24,32,36))2  2.190033  1.045317 408   2.09509  0.0368
SM*ns(time, knots = c(12,18,24,32,36))3  2.882094  0.928684 408   3.10342  0.0020
SM*ns(time, knots = c(12,18,24,32,36))4  1.351705  0.630110 408   2.14519  0.0325
SM*ns(time, knots = c(12,18,24,32,36))5  6.219119  2.090149 408   2.97544  0.0031
SM*ns(time, knots = c(12,18,24,32,36))6  0.966519  0.849722 408   1.13745  0.2560
```

Table.3.14: REML estimate for the model fitted to haemoglobin concentration during pregnancy stratified by singleton /multiple in Group 2.

```
Linear mixed-effects model fit by REML

      AIC      BIC    logLik
  2077.77 2196.982 -1013.885

Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 0.71145696 (Inter
       time 0.02648804 -0.664
   Residual 0.54937845
Fixed effects: HAEM ~ Vg * ns(time, knots = c(12, 18, 24, 32, 36))
                                          Value  Std.Error  DF  t-value p-value
                                (Intercept)  12.31711 0.3642469 5713 3.81528  <.0001
                                   In past   0.34589 0.3330695 299  1.03848  0.2999
                                At present  -0.41102 0.3093499 299 -1.32865  0.1850
    ns(time, knots = c(12,18,24,32,36))1  -0.42216 0.3985421 571 -1.05926  0.2899
    ns(time, knots = c(12,18,24,32,36))2  -2.03444 0.4648937 571 -4.37613  <.0001
    ns(time, knots = c(12,18,24,32,36))3  -1.69474 0.3892975 571 -4.35333  <.0001
    ns(time, knots = c(12,18,24,32,36))4  -0.91140 0.2692930 571 -3.38441  0.0008
    ns(time, knots = c(12,18,24,32,36))5  -0.91976 0.8974197 571 -1.02490  0.3058
    ns(time, knots = c(12,18,24,32,36))6  -0.76284 0.3080375 571 -2.47646  0.0136
vg1*ns(time, knots = c(12,18,24,32,36))1  -0.21835 0.3423934 571 -0.63771  0.5239
vg2*ns(time, knots = c(12,18,24,32,36))1   0.66048 0.3460607 571  1.90856  0.0568
vg1*ns(time, knots = c(12,18,24,32,36))2  -0.08384 0.4225348 571 -0.19842  0.8428
vg2*ns(time, knots = c(12,18,24,32,36))2  -0.02139 0.3957452 571 -0.05404  0.9569
vg1*ns(time, knots = c(12,18,24,32,36))3  -0.78546 0.3772831 571 -2.08188  0.0378
vg2*ns(time, knots = c(12,18,24,32,36))3   0.70143 0.3226531 571  2.17396  0.0301
vg1*ns(time, knots = c(12,18,24,32,36))4   0.01407 0.2604322 571  0.05402  0.9569
vg2*ns(time, knots = c(12,18,24,32,36))4   0.05017 0.2234064 571  0.22458  0.8224
vg1*ns(time, knots = c(12,18,24,32,36))5  -1.19874 0.7992963 571 -1.49975  0.1342
vg2*ns(time, knots = c(12,18,24,32,36))5   1.23356 0.7696778 571  1.60270  0.1096
vg1*ns(time, knots = c(12,18,24,32,36))6  -0.29695 0.2835766 571 -1.04717  0.2955
vg2*ns(time, knots = c(12,18,24,32,36))6   0.38913 0.2609251 571  1.49133  0.1364
```

Table.3.15: REML estimates for the model fitted to haemoglobin concentration during pregnancy stratified by vegetarianism of mothers in Group 3.

```
Linear mixed-effects model fit by REML
 Data: edu32
       AIC      BIC    logLik
  2090.211 2162.163 -1030.105

Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 0.67581943 (Inter
       time 0.02544767 -0.637
   Residual 0.55145226

Fixed effects: HAEM ~ edu + ns(time, knots = c(12, 18, 24, 32, 36))
                                          Value Std.Error  DF   t-value  p-value
                                (Intercept)  12.27624 0.1916140 593  64.06755   <.0001
                                        CSE   0.03545 0.0788332 302   0.44964   0.6533
                                        VOC   0.08532 0.0321734 302   2.65195   0.0084
                                    O level   0.04767 0.0238031 302   2.00270   0.0461
                                    A level   0.00136 0.0242648 302   0.05611   0.9553
ns(time, knots = c(12,18,24,32,36))1  -0.76847 0.2053602 593  -3.74204   0.0002
ns(time, knots = c(12,18,24,32,36))2  -1.92231 0.2516510 593  -7.63880   <.0001
ns(time, knots = c(12,18,24,32,36))3  -1.59594 0.2154475 593  -7.40757   <.0001
ns(time, knots = c(12,18,24,32,36))4  -0.93228 0.1591262 593  -5.85876   <.0001
ns(time, knots = c(12,18,24,32,36))5  -0.87912 0.4662108 593  -1.88567   0.0598
ns(time, knots = c(12,18,24,32,36))6  -0.89757 0.2035224 593  -4.41020   <.0001
```

Table.3.16: REML estimates for the model fitted to haemoglobin concentration during pregnancy stratified by education level of mothers in Group 3.

```
Linear mixed-effects model fit by REML

      AIC     BIC    logLik
  2075.678 2133.28 -1025.839

Random effects:
 Formula:  ~ time | sub
 Structure: General positive-definite
                StdDev   Corr
(Intercept) 0.70566459 (Inter
       time 0.02596994 -0.66
   Residual 0.55018628

Fixed effects: HAEM ~ Parity+ ns(time, knots = c(12, 18, 24, 32, 36))
                                          Value Std.Error  DF   t-value   p-value
                                (Intercept)  12.64429 0.2253280 593  56.11502   <.0001
                                       >=2  -0.17977 0.0750917 305  -2.39403   0.0173
ns(time, knots = c(12, 18, 24, 32, 36))1  -0.77723 0.2070770 593  -3.75335   0.0002
ns(time, knots = c(12, 18, 24, 32, 36))2  -1.96063 0.2534752 593  -7.73500   <.0001
ns(time, knots = c(12, 18, 24, 32, 36))3  -1.62501 0.2169630 593  -7.48982   <.0001
ns(time, knots = c(12, 18, 24, 32, 36))4  -0.94605 0.1600702 593  -5.91019   <.0001
ns(time, knots = c(12, 18, 24, 32, 36))5  -0.96609 0.4712310 593  -2.05015   0.0408
ns(time, knots = c(12, 18, 24, 32, 36))6  -0.88659 0.2035694 593  -4.35521   <.0001
```

Table.3.17: REML estimates for the model fitted to haemoglobin concentration during pregnancy stratified by mother's parity in Group 3.

Name of covariates	Lower	Est.	Upper	SE	T-value	P-value
Parity:						
≥ 2	-0.19	0.06	0.32	0.13	0.49	0.63
Par*ns(time, knots = c(12, 18, 24, 32, 36))4	-0.48	-0.29	-0.10	0.1	-2.94	0.003
Par*ns(time, knots = c(12, 18, 24, 32, 36))6	0.10	0.40	0.70	0.15	2.64	0.008
Vegetarianism of mother:						
In the past	-0.17	-0.09	-0.02	0.04	-2.45	0.014
At the present	-0.09	-0.03	0.04	0.03	-0.76	0.45

Table.3.18: Coefficient values for the final model fitted to haemoglobin concentration during pregnancy in Group 1.

Name of covariates	Lower	Est.	Upper	SE	T-value	P-value
Number of smoked in 1st 3 months of pregnancy:						
≥ 1	-0.01	0.11	0.24	0.06	1.76	0.08
Ethnic of mothers:						
Non-White	-0.53	-0.26	0.02	0.14	-1.82	0.07
Singleton/Multiple:						
≥ 1	-8.65	-5.02	-1.39	1.85	-2.71	0.007
SM*ns(time, knots = c(12, 18, 24, 32, 36))1	1.15	4.37	7.59	1.64	2.66	0.008
SM*ns(time, knots = c(12, 18, 24, 32, 36))2	0.44	4.47	8.51	2.06	2.17	0.031
SM*ns(time, knots = c(12, 18, 24, 32, 36))3	2.27	5.85	9.44	1.83	3.2	0.001
SM*ns(time, knots = c(12, 18, 24, 32, 36))4	0.26	2.69	5.12	1.24	2.17	0.031
SM*ns(time, knots = c(12, 18, 24, 32, 36))5	4.52	12.59	20.65	4.12	3.06	0.002

Table.3.19: Coefficient values for the final model fitted to haemoglobin concentration during pregnancy in Group 2.

Name of covariates	Lower	Est.	Upper	SE	T-value	P-value
Parity:						
≥ 2	-0.18	-0.10	-0.83	0.04	-2.07	<.0001
Vegetarianism of mother:						
In the past	-0.27	0.42	1.07	0.33	1.25	0.21
At the present	-0.20	0.40	0.99	0.3	-1.31	0.19
In the past*ns(time, knots = c(12, 18, 24, 32, 36))3	-1.70	-0.95	-0.19	0.39	-2.46	0.014
At the present*ns(time, knots = c(12, 18, 24, 32, 36))3	0.05	0.71	1.34	0.32	2.23	0.026
In the past*ns(time, knots = c(12, 18, 24, 32, 36))5	-3.41	-1.78	-0.16	0.83	-2.15	0.032
In the past*ns(time, knots = c(12, 18, 24, 32, 36))6	-2.14	-1.16	-0.19	0.5	-2.33	0.020
At the present*ns(time, knots = c(12, 18, 24, 32, 36))6	0.13	0.74	1.34	0.31	2.4	0.017
Education of mother:						
Vocational	-1.09	-0.21	0.66	0.45	-0.48	0.63
O level	-0.22	0.13	0.47	0.18	0.72	0.50
A level	-0.10	0.14	0.37	0.12	1.16	0.25
Degree	-0.82	-0.40	0.01	0.21	-1.92	0.06
Deg*ns(time, knots = c(12, 18, 24, 32, 36))1	0.08	0.48	0.88	0.21	2.34	0.02
Deg*ns(time, knots = c(12, 18, 24, 32, 36))3	0.06	0.48	0.89	0.21	2.23	0.03
Voc*ns(time, knots = c(12, 18, 24, 32, 36))6	0.20	1.26	2.33	0.55	2.32	0.021
Deg*ns(time, knots = c(12, 18, 24, 32, 36))6	0.02	0.29	0.57	0.14	2.08	0.038

Table.3.20: Coefficient values for the final model fitted to haemoglobin concentration during pregnancy in Group 3.

CHAPTER 4

Haemoglobin Concentration in Children followed Longitudinally from 8 months to 7 years.

4.1. Introduction:

Children and Infants require more Iron compared with adults, primarily due to their fast level of growth. They are completely dependent on dietary Iron after four to six months old, in order to meet their physiological requirements. An Insufficient Iron supply will result in the decrease of Iron stores in the body which leads to Iron deficiency anaemia(Wharf, Fox, Fairweather, & Cook 1997).

Some nutritional problems in early life of children are highlighted by recent national data. In the United States, Iron deficiency anaemia is still relatively common in children aged 1 to 2 years old (Looker, Dallman, & Carroll 1997) .

A recent World Health Organization document on infant nutritional requirements has described Iron, Zinc and Calcium as nutrients problem in both the developing world and the developed world (WHO 1998).

When looking at childhood, Iron deficiency anaemia is even found amongst the economically well off. Low birth weight, early intake of cow's milk, rapid growth and inadequate dietary Iron intake, have been found to be the most important risk factors (Oski 1989) . However there are some effective preventive measures, such as Iron enriched infant formula milk as well as Iron fortified cereals. Preventing Iron deficiency is difficult throughout the developing countries, where Iron deficiency anaemia is epidemic and Iron enriched infant formula milk, and cereals are usually not available (Walter et al. 1993).

An appropriate Iron status have been found in infants who exclusively breast feed for long period of time (Pisacane et al. 1995). However this is not always possible a result Iron deficiency can occur, as Iron supplement may not be an option either for economical or cultural reasons. Iron status at birth present great variations amongst individuals, these differences can still be found in the Iron status of the same individuals at age 6, 9, and 12 months old (Michaelsen, Milman, & Amuelson 1995).

This might be an explanation why the Iron levels of some infants appears to be and remain sufficient, throughout their first year of life, even if they do not ingest enough daily Iron and Iron supplement.

In the first year of life, it seems that dietary Iron, is only one of the factors that influence Iron levels. It may be that the Iron absorption from Iron enriched infant formula milk and cereals is moderate and is inhibited by a number of components in the diet, such as poly phenols present in fruit and vegetables (Duggan et al. 1991).

4.2. Statistical method:

The aims of this part of the study were:

- To investigate various haemoglobin levels during childhood from age 8 months to 7 years old;

- To determine the effect of different time independent covariates, including maternal and child covariates, on haemoglobin levels;

- To determine the effect of some time-dependent covariates such as nutrient intakes on the haemoglobin levels of these children.

As before, the data for this study was also obtained from the Avon Longitudinal Study of Pregnancy and Childhood (ALSPAC).

The children from *Children in Focus* were invited to attend a research clinic at 8,12, 18,31,43,61 and 84 months old (the age was recorded in weeks). The children were measured and had a blood sample taken (Cowin, Emond , Emmett, & The ALSPAC Study Team 2001).

Special statistical methods are essential in order to accommodate individual correlations and illustrate repeated haemoglobin measures on each of the children.

Ignoring this correlation may result in inappropriate estimates of regression coefficients and invalid inferences concerning the question (Gregoire et al. 1997).The mixed model approach allows inclusion of fixed effects and random effects. Fixed effects model the influence of covariates on mean haemoglobin, while random effects model the haemoglobin correlation structure.

This model does not need equally spaced or equal numbers of haemoglobin observations per child (Manor & Kark 1996).

Missing data is a problem for statistical analysis in many fields of research, and we need to decide how to deal with the missing data appropriately each time.

Data is not often missing completely at random (MCAR), but the data in this study may be classifiable as missing at random (MAR), see Chapter 2. For this data, the missing value is likely to be unrelated to the value of outcome after controlling for another variable such as sex, ethnicity.

One of the most popular methods for handling missing data is list-wise data deletion that is used for handling database. That means, if a record has missing data for any one variable, that record is dropped from the analysis.

This approach is implemented as the default method for handling missing data by most statistical software packages (e.g., SAS, SPSS). This practice reduces the sample size, lowering the power of any tests, which have been used.

However, S-plus (a statistical package) has a good facility to deal with missing values without deleting whole cases, so S-plus was used in the present study to deal with missing values.

That means, to model the variables studied, between child haemoglobin levels and covariates for this cohort study, we used mixed effect models based on cubic spline estimated by the maximum likelihood method.

The fitted model predicts the average haemoglobin for the population at a given age, while controlling for statistically relevant covariates.

We used cubic spline to fit a curve to children's haemoglobin levels, this procedure produced a curve showing haemoglobin level changes, from base line to 7 years old (84 months) of follow up. We selected this analytic approach for reasons that were explained in Chapter 3.

4.3. Results:

A total of 1432 children were invited to the clinic at 8, 12, 18,31,43,61 and 84 months. As you can see in Table 4.2, of them 75.1%(1074), 64.2%(920), 56.9%(815), 40.8%(584), 50.1%(717), 44.7%(641) and 47.3%(677) percent attended respectively.

Table 4.1 shows that age of each child (to the nearest week) was recorded on each occasion, as was the child's haemoglobin level.

Covariates of interest in this phase of our investigation included maternal covariates and children covariates. Categorical variables were created for several maternal covariates that did not change during the study (time stationary covariates).

These included maternal age, vegetarianism of mother, education, number of cigarettes smoked per day in 1st 3 months, last 2 weeks, 32 weeks of pregnancy, parity and singleton /multiple pregnancy.

As a positive correlation between maternal haemoglobin and their children iron status has been found in a number of studies (Ziaei, Hatefian, & Togeh 2002), it was decided to assess the relation between haemoglobin levels in children in this cohort study and their mothers during pregnancy.

Therefore the mother-specific random intercept and random slope terms from the cubic spline model, which was fitted to the maternal data during pregnancy (Chapter 3), were included as time stationary covariates to evaluate the role of mother's haemoglobin levels on iron status during childhood.

Child covariates modelled as time stationary covariates include birth weight, sex and ethnicity of children.

Also in this chapter, much statistical interest centers on covariates that, change with time (time varying covariates), and contributes to the variation in haemoglobin levels.

Therefore time varying covariates were examined. These included weight of children and nutrient intake such as, vitamin C intake, Haem iron intake, Fat intake, Saturated fat intake, Monounsaturated fat intake, Energy intake, Calcium intake, Iron intake and Non-Starch Polysaccharide intake (NSP) as time varying covariate. These measurements were not available in all times point; the nutrient intakes were only made at 18 and 43 months.

In the data collection process, these data exist just at two time points across all ages for the observations. Unavailable data for these children at other ages can cause some restrictions in the analysis of data. The analyses were run, but it was anticipated that some results would not be statistically significant because of the small of amount of input data.

A scatter plot of all haemoglobin measures contributed between 8 months (33 weeks) and 7 years (426 weeks) is shown in Figure 4.1. A smooth curve fitted to observed haemoglobin measurements for children using cubic spline method in this Figure.

Haemoglobin levels in children slightly decreased from 33 to around 85 weeks then there was a substantial increase in children's haemoglobin levels until about age 7 years (426 weeks).

In Figure 4.1 can be observed the existence of change in children's haemoglobin over time.

Table 4.1 is used to convert weeks of attendance of children at clinic to month and year. For example, as the table shows, seven-year data is actually for children aged 357 weeks to 426 weeks.

As you can see in Table 4.3, of the 1432 children in this study 773 (54%) were male and 659 (46%) were female. There were 1394 (97.3%) singleton births and 38 (2.7%) twins in the data set. Only 3.6% (48) of children were non-white and the rest of them, that means 96.4% (1296) were white. The mean (SD) birth weight for these children was 3433g (529.6).

Of the total of mothers invited to the clinic, 196 (14.5%), 132 (9.8%), 481(35.7%), 343 (25.4%) and 196 (14.5%) were with CSE, Vocational, O level, A level, Degree level of education respectively, There are no answer for 65 of mothers for this covariate.

Of all mothers in this study 53.9% (733) of them were with a parity of two or more than born children and 46.1% (628) of mothers were with first-born children.

Majority of mothers (84.9%) were non-smoker in last 2 weeks of pregnancy and 15.1% were smoker with more than one cigarettes smoke per day, of the total mothers in this study, 80.3% did not smoked and 19.7% did smoked more than one cigarettes per day in 1st 3 months of pregnancy.

Eighty five percent (1108) of mothers did not smoke per day in 32 weeks of pregnancy and the rest of them did smoke that were (15.0%) of mothers. As you can see in Table 4.3 mean and standard deviation of maternal age in this study were 28.9 and 4.7 respectively.

Vegetarianism was not popular among the mothers, because eighty-five present of them are not vegetarian and 15% were vegetarian in the past or at the present.

In order to find the best position and number of knots for fitting a cubic spline model to these data, as in the previous chapter, several different combinations were tested.

First, a model including 6 knots at 46, 69, 108, 162, 228 and 324 weeks was tested. All cubic spline segments but 4 were significant. We continued by fitting models with different positions and numbers of knots but all models had at least one cubic spline segment which was not significant.

The only model that was significant on each consecutive pair of knots was the model with three (interior) knots at 69, 228 and 324 weeks.

Hence the knots at 69, 228 and 324 weeks are used in the cubic spline to capture the observed relationship between age and haemoglobin levels in this study population.

The use of the cubic spline, with three knots at 69, 228 and 324 weeks, produced the best fitting model to children's haemoglobin levels over the this period of time from 8 months to 7 years old age (33 weeks to 426 weeks).

Time independent covariates were first introduced individually to the fundamental model of haemoglobin levels over time, and then compared with the reduced model (model with time only). The log-likelihood statistic is used to find a significant effect of a covariate on haemoglobin levels.

The *lme* function in S-Plus provides two methods to estimate Mixed-effects Models: Maximum-Likelihood (ML) estimation of the model maximizes the likelihood with respect to all of the parameters of the model at the same time; include fixed effects and variance components.

Restricted Maximum-Likelihood (REML) estimation incorporates the fixed effects out of the likelihood and estimates the variance components and then, given the estimates of the variance components, estimates of the fixed effects are recovered.

Either ML or REML to construct this statistic model can be used that tests whether the full model is necessary or the reduced model.

As REML is used to construct the models in this study and because REML estimates are calculated incorporating out the fixed effects, one cannot legally do likelihood-ratio tests across models with different fixed effects when the models are estimated by REML. So for using ANOVA procedure, refitting the model to data by ML is necessary.

Significant covariates were then included as interaction terms with time and again tested by comparing the log-likelihood statistic of the reduced model with full model to study whether covariates modify the effect of time.

The model was then run again to find the effects of time-varying covariates of interest on haemoglobin levels in children. As the nutrient data are only available at two time points (18 and 43 months), it would be appropriate to delete the data for the rest of the times and fit an adjusted model for just these two times.

Differences by sex appeared in this children, as you can see in table 4.4 and also in Figure 4.2, boys had mean haemoglobin levels significantly below the girls means (P=0.007).

As you can see in Table 4.4, the mean haemoglobin level was also negatively associated with birth weight of children. The haemoglobin levels were similar for the all groups education of mothers, but there were significant negative interaction between children haemoglobin levels born to mothers with Vocational level of education in knots (,69) and children haemoglobin level born to mothers with Degree level in (228,324) with time respectively.

Figure 4.4 and Table 4.4 show children born to mothers with Vocational and Degree level of education had a significant decrease in mean haemoglobin than children born to mothers with other levels of education in these knots (P-Value=0.030, P-Value =0.012).

From the Figure 4.3 and Table 4.4, it is clear that parity of mother was associated with hemoglobin levels of children. Children born to mothers with first parity had a higher mean hemoglobin levels compared with children born to mothers with parity of two or more in this part of the study.

No differences in haemoglobin levels of children over time were detected for ethnicity of children, vegetarianism of mothers, Singleton /Multiple pregnancy, maternal age, number of cigarettes smoked per day in 1st 3 months, number of cigarettes smoked per day in last 2 weeks and number of cigarettes smoked per day in 32 weeks during pregnancy.

For assessment of time varying covariates, the data that were presented pertain to children at 18 and 43 months on whom data on nutrient intake were available. A cubic spline with a single knot at 135 weeks was used in these data, which this was the best fitting model to children's data.

Table 4.5 presents the coefficient values for all time varying covariates in Linear Mixed Model (nutrient intakes and weight) and hemoglobin levels in children.

As you can see in this Table, vitamins C intakes were positively associated with hemoglobin levels. Both Non Starch Polysaccharide intake (NSP) and Calcium intake were significantly positively associated with haemoglobin levels.

Weights of children also were positively associated with haemoglobin levels. That means, the lower weight groups of children may be at more risk of being anaemic. Neither iron intake nor haem iron intake were significantly associated with haemoglobin levels.

There was no association between Fat intake and haemoglobin levels. Saturated fat intake, Energy intake and Monounsaturated fat intake were not associated with haemoglobin levels.

There were no significant interaction between all covariates such as time varying covariates and time stationary covariates but education level of mothers and time in these children.

In this study, it is of interest to establish whether haemoglobin levels in mothers affect the iron status of their infants.

Both cross-sectional and longitudinal studies in the U.S. and also Europe have illustrated that even moderate anaemia can be negatively associated with obstetrical outcomes, including preterm delivery and low birth weight(Murphy et al. 1986).

Also there is no doubt that iron supplement is effective in improving the iron levels of mothers during pregnancy (Dawson & McGanity 1987) whereas infants born to women with iron deficiency also have a high prevalence of anaemia in the first 6 months (Preziosi et al. 1997).

Therefore, it is necessary to investigate the effect of haemoglobin levels of mothers during pregnancy on iron status of their children separately for mother in different iron supplement groups.

Table 4.6 illustrates some results, which are about the association between mother's haemoglobin levels during pregnancy and children's haemoglobin levels.

These were done separately for the three groups of mothers according to iron supplements.

This analysis took the coefficients (random intercept and random slope) from the model, which was fitted to the maternal data, and investigated whether these were related to the children's haemoglobin levels subsequently.

The random coefficients for individual mothers were added as two separate covariates to the cubic spline for the children's haemoglobin levels.

The results have shown that random slopes were significantly positively associated with children's haemoglobin level in Group 1 (Non iron supplement group) but random intercepts were not associated with children's haemoglobin level in this group.

This suggests that, in Group 1, mothers who had a relatively slow decrease in haemoglobin during pregnancy tended to have children with relatively higher haemoglobin levels.

We went on to look at the association between these coefficients in mothers in-group 2 and 3 (mothers who took iron supplement in week 18 and in 32 week) with children's haemoglobin level but neither of these coefficients was significantly associated with children's haemoglobin level. (This could indicate that iron supplementation during pregnancy, even if started quite late, is successful in bringing the mother's haemoglobin level back up to a point where it does not adversely affect the child's iron level).

Another risk of anaemia and iron deficiency is that mothers with these conditions may give birth to children with anaemia or iron deficiency and this may result in abnormal child development.

However, most research indicates that mothers who are iron deficient during pregnancy are more probable to give birth to iron deficient infants than mothers who have good iron status (Lao et al. 1991).

For making it clear whether haemoglobin levels in mothers during pregnancy are associated with iron status in children in early life, Table 4.7 shows the correlation between haemoglobin levels at 8 months and random intercepts and slopes of the models that were fitted to the maternal data during pregnancy.

Random intercepts and random slopes were not associated with haemoglobin levels at 8 months in Group 1.

In Group 2, random intercepts were positively associated with haemoglobin levels at 8 months although random slopes were not associated with haemoglobin levels.

Random intercept was significantly positively associated with haemoglobin concentration at 8-month in-group three whereas random slopes were not associated with haemoglobin levels at 8 months in same group.

4.4. Final Model:

The effect of all the different significant covariates including maternal education, parity, sex and birth weight, were tested as time stationary covariates in a final model. The relationship did not reach statistical significance for birth weight in the first model fitted. A model without birth weight was then fitted to test the rest of covariates. A statistically significant difference was observed depending on the maternal education, parity and sex, as you can see in Table 4.8.

New model fitted to data revealed a significant interaction between maternal education and time. A significant interaction was observed that indicated a negative effect of education at Voc level at knots (,69) and Degree level at knots (228,324) respectively.

As you can see in Table 4.8 and Figure 4.3, also there is a slightly positive significant interaction between parity and time in knots (228,324). That means that the difference in children's haemoglobin levels in mother with parity more than one tends to increase during this period of time.

No significant interaction existed between children's sex and time, so this term was dropped from the statistical analysis. Significance testing between different models to find the best final models was done using the likelihood ratio test.
In addition, the vitamin C, Calcium, NSP intakes and weight were included as time varying covariates in linear mixed model in order to find a final model. In the final model, there was statistically significant positive correlation between the mean haemoglobin level and Vitamin C intake, weight, while Calcium and NSP intakes did not change significantly.

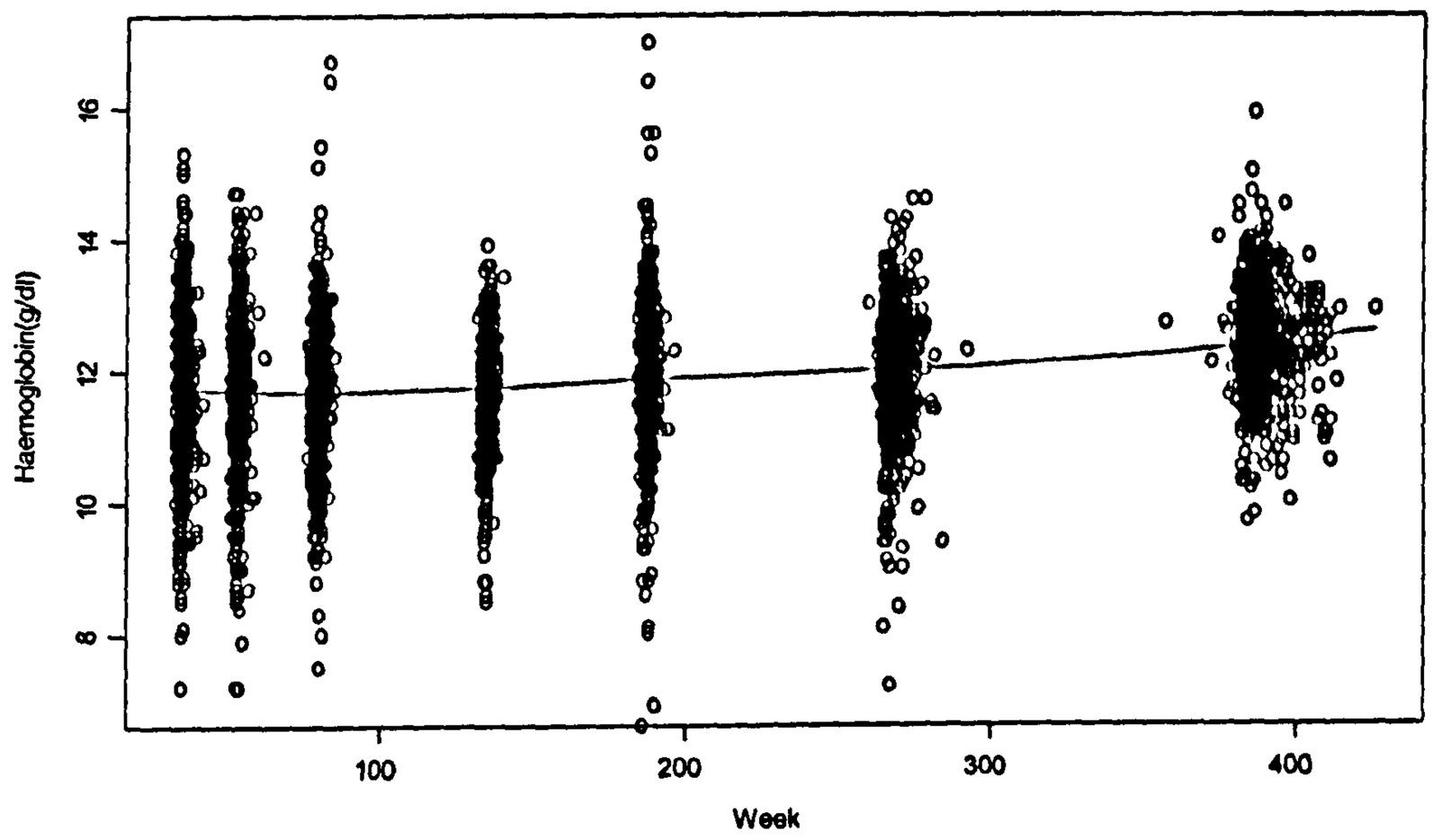


Figure.4.1: Cubic spline curve for children's haemoglobin levels against age (week).

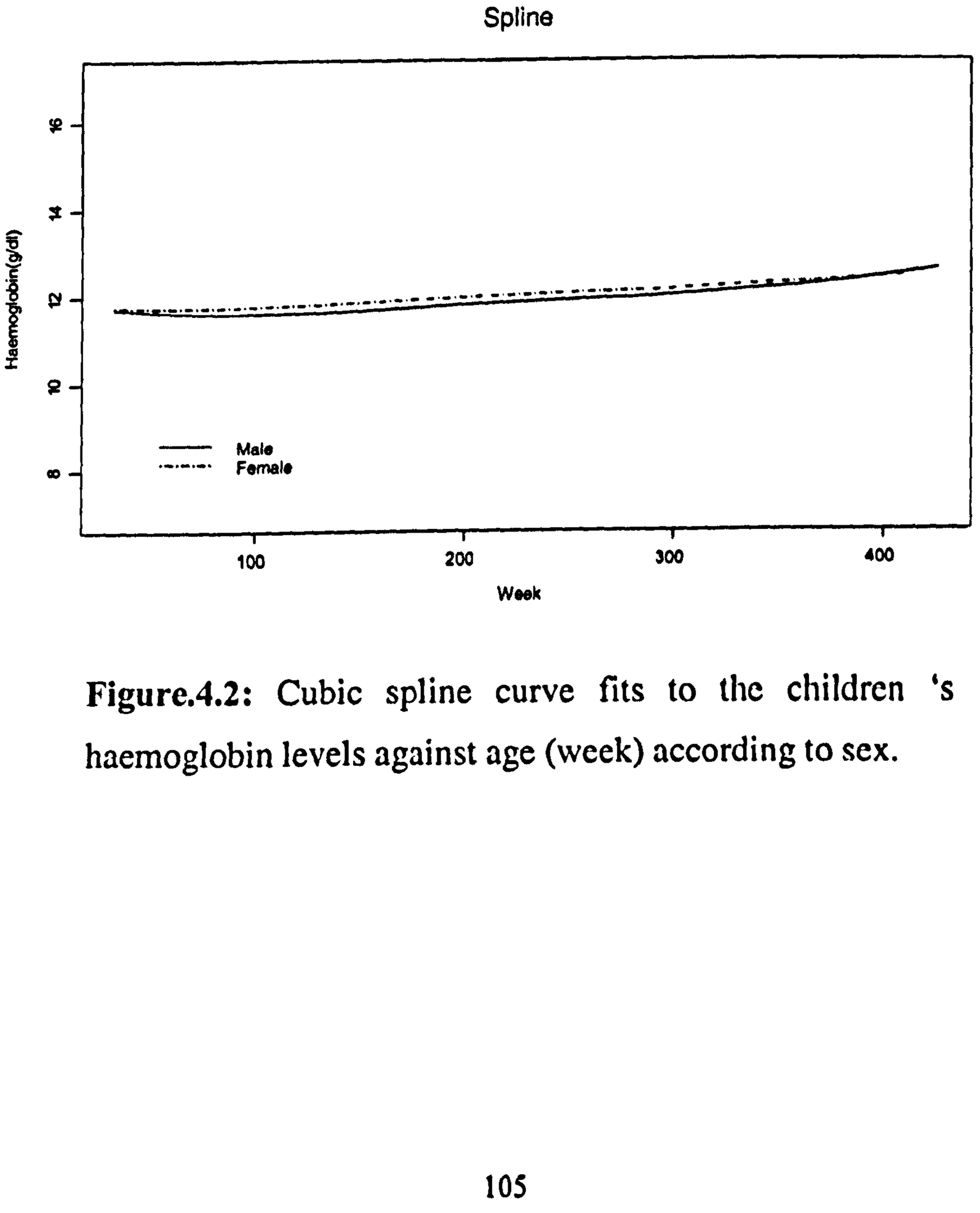


Figure.4.2: Cubic spline curve fits to the children 's haemoglobin levels against age (week) according to sex.

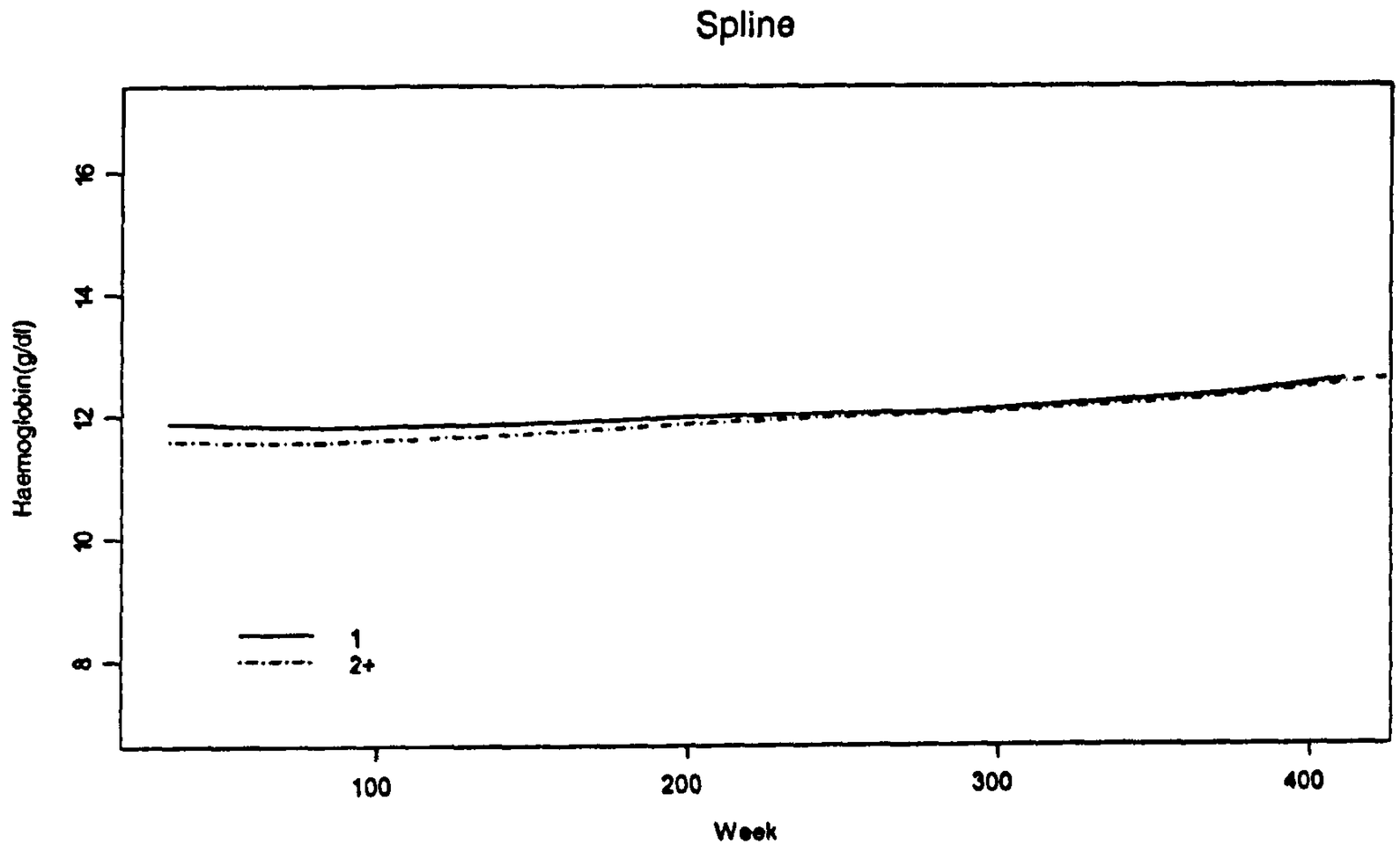


Figure.4.3: Cubic spline curve fits to the children's haemoglobin levels against age (week) according to parity of mothers.

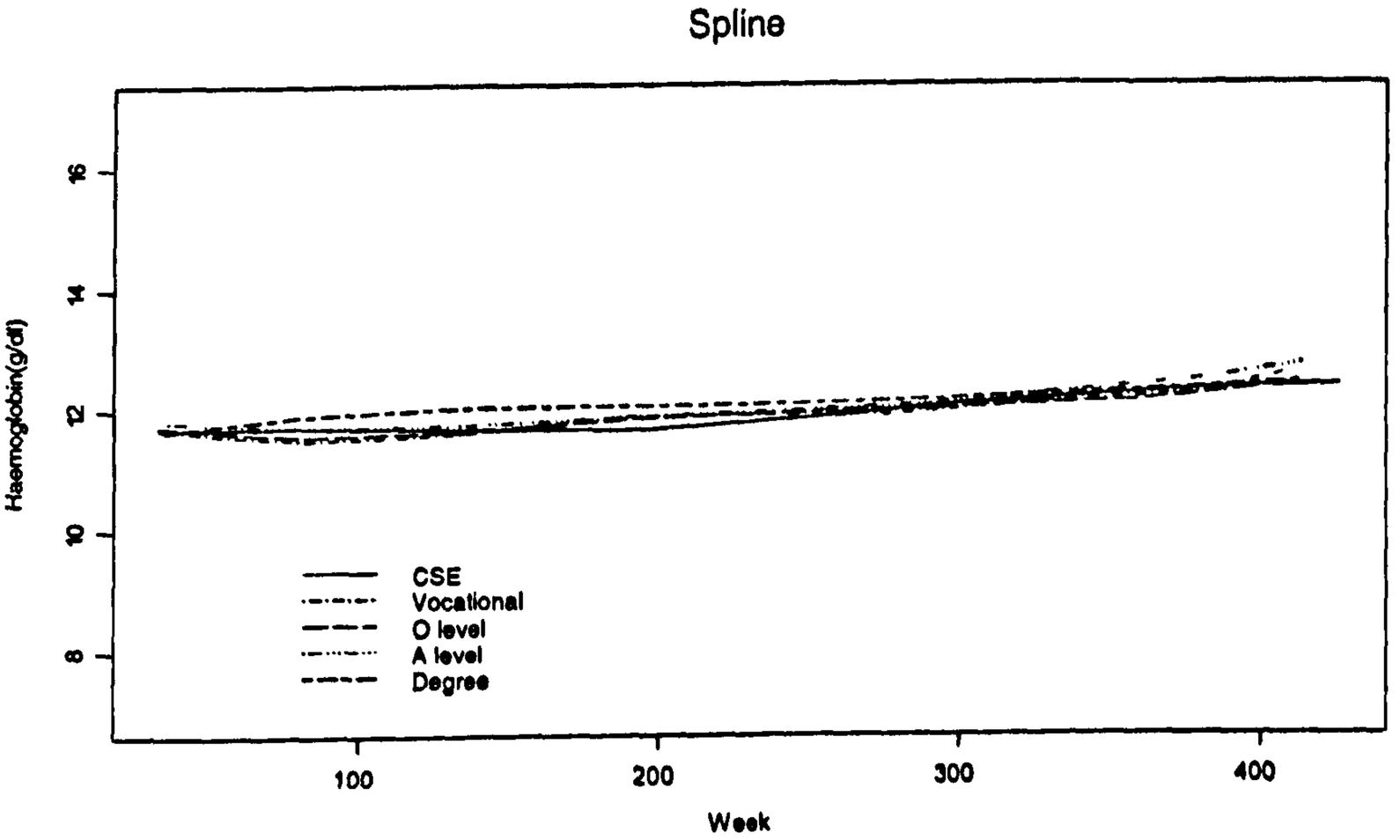


Figure.4.4: Cubic spline curve fits to the children 's haemoglobin levels against age (week) according to maternal education.

Week	Month	Year
33-42	8	-
51-62	12	1
76-85	18	1.5
132-140	31	~2.5
184-196	43	~3.5
260-292	61	~5
357-426	84	7

Table.4.1: Week's conversion to month and year.

		Hb level at 8 months	Hb level at 12 months	Hb level at 18 months	Hb level at 31 months	Hb level at 43 months	Hb level at 61 months	Hb level at 84 months
N	Valid	1074	920	815	584	717	641	677
	Missing	358	512	617	848	715	791	755
Mean		11.70	11.75	11.68	11.61	11.89	11.97	12.38
S.D		1.13	1.04	0.95	0.82	1.03	0.91	0.80

Table.4.2: Frequency distribution, mean and standard deviation for Children in Focus in 7 measurements.

Covariates	Frequency	Valid Percent*
Child Ethnicity;		
White	1296	96.4
Non-White	48	3.6
Missing	88	
Total	1432**	100.0
Birth Weight (g)		
Mean (SD)	3433 (529.6)	
Missing	13	
Sex;		
Male	773	54.0
Female	659	46.0
Total	1432	100.0
Singleton/Twin:		
Singleton	1394	97.3
Twin	38	2.7
Total	1432	100.0
Mothers Highest Education;		
CSE	196	14.5
Vocational	132	9.8
O level	481	35.7
A level	343	25.4
Degree	196	14.5
Missing	65	
Total	1413***	100.0
Maternal age		
Mean (SD)	28.85 (4.69)	
Number of smoked / day in 32 Weeks;		
None	1108	85.0
≥ 1	195	15.0
Missing	110	
Total	1413	100.0
Number of smoked / day in last 2 Weeks;		
None	1166	84.9
≥ 1	208	15.1
Missing	39	
Total	1413	100.0
Number of smoked / day in 1st 3 months;		
None	1103	80.3
≥ 1	271	19.7
Missing	39	
Total	1413	100.0
Parity;		
1	628	46.1
≥ 2	733	53.9
Missing	52	
Total	1413	100.0
Vegetarianism of mother		
Never	1064	85
In the past	111	9
At present	71	6
Missing	167	
Total	1413	100.0

Table.4.3: Frequency and percentage of time stationary covariates.

* The percentages are based only on the cases who actually answered the question.

**n=1432(children)

***n=1413(mothers)

Name of covariates	Lower	Est.	Upper	SE	T-Value	P-value
Ethnic of children: Non-White	-0.16	-0.06	0.05	0.05	-1.02	0.31
Birth weight (Kg):	-0.15	-0.09	-0.03	0.03	-2.94	0.003
Sex of children: Female	0.01	0.05	0.09	0.019	2.72	0.007
Singleton/multiple: ≥ 1	-0.08	0.02	0.13	0.05	0.43	0.67
Education of mother:						0.02
Vocational	-0.12	0.01	0.13	0.07	0.14	0.89
O level	-0.04	0.02	0.07	0.03	0.54	0.59
A level	-0.01	0.03	0.07	0.02	1.50	0.13
Degree	-0.06	-0.02	0.02	0.02	-1.15	0.25
Voc*ns (time, knots = c (69, 228, 324)) 1	0.04	0.39	0.74	0.18	2.17	0.03
Deg*ns (time, knots = c (69, 228, 324)) 3	0.03	0.13	0.24	0.05	2.53	0.01
Maternal age:	- 0.004	0.003	0.01	0.004	0.71	0.48
No of Cig smoked in 32 wks: ≥ 1	-0.03	0.02	0.08	0.03	0.81	0.42
No of cig smoked in 1st 3 months: ≥ 1	-0.02	0.03	0.07	0.02	1.08	0.28
No of Cig smoked in last 2 wks: ≥ 1	-0.04	0.02	0.07	0.03	0.56	0.58
Parity: ≥ 2	-0.24	-0.17	-0.10	0.04	-4.52	<.0001
Vegetarianism of mother:						0.65
In the past	-0.05	0.02	0.09	0.03	0.5	0.62
At the present	-0.07	-0.008	0.05	0.03	-0.27	0.79

Table.4.4: Coefficient values for time stationary covariates in Linear Mixed Model.

Name of time varying covariates	Lower	Est.	Upper	SE	T-Value	P-value
Vitamin c intake	0.0007	0.002	0.003	0.0006	3.14	0.002
Haem iron intake	-0.13	0.10	0.32	0.12	0.84	0.4
Fat intake	-0.003	0.001	0.005	0.002	0.56	0.57
Saturated fat intake	-0.005	0.003	0.012	0.004	0.71	0.48
Monounsaturated fat intake	-0.012	-0.0003	0.002	0.006	-0.05	0.96
Energy intake	-0.002	0.0002	0.002	0.0001	1.37	0.17
Calcium intake	0	0.0002	0.0004	0.0001	1.98	0.049
Iron intake	-0.009	0.018	0.05	0.01	1.25	0.21
Non Starch Polysaccharide intake (NSP)	0.003	0.03	0.05	0.01	2.31	0.02
Weight	0.016	0.05	0.08	0.02	2.89	0.004

Table.4.5: Coefficient values for time varying covariates in Linear Mixed Model.

Name of the covariates	Lower	Est.	Upper	SE	T-Value	P-value
Group 1;						
Random intercept	-0.02	0.002	0.02	0.01	0.15	0.88
Random slope	0.79	3.75	6.72	1.51	2.48	0.014
Group 2;						
Random intercept	-0.05	-0.02	0.01	0.01	-1.30	0.19
Random slope	-3.79	-0.72	2.35	1.57	-0.46	0.65
Group 3;						
Random intercept	-0.01	0.02	0.04	0.02	1.03	0.31
Random slope	-2.34	2.63	7.59	2.54	1.04	0.30

Table.4.6: Coefficient values for random intercept and random slope of the model fitted to maternal data, as covariates in Linear Mixed Model fitted to children's haemoglobin.

	Random Intercept	Random slope
Haemoglobin levels at 8 months in Group 1	0.074 *P-value* = 0.0795	0.073 *P-value* = 0.082
Haemoglobin levels at 8 months in Group 2	0.163 *P-value* = 0.028	-0.143 *P-value* = 0.054
Haemoglobin levels at 8 months in Group 3	0.130 *P-value* = 0.042	0.048 *P-value* = 0.453

Table.4.7: Pearson correlation between Haemoglobin level at 8 moths and the coefficients (random intercept and random slope) from the model, which was fitted to the maternal data.

Name of time stationary covariates in final model	Lower	Est.	Upper	SD	T-Value	P-valu
Education of mother:						
Vocational	-0.16	-0.03	0.10	0.07	-0.47	0.64
O level	-0.04	0.01	0.06	0.03	0.41	0.68
A level	-0.01	0.02	0.06	0.02	1.27	0.20
Degree	-0.06	-0.03	0.01	0.02	-1.39	0.16
Voc*ns(time, knots = c(69, 228, 324))1	0.02	0.38	0.73	0.18	2.09	0.04
Deg*ns(time, knots = c(69, 228, 324))3	0.04	0.14	0.24	0.05	2.62	0.01
Parity:						
>1	-0.2	-0.14	-0.08	0.03	-4.57	<.0001
>1 *ns(time, knots = c(69, 228, 324))3	0.02	0.20	0.38	0.09	2.15	0.032
Sex:						
Female	0.01	0.05	0.09	0.02	2.56	0.01

Table.4.8: Coefficient values of time stationary covariates in Final Model

Name of time varying covariates in final model	Lower	Est.	Upper	SD	T-Value	P-value
Vitamin C intake	0.0007	0.002	0.003	0.0006	3.17	0.002
Calcium intake	-0.00006	0.00014	0.0003	0.0001	1.34	0.18
Non Starch Polysaccharide intake (NSP)	-0.01	0.02	0.04	0.011	1.48	0.14
Weight	0.02	0.06	0.09	0.019	2.93	0.004

Table.4.9: Coefficient values of time varying covariates in Final Model.

CHAPTER 5

Constructing Reference Curves for Haemoglobin Levels in Children

5.1. Introduction:

In this chapter we aim to introduce a longitudinal reference curve for haemoglobin levels in childhood, from age 8 months to 7 years.

A reference curve is a graph illustrating how the distribution of some interesting observations changes with age. This may be achieved by displaying the median and different percentiles over ages. An $\alpha\%$ age-related curve is defined by the range between two reference curves, which include $\alpha\%$ of the data at each individual age. For example measurements between the 15^{th} and 85^{th} centile curves lie within a 70% age-related range (Wright & Royston 1997).

Until the decade of the 1970's the normal range was commonly used instead of a reference curve, each value outside the range between the 2.5^{th} and 97.5^{th} centiles was abnormal data. One mistake of this definition is that 5% of normal population values will be outside the normal range(Wright & Royston 1999).

Wright and Royston in 1997 reviewed methods for constructing such reference curves. After an appropriate transformation of the measurements which are approximately normally distributed, the median and mean become equivalent and the percentiles are simple functions of the mean and standard deviation (Scheike, Zhang, & Juul 1999).

Nevertheless, when such a reference curve is used to follow up individuals, a number of problems appear as a result of the how centiles have been estimated.

In general cross-sectional reference curves ignore information on the prior observation of the individual and the individual measurement at each time is assessed separately. It seems that when longitudinal data are available, better results could be reached by conditional chart method.

The approach used in this chapter combines data measured on an individual at previous time points to provide a means of evaluating a later observation. As an example of the type of problem that might arise, once the previous individual information is ignored, assume that a woman's haemoglobin levels during pregnancy have been continuing along at the 75^{th} percentile for number of weeks and then suddenly drops to the median at the next clinic visit.

From a cross-sectional reference curve, this mother is still well within any reference curve and so there is no reason for alarm. However, given the history of longitudinal data of the mother up to that time, the sudden drop may be a warning sign. A conditional reference curve, which adjusted an individual's expected route according to previous measurement, would better be able to draw attention to this event.

Likewise, if haemoglobin gain of the woman went above the 90th percentile occasionally, it would not necessarily be construed as an abnormal haemoglobin gain (Fatti, Senaoana, & Thompson 1998).

This Chapter describes the extension of the reference curves approach of Cole and Green(Cole & Green 1992) to generate conditional reference curves (Cole 1994) for haemoglobin levels between 8 months and 7 years old in Children in Focus.

In this study, we construct cross-sectional and longitudinal reference curve for haemoglobin levels in children using LMS methods.

Firstly, cross-sectional reference curves were fitted to the data using Cole's LMS method (Cole & Green 1992) ,and then cross-sectional approach have been developed to conditional reference curve where haemoglobin levels (time t) was predicted from haemoglobin levels one previous time point (time $t-1$) and then it has been rearranged to give an other conditional reference curve by extending one previous time point to all previous times point , this method were derived using Cole's method (Cole 1994) .

After cross-sectional reference curve for haemoglobin levels were constructed and also conditional reference curves for haemoglobin levels calculated from observations obtained from Cole's method, then these reference curves were compared with each others.

5.2. Methods:

The references curves for haemoglobin levels were derived using the LMS method described by Cole and Green, 1992.

A normal distribution for constructing reference curves at every age in sample values is unnecessary in the LMS method assumption (Cole & Green 1992) since the data can be transformed to an approximately normal distribution via the Box-Cox power transformations.

The Box-Cox power transformation (Box & Cox 1964) :

$$\begin{cases} g(Y) = (Y^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ g(Y) = \mathrm{Ln}(Y) & \lambda = 0 \end{cases}$$

It is almost inevitably the popular transformations system to normality. Once the distribution of Y is negatively skewed, values of $\lambda > 1$ are used, when Y is normal, $\lambda = 1$ and when Y is positively skewed $\lambda < 1$ is used.

Using the LMS method in its original version(Cole 1988), firstly the Box-Cox distribution parameters for Y are estimated within each group of age . A useful parameterization in which the three parameters L, M and S are the skewness, median and coefficient of variation of Y respectively was proposed by Cole in 1988. The parameters were estimated with in each group of age by maximum likelihood separately and then smoothed across age. Regression smoothing procedure such as polynomial regression may be applied.

Under the LMS method, the Z score is as follow:

$$Z = \frac{\left[Hb/M\right]^{L} - 1}{L\,S} \qquad (5.1)$$

These Z scores have approximately a standard normal distribution. As haemoglobin level is a positive value, there is no difficulty in fitting Box–Cox power transformations to these data.

A nonparametric aspect was added by Cole and Green in 1992 to the original LMS method, using maximum penalized likelihood to estimate the age-related curves for each of the parameters L, M and S by natural cubic splines.

The smoothness of curve fitting across age is controlled by the values of the parameters called equivalent degree of freedom (e.d.f) for three parameters.
The following is formula which defined the points on each reference curve (Cole, Freeman, & Preece 1998):

$$M(1+LSz_{\alpha})^{1/L} \qquad (5.2)$$

Where L, M and S are values of the fitted curves at individual age, Z is score for the centile that was required (The normal equivalent deviate).
It is a profit of the LMS methods that the haemoglobin levels can be converted to Z scores that is the distance SD score concerning to the haemoglobin from the values of L, M and S at each age, with the formula (5.1) (Cole 1994).
The disadvantage of this method is that the choice of e.d.f.'s is somewhat subjective because statistical inference for use with penalized likelihood is not well understood, and determination of the best fitting curves may not be unique (Wright & Royston 1997) .
The conditional reference curve for one previous time point obtained using linear regression approach where haemoglobin levels at time *t* is predicted from haemoglobin measurement at time *t-1*, then:

$$\textbf{Haemoglobin}_{t} = a_t + b_t\,\textbf{Haemoglobin}_{(t-1)} + \textbf{error} \qquad (5.3)$$

Where

b_t =The slope coefficient

a_t =The intercept

Error term~ $N(0, \sigma_t^2)$

For using the SD score, the formula in (5.3) becomes:

$$Z_t = a'_t + b'_t Z_{t-1} + error \qquad (5.4)$$

Z_t =The SD scores at age t

Z_{t-1} =The SD score at age t-1

b'_t = The slope coefficient

a'_t = Intercept

The means and standard division of Zs are zero and unit respectively, however b'_t is the correlation between Z_t and Z_{t-1}, and a'_t is the intercept that are equal to r_t and zero respectively.

Then (5.4) changes to

$$Z_t = r_t Z_{t-1} + error \qquad (5.5)$$

The meaning is $E(Z_t | Z_{t-1}) = r_t Z_{t-1}$.

For data with a normal distribution the two approaches are the same, but in a distribution that is skewed, the SD score scale is more suitable (Cole 1994).

We have discussed now haemoglobin levels might be predicted from haemoglobin levels at one previous time point. Now, we could extend equation (5.3) to include haemoglobin levels in all previous times as follows:

$$\mathbf{Hb}_t = a_t + a_{t-1}\mathbf{Hb}_{t-1} + a_{t-2}\mathbf{Hb}_{t-2} + \ldots + a_1\mathbf{Hb}_1 + \mathbf{error} \qquad (5.6)$$

Where

$a_{t-1}, a_{t-2}, \ldots, a_1$ =The slope coefficients

a_t = The intercept

Error term~ $N(0, \sigma_t^2)$

Significant slope coefficient at different times show that haemoglobin level in these times could be an important predictor for conditional reference curves.

Multiple regression analyses were performed to examine the role of haemoglobin at various times. The variables, which were entered as independent variables are as follow:

Haemoglobin levels at time $t-1, t-2, ..., 1$. Significant explanatory variables were selected using stepwise regression modelling.

To obtain conditional reference curves by using regression base approach, the intercept terms were ignored when these were not statistically significant or were indeed very weak as you can see in Table 5.2.

The LMS method with smoothing by maximum penalised likelihood was carried out using the Windows 2000 supported LMS software (Cole, personal communication, 1999).

The reference curves representing the 3rd, 10th, 25th, 50th, 75th, 90th, and 97th centiles between 8 and 84 months old are given.

The convenience and flexibility of the recent method are hugely better than original method (Wright & Royston 1997).

The *e.d.f* for each curve is comparable with the degree of freedom of a polynomial, and range from 2 upwards. The *e.d.f* equal 2 corresponds to a straight-line smoothed curve, the *e.d.f* 3 find a quadratic curve, as the *e.d.f* increase the curves become rougher spline curves.

In our study we used 3, 5 and 3 as *e.d.f*, which gave a convenient amount of smoothing for the data.

Figure 5.1 shows the L curves for haemoglobin levels in children from 8 month to 7 years, over a range of fitted *e.d.f* between 2 and 4.

The L curve with 2 *e.d.f* in Figure 5.1(a) corresponds to a straight line and it is unlikely except for large data with wide age range.

In comparison with Figure 5.1 (c), the Figure 5.1 (b) is reasonably good smoothed over the whole rang of age.

Cole and Green 1992 believe that the *e.d.f* required for the M curve is greater than for the L and S curves.

As you can see in Figure 5.2, the benefit of fitting the M curve with 5 *e.d.f* is that *e.d.f* with this value allows for the fitted seven suitable smooth reference curves to the data while with increase *e.d.f* reference curve become raggedness.

Figures 5.3 present the fitted S curves obtained by the various *e.d.fs* included 2, 3 and 4 respectively. The smallest *e.d.f* is 2 for the S curve, corresponding to a straight line and it provides a low quality fitted to the data as shown in Figure 5.3(a).

The S curve in Figure 5.3(b) demonstrates a smooth curve and the variation of shape in the curve is clear, whereas in Figure 5.3(c) the curve is a ragged curve and became more complex compared with Figures 5.3(a) and 5.3(b).

5.3. The Results:

Table 5.1 shows the correlation matrix for haemoglobin SD score at each of the seven measurement age groups from 8 month (33 weeks) to 84 months (426 weeks) based on Children in Focus.

As you can see in this Table, in the most cases, the nearest SD scores in different time points were more correlated.

Firstly, a linear regression model was fitted to find slope coefficients to produce a conditional reference curve. The results in this point illustrated that the coefficient of variation (S) for conditional reference curve using data at all previous times was bigger than the coefficient of variation in conditional reference curve using one previous time point.

Also conditional reference curves (one previous time) were narrower than the conditional reference curve with all previous time point.

It was thought that this effect might have been due to patterns of missing data. For dealing with missing values, a single imputation method was used, but it was not a beneficial solution for this problem because the coefficient of variation (S) for the conditional reference curve using data at all previous times was still bigger than the coefficient of variation in the conditional reference curve using one previous time point. This explains why the conditional reference curves (all previous times) were wider than the conditional reference curves with one previous time.

As an alternative explanation, it was considered whether there were any gross outliers that might affect the regression analysis underlying the construction of the longitudinal reference curves. On investigation, we discovered several outliers that were more than 4 standard deviations from the sample mean. We repeated the analysis excluding these cases. Again, the conditional reference curves were produced with and without the use of an imputation method to deal with the missing data.

Figures 5.1, 5.2 and 5.3 show the L, M and S curves for children from 8 month (33 weeks) to 84 months (426 Weeks), with some variation in the *e.d.fs*.

Figures 5.5, 5.6 and 5.7 demonstrate the conditional fitted L, M and S curves obtained by using one previous time for children and figures 5.9, 5.10 and 5.11 show the L, M and S (all previous times), setting the *e.d.f* 3, 5 and 3 respectively.

Figures 5.13, 5.14, 5.15, 5.17, 5.18 and 5.19 show the L, M and S obtained by using one previous time and all previous times after used an imputation method (mean of series method) respectively.

As you can see in Figures 5.1(b), 5.5, 5.9, 5.13 and 5.17 the amount of skewness were measured by sample L curves, which is different across the methods. The L values in cross-sectional method are consistently between 1.5 and 3 and this curve in Figure 5.1(b) shows a peak during 250 to 300 weeks.

Figure 5.5 shows the conditional L curves fitted by one previous time. A increase the skewness to around 4 in this Figure indicate data that are skewed left up to 200 week (50 months) and after that there is a sharply decrease in the L curve in the whole rest of age.

Conditional Box-Cox power fitted by all previous times demonstrate in Figure 5.9 is between around 1 and 2.5 and the highest value is in the early of childhood then the L curve decreased continuously up to 7 years old.

Figure 5.13 and 5.17 illustrate refitted L curves obtained by using one previous time and all previous times after using the imputation method and these curves show clear trend in skewness with age.

In Figure 5.13, There is a rise in the Box-Cox power from the beginning of childhood up to around 130 weeks old age and then levelled off at a value about 3.8 until near 230 weeks after that the L curve decline steeply to minimum values in this Figure.

Figure 5.17 represents same manner with conditional L curve fitted by all previous time before used the imputation method in Figure 5.9, except in extreme point of children age in this study that there is an increase in the L curve.

Figures 5.2, 5.6,5.10, 5.14 and 5.18 present the M curves for haemoglobin levels whose are fitted by different methods in Children in Focus .The all M curves demonstrate high haemoglobin levels in early of measurements, followed by a fall and then a rise continuing until end of measurements. The minimum of median haemoglobin was about 11.7 (g/dl) around 130 weeks of the age.

Figure 5.6 demonstrates that the M curve from early of children life up to around 130 weeks tends to fall, but after that, there is rising sharply to the rest of the ages. The velocity in median haemoglobin for conditional curve (one previous time) in Figure 5.6 was larger than the M for others, particularly at older children.

The LMS method estimates the coefficient of variation (S curve) for haemoglobin levels during the centile fitting process, Figures 5.3, 5.7, 5.11, 5.15 and 5.19 have shown the S curves for the different methods used in this study.

The coefficient of variation in the cross-sectional method is greater than in the others, as expected.

The conditional S curve fitted with all previous time is smaller than other S curves.

Whereas the conditional coefficient of variation fitted with all previous time after used the imputation method is greater than other conditional coefficient of variation.

As you can see in Figure 5.3(b), The S curve demonstrates that the coefficient of variation is decreased during whole ages and the S curve varies between 0.06 and 0.1.

Otherwise in Figure 5.7 the S curve started to increase in the beginning of childhood until near 200 weeks (50 months) and it decline over rest of ages but not very variable. The varying of S curve is narrowly from above 0.02 to under 0.03.

The coefficient of variation, which is fitted by all previous times, shows in Figure 5.11. The overall trend is as same as the coefficient of variation in Figure 5.7, in the beginning is to increase with age up to 200 week (50) then show decrease with age.

The conditional S curve which is fitted by one previous time after used the imputation method seem to be lower than the other conditional coefficients of variation (S), as shown in Figure 5.15.

Figures 5.4, 5.8, 5.12, 5.16 and 5.20 demonstrate the reference curves were fitted using Cole's LMS method, which adjusts haemoglobin levels distribution at different methods.

In these Figures you can see seven centiles, from 3^{rd} to the 97^{th} for haemoglobin levels in Children in Focus, obtain from the LMS curves using equation (5.2).

The reference curves are reasonably well smoothed over the whole age range. However the corresponding centiles for older children are convincingly smooth, showing that the extent of smoothing is greater for the old children than the young children.

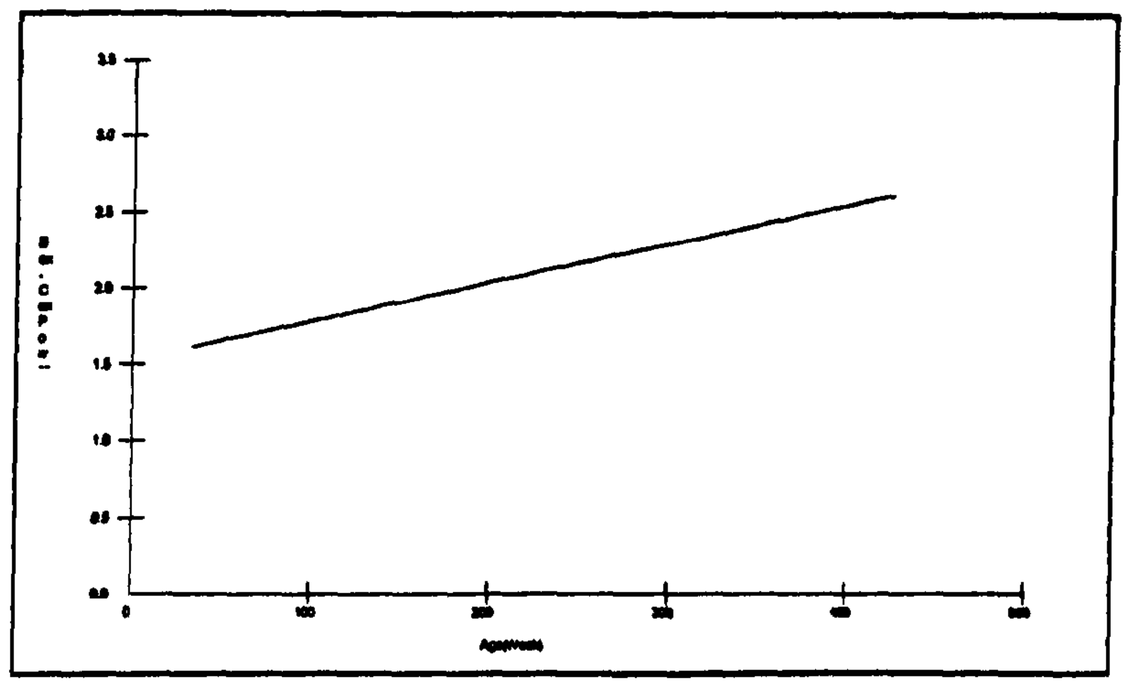
The 3rd, 10th, 25th, 50th, 75th, 90th,and 97th centiles for haemoglobin levels are given in Figure 5.4. This is for children between 8 months to 7 years old in Children in Focus as a cross-sectional study.

At early childhood in cross-sectional reference curve, the references curves are wider, but thereafter the centiles get closer to the median.

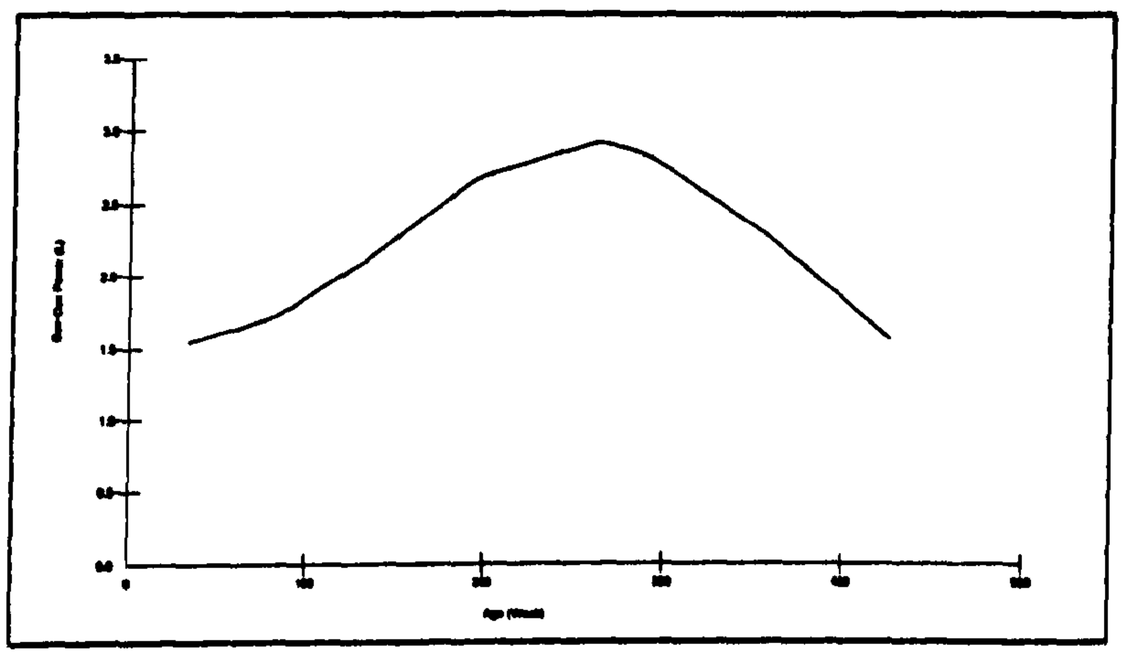
Figure 5.8 shows seven conditional reference curves fitted by one previous time; each conditional centile represents the conditional median pattern of haemoglobin levels of children.

In Figure 5.8 is evident that, these centiles cross all distance and most obviously during early age up to 150 weeks (37 months) are almost parallel and then became wider in the older children.

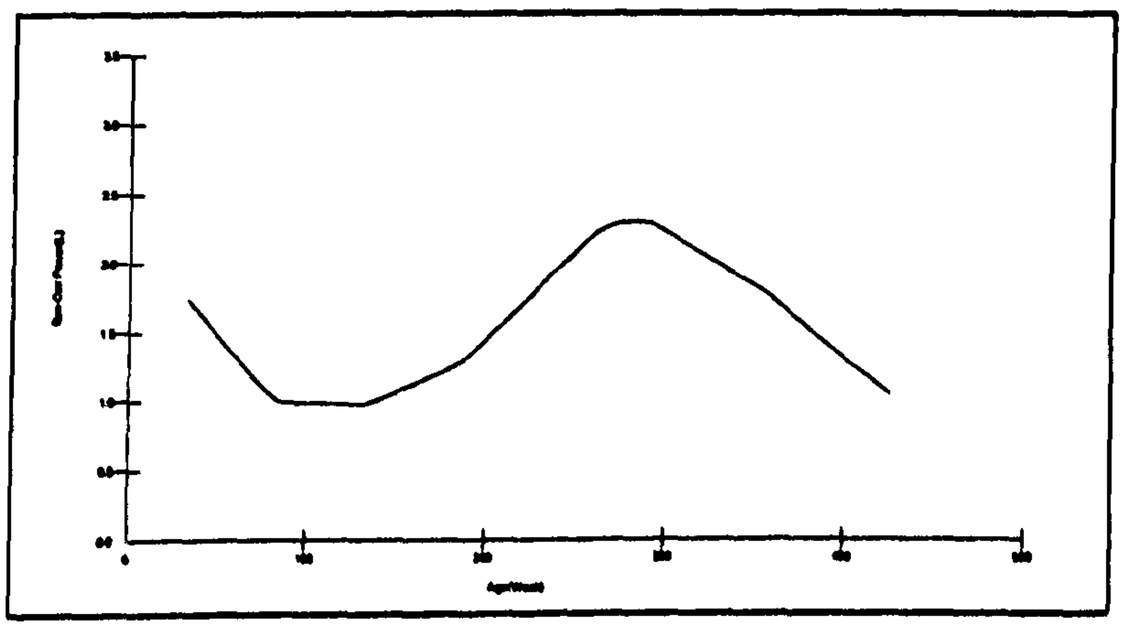
The reference curves for haemoglobin levels in this Figure are similar to the reference curves in Figure 5.12, 5.16 and 5.20, This means that all conditional reference curves are much closer to the conditional median than cross-sectional reference curves, particularly in the conditional reference curves which use data from all the previous times without use the imputation method.



(a)

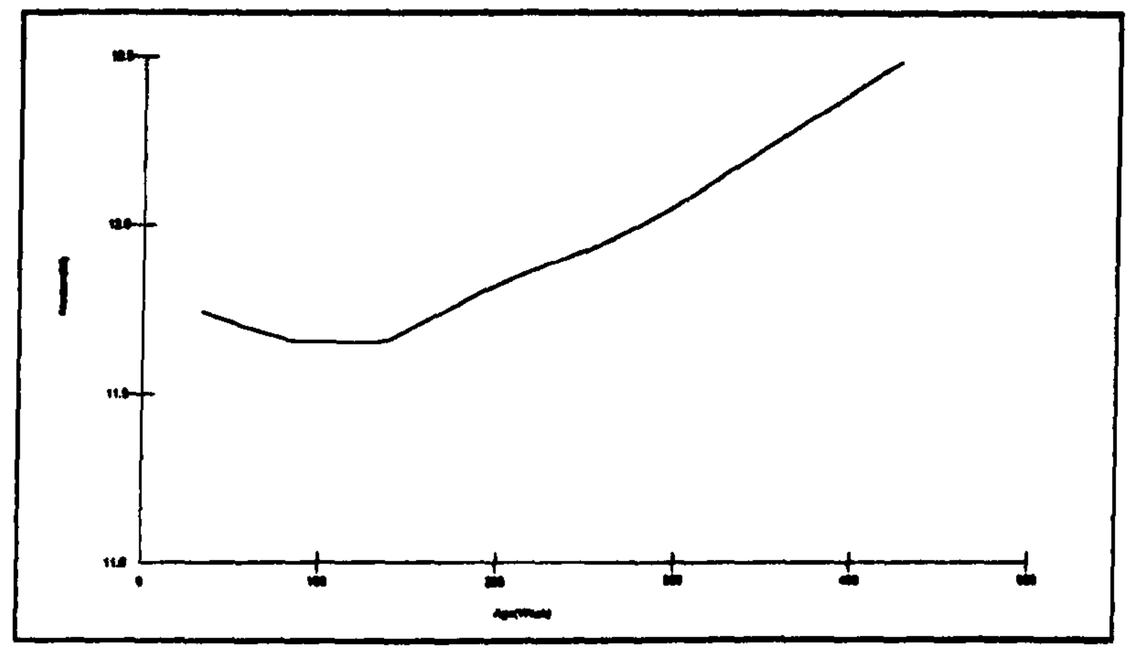


(b)

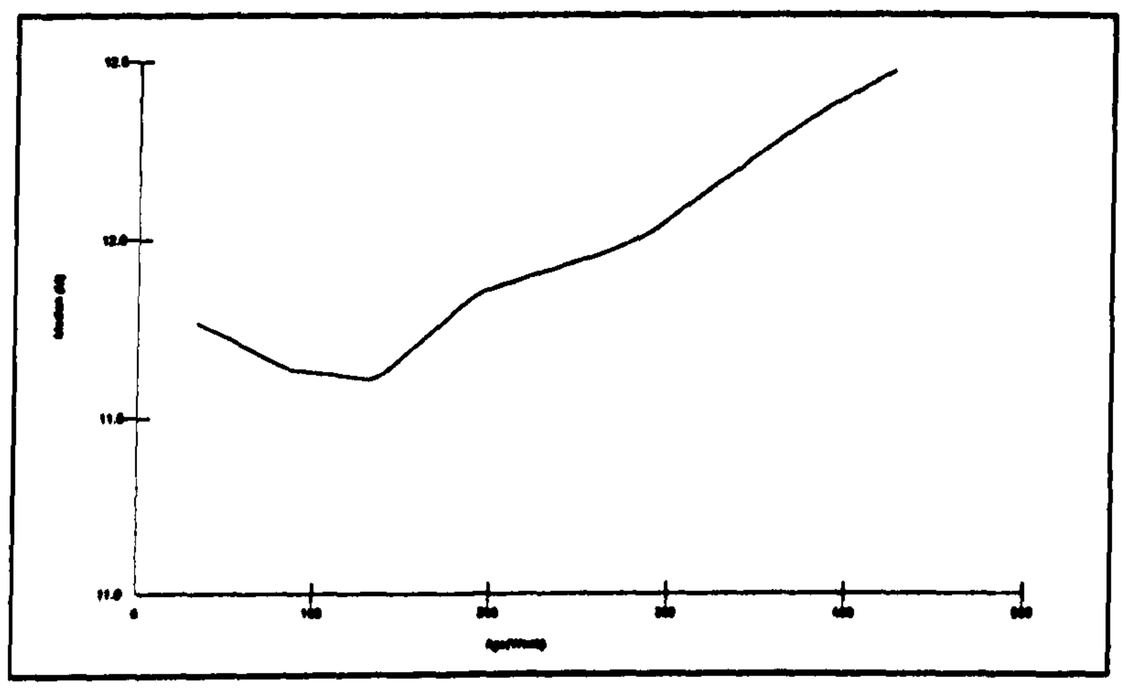


(c)

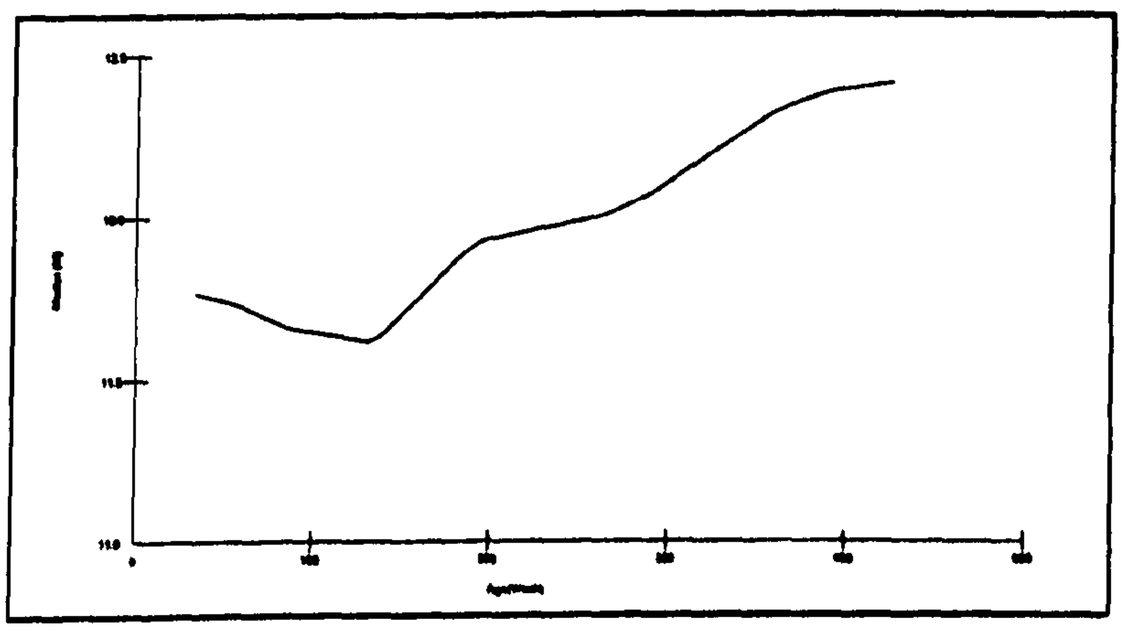
Figure.5.1: Box-Cox power (L) fitted by spline curve with 2, 3 and 4 equivalent degrees of freedom in Children in Focus respectively in figures (a), (b) and (c).



(a)

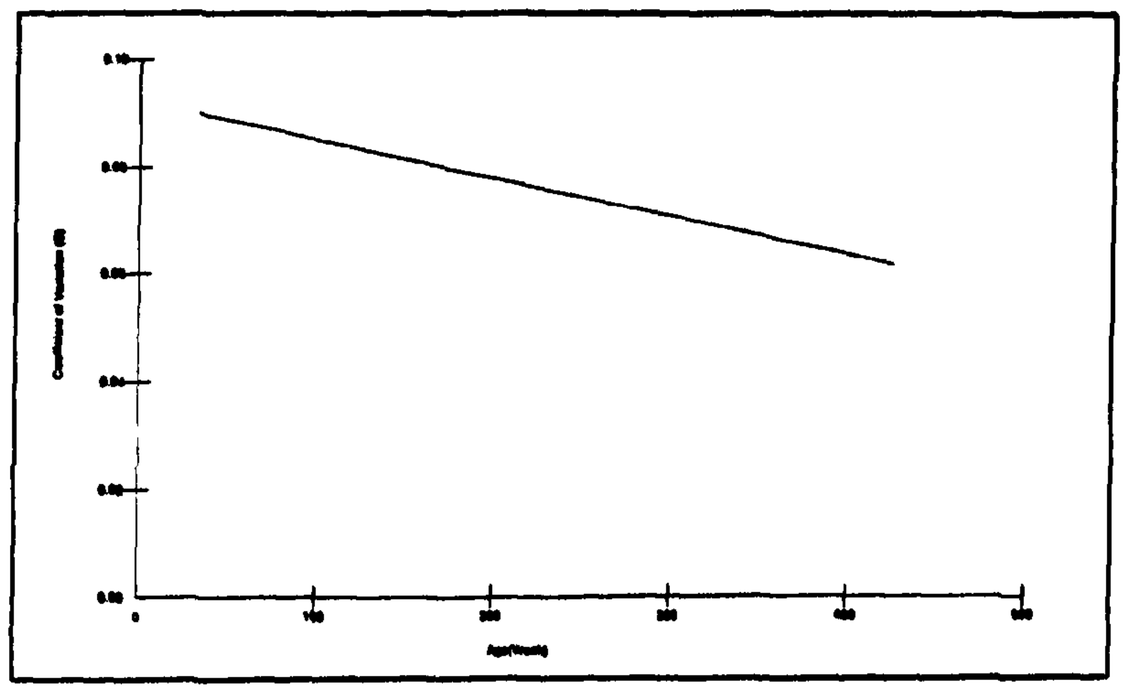


(b)

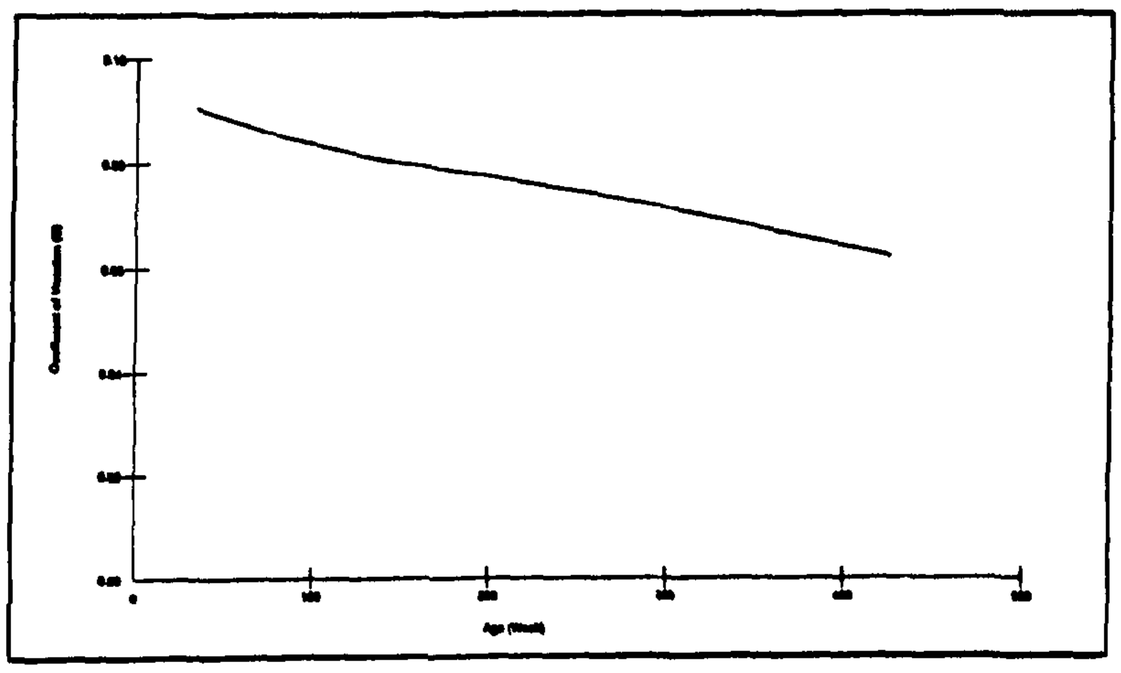


(c)

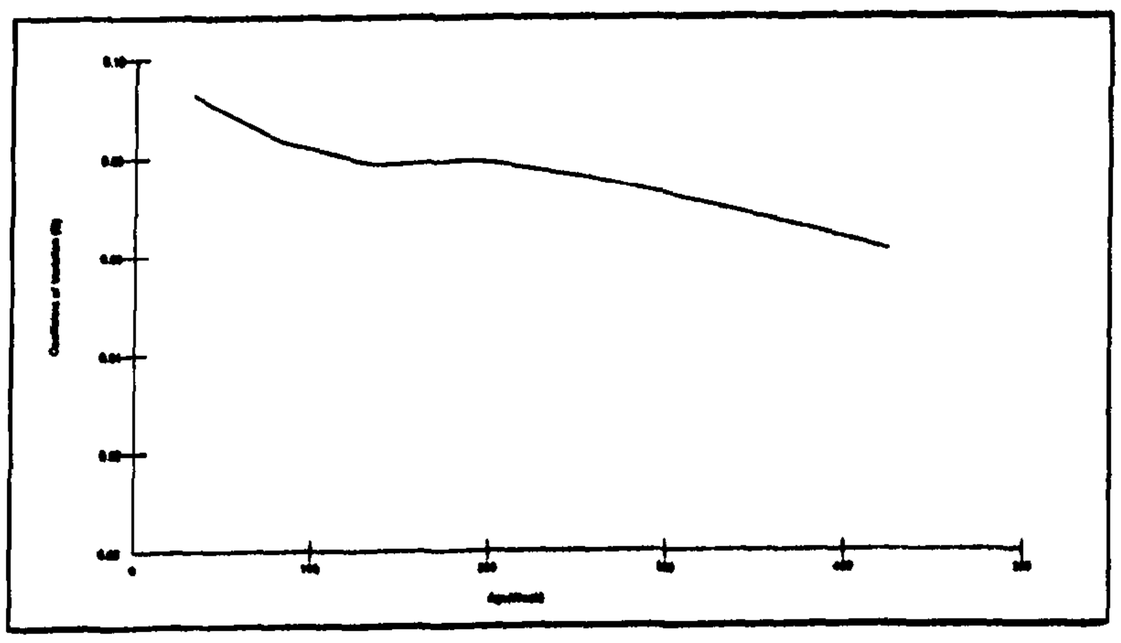
Figure.5.2: Median (M) fitted by a spline curve with 4,5 and 6 equivalent degrees of freedom in Children in Focus respectively in Figures (a), (b) and (c).



(a)



(b)



(c)

Figure.5.3: Coefficient of variation (S) fitted by a spline curve with 2,3 and 4 equivalent degrees of freedom in Children in Focus respectively in Figures (a), (b) and (c).

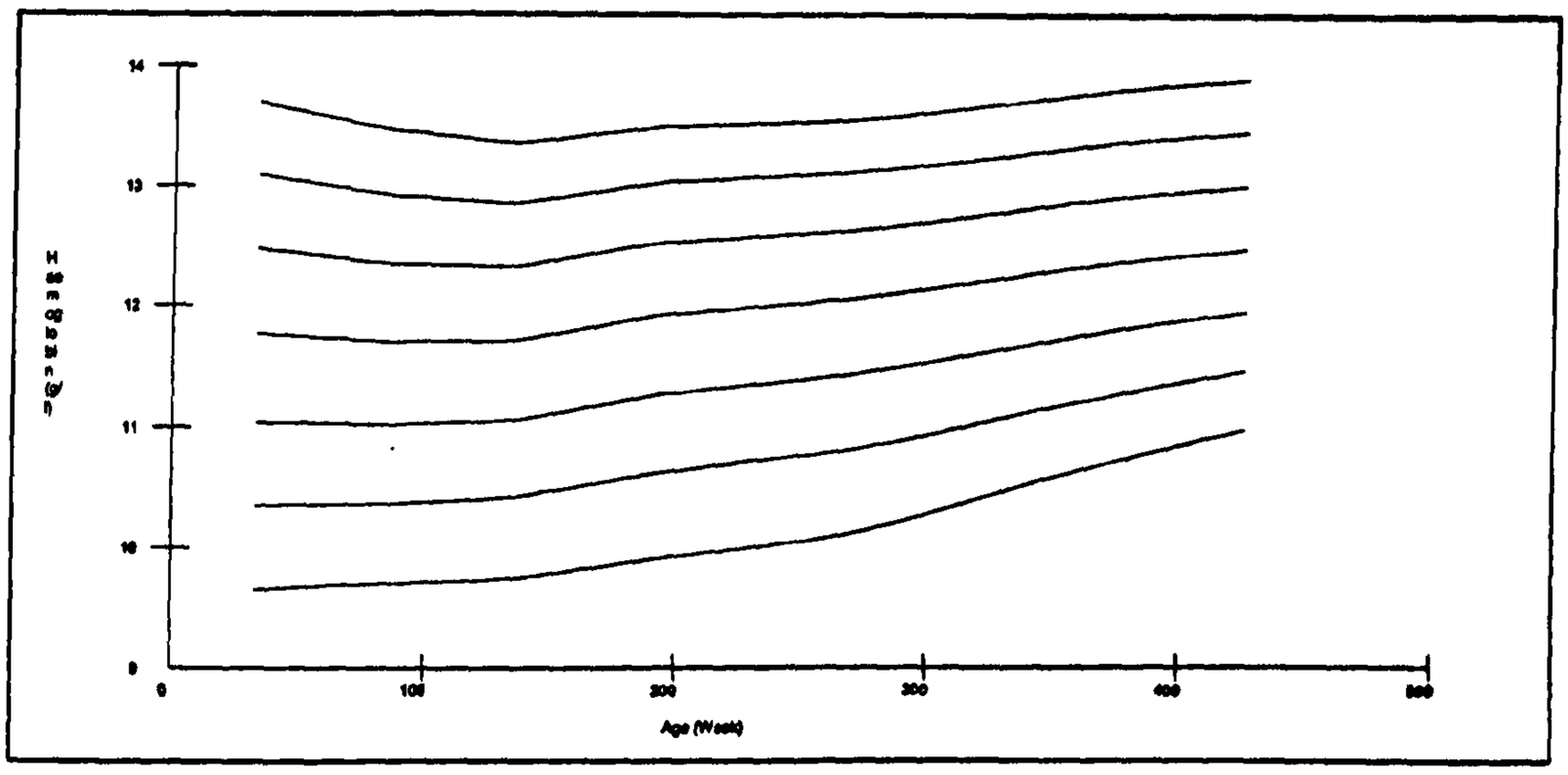


Figure.5.4: Seven reference curves of haemoglobin concentration in Children in Focus from 8 to 84 months of age based on the LMS curves within Figures 1-b, 2-b and 3-b. The 3^{rd}; 10^{th}; 25^{th}; 50^{th}; 75^{th}; 90^{th} and 97^{th} .

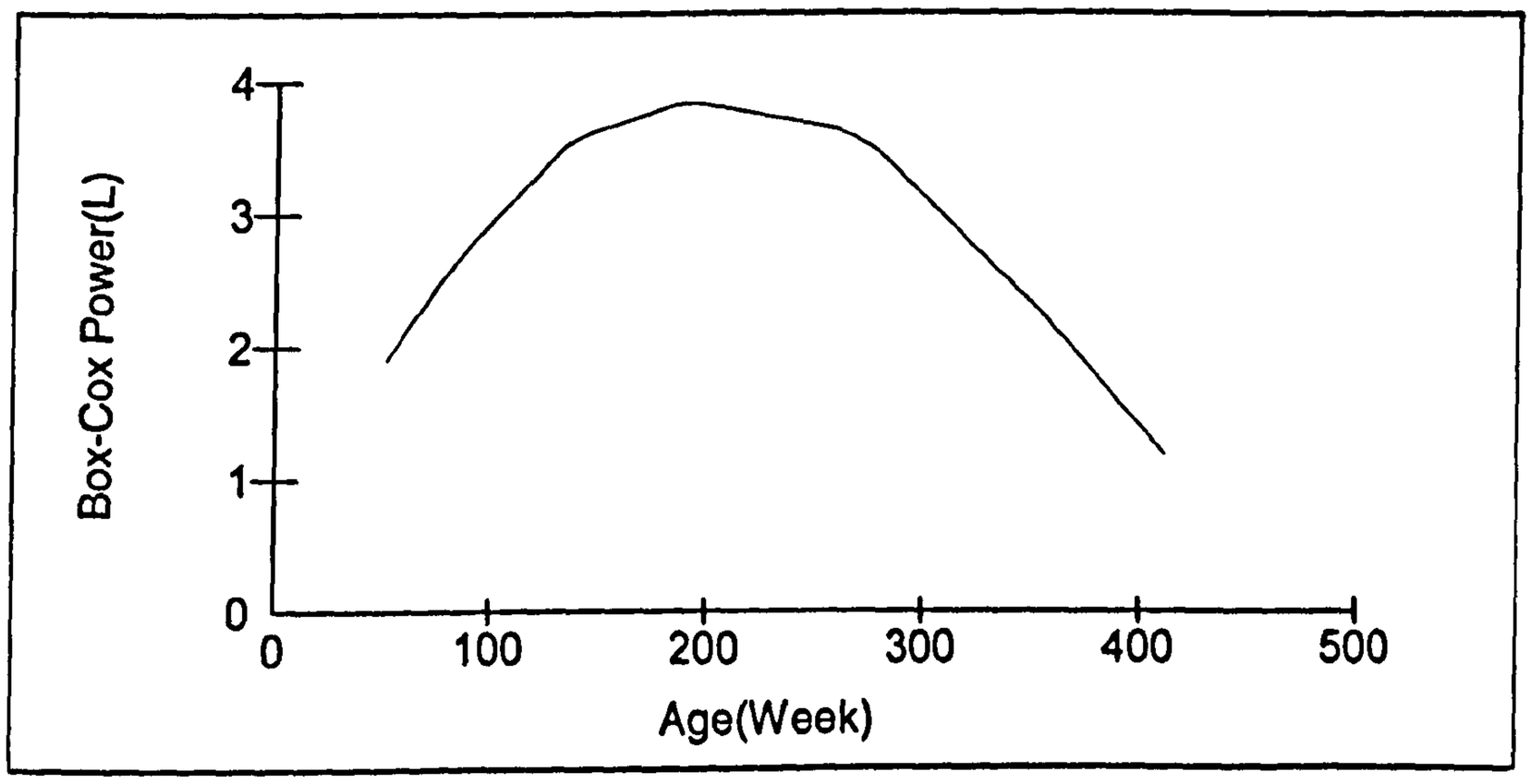


Figure.5.5: Conditional (one previous time) Box-Cox power (L) fitted by a spline curve with 3 equivalent degrees of freedom in Children in Focus.

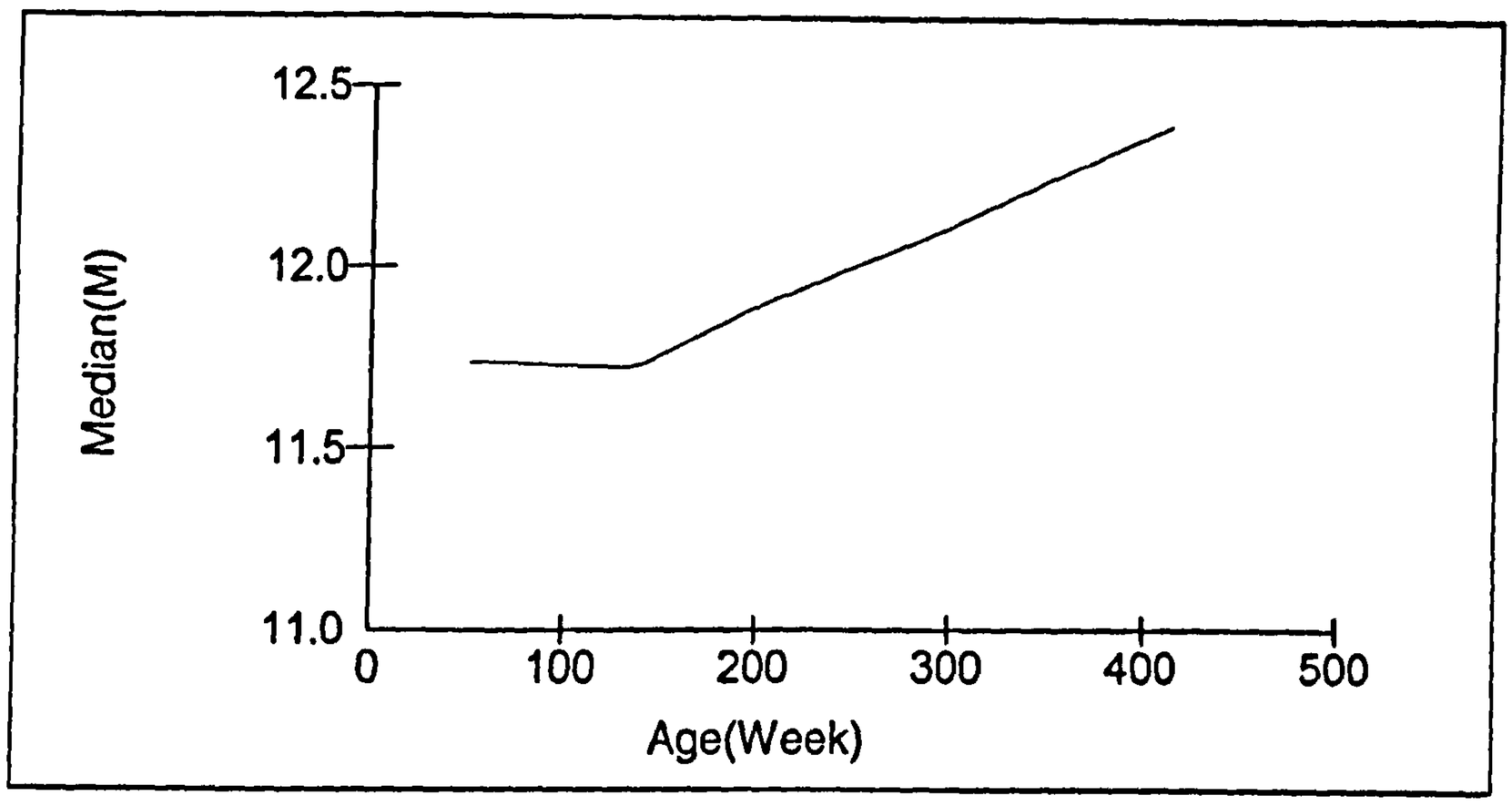


Figure.5.6: Conditional (one previous time) Median (M) fitted by a spline curve with 5 equivalent degrees of freedom in Children in Focus.

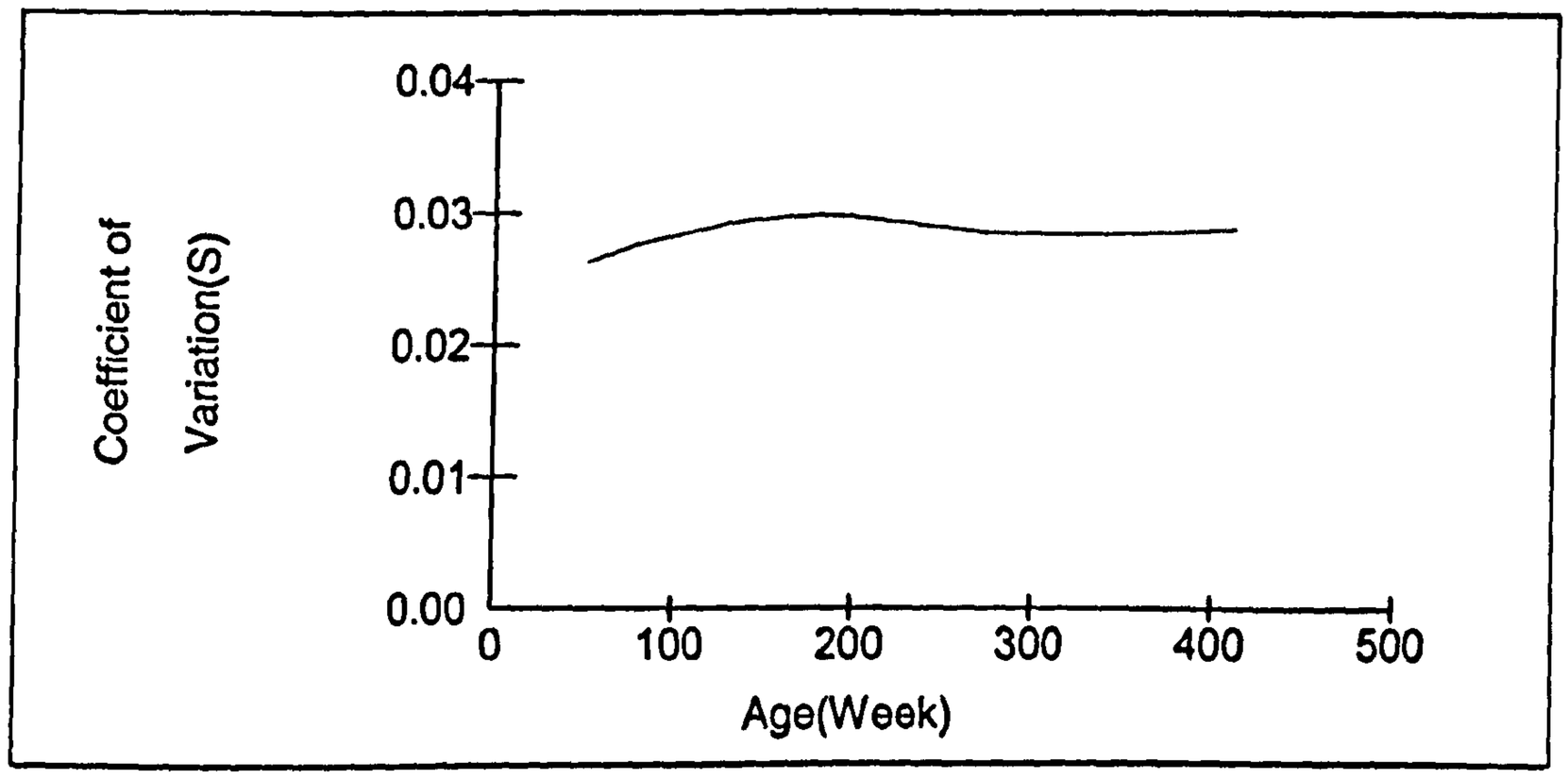


Figure.5.7: Conditional (one previous time) Coefficient of variation (S) fitted by a spline curve with 3 equivalent degrees of freedom in Children in Focus.

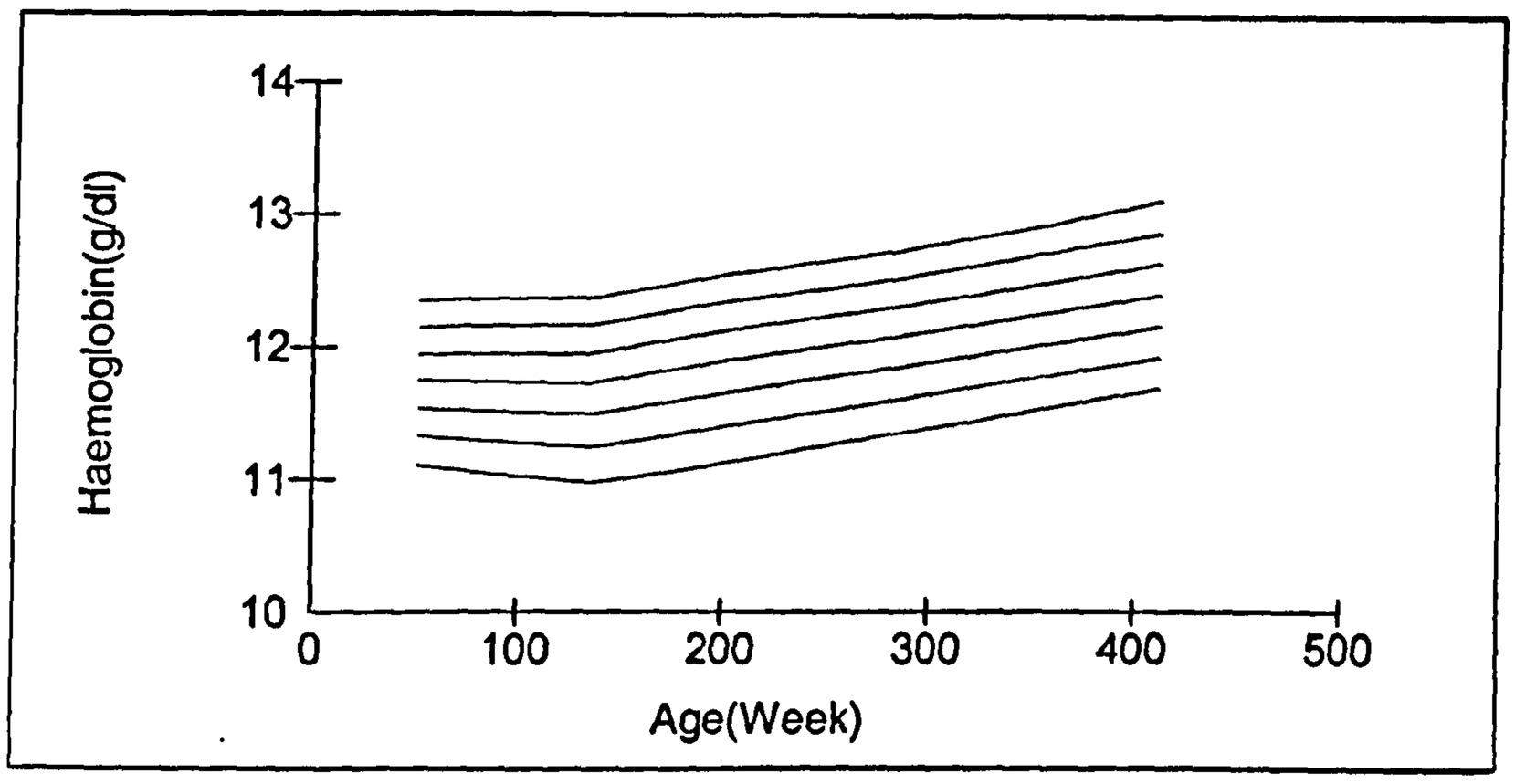


Figure.5.8: Seven Conditional (one previous time) reference curves of haemoglobin concentration from 8 to 84 weeks of age based on the LMS curves within Figures 5.5,5.6 and 5.7. The 3^{rd}; 10^{th}; 25^{th} ; 50^{th} ;75^{th} ;90^{th} and 97^{th} centiles.

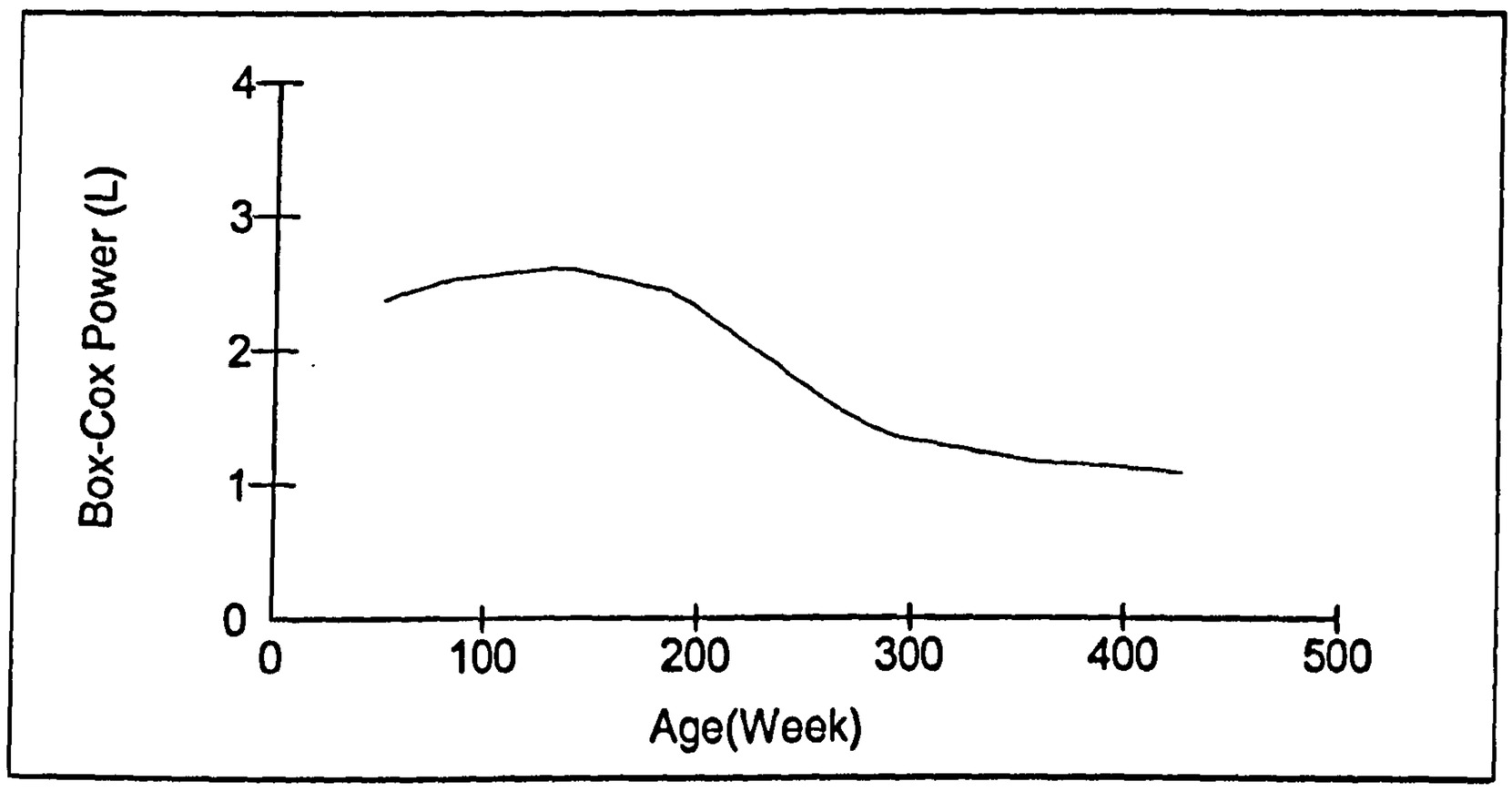


Figure.5.9: Conditional (all previous time) Box-Cox power (L) fitted by a spline curve with 3 equivalent degrees of freedom in Children in Focus.

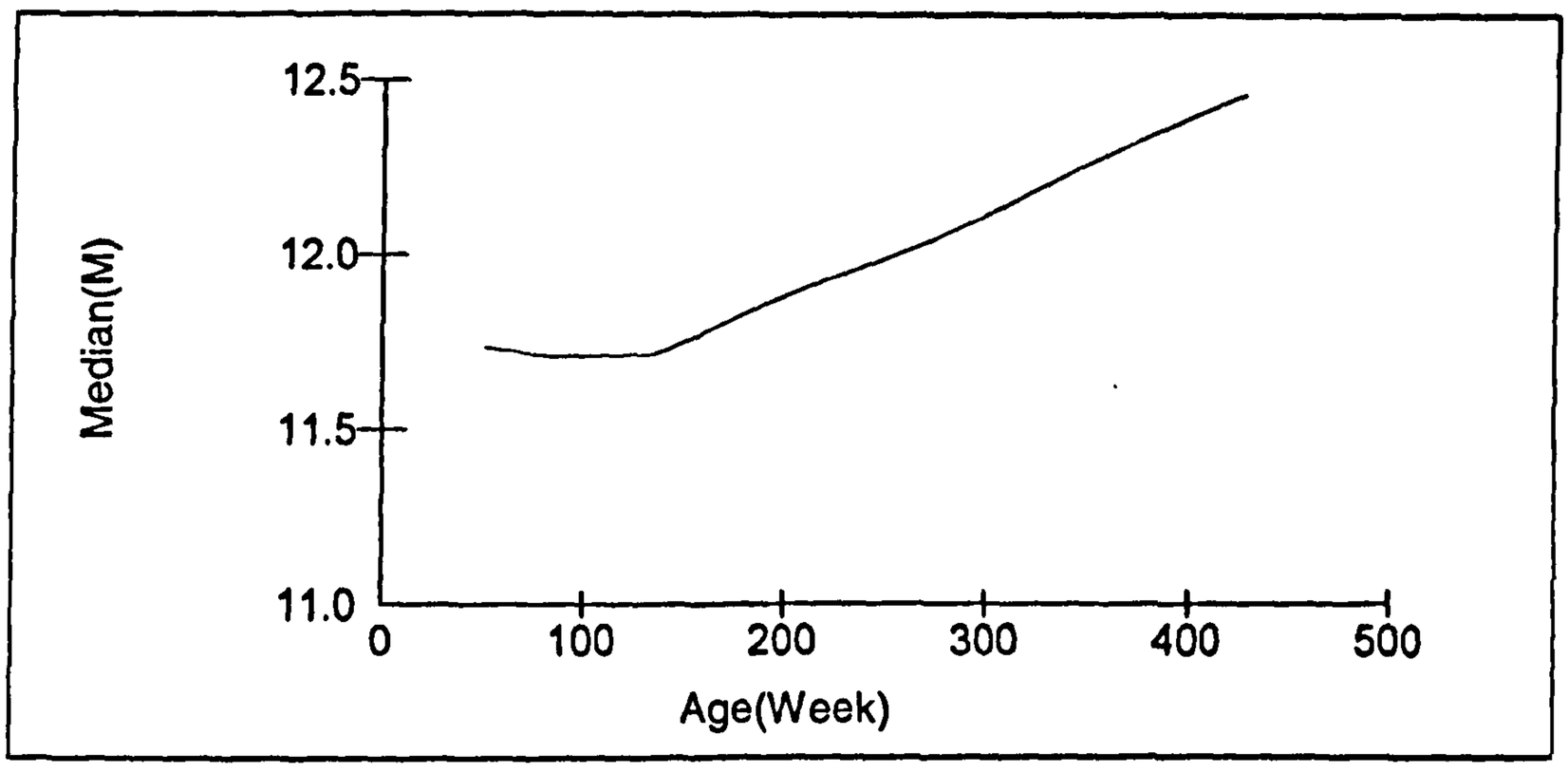


Figure.5.10: Conditional (all previous time) Median (M) fitted by a spline curve with 5 equivalent degrees of freedom in Children in Focus.

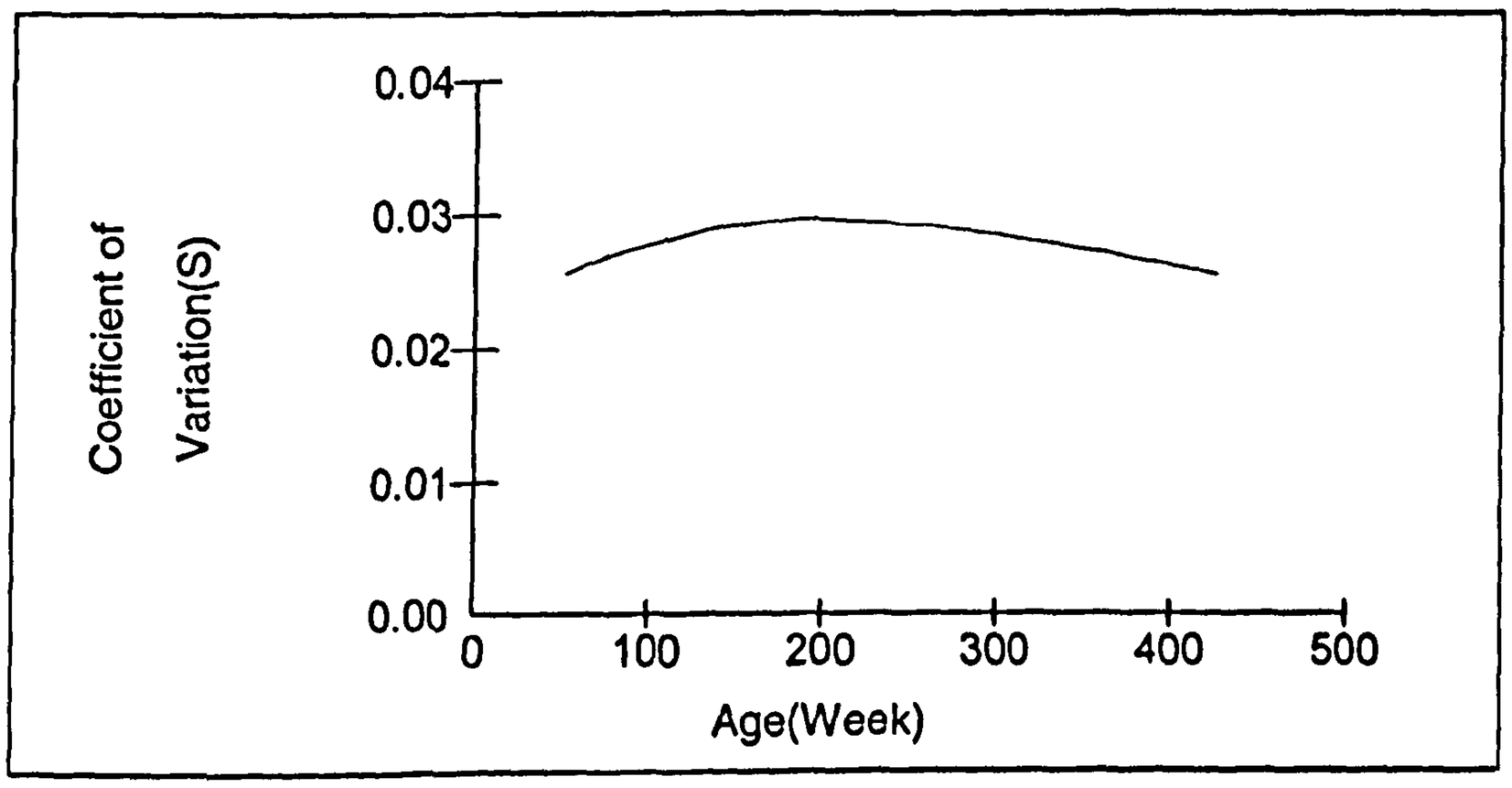


Figure.5.11: Conditional (all previous time) Coefficient of variation (S) fitted by a spline curve with 3 equivalent degrees of freedom in Children in Focus.

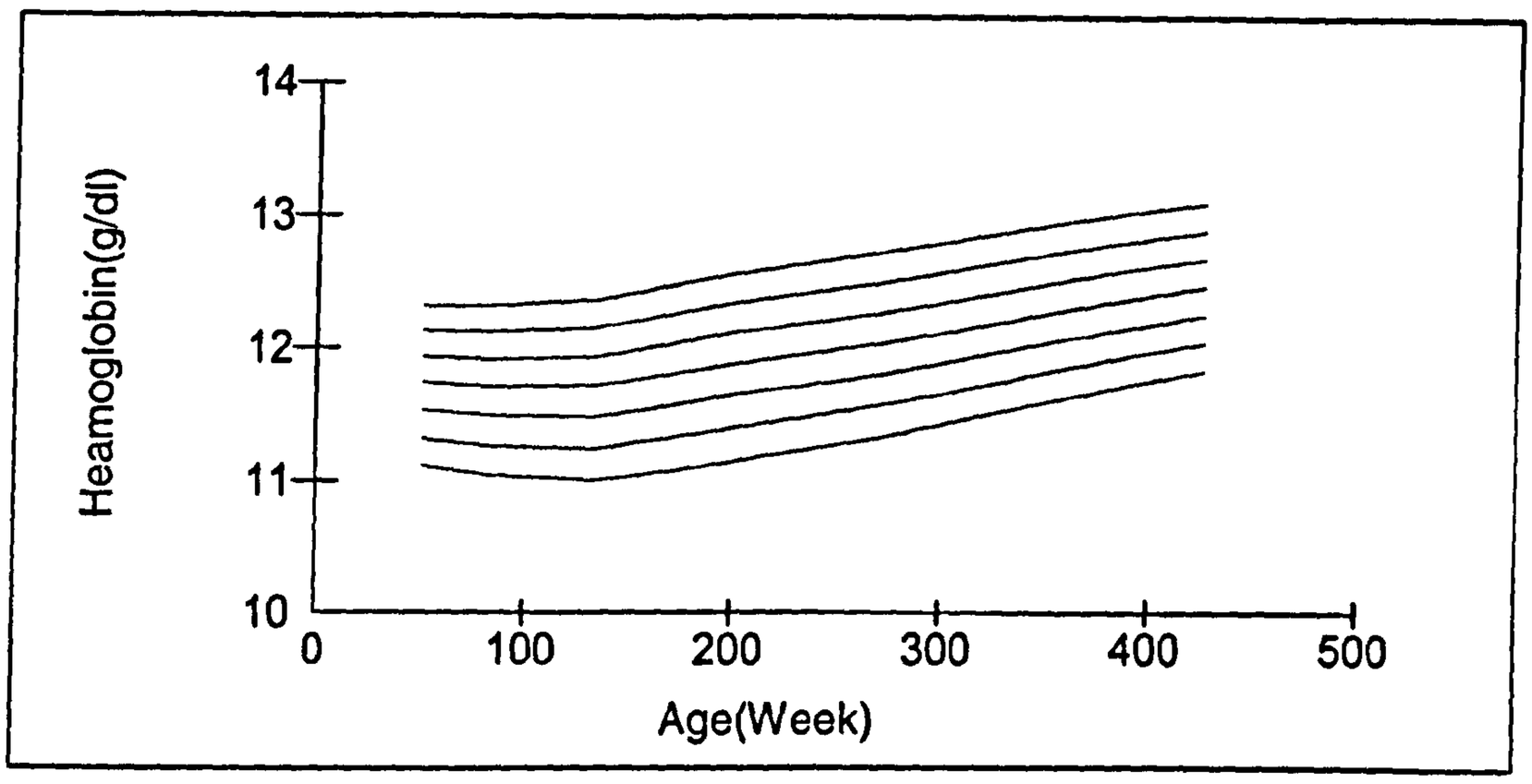


Figure.5.12: Seven conditional reference (all previous times) charts of haemoglobin concentration in Children in Focus from 8 to 84 weeks of age based on the LMS curves within Figures 5.9,5.10 and 5.11. The 3rd; 10th; 25th ; 50th ;75th ;90th and 97th .

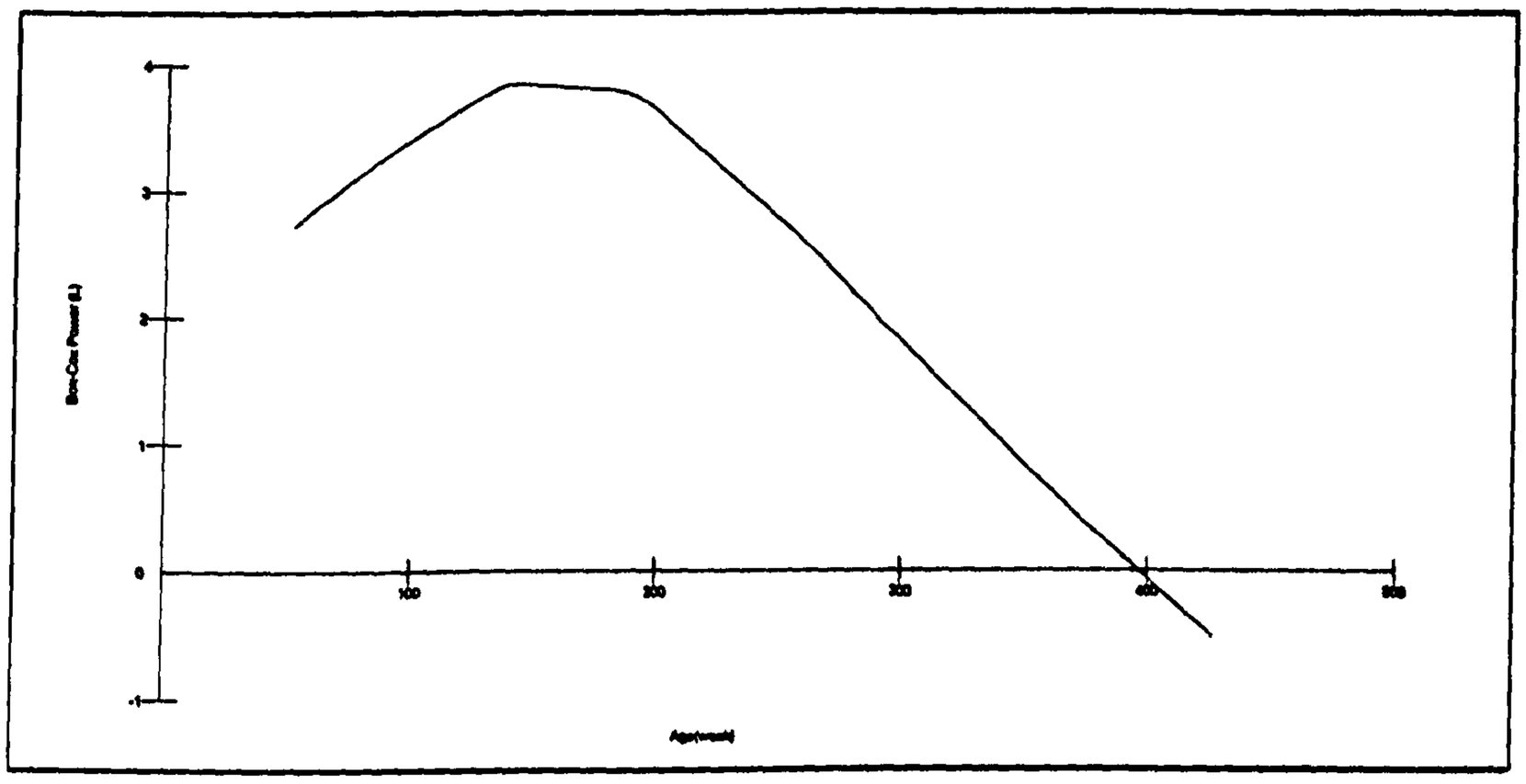


Figure.5.13: Conditional (one previous time) Box-Cox power (L) fitted by a spline curve with 3 equivalent degrees of freedom after used the imputation method in Children in Focus.

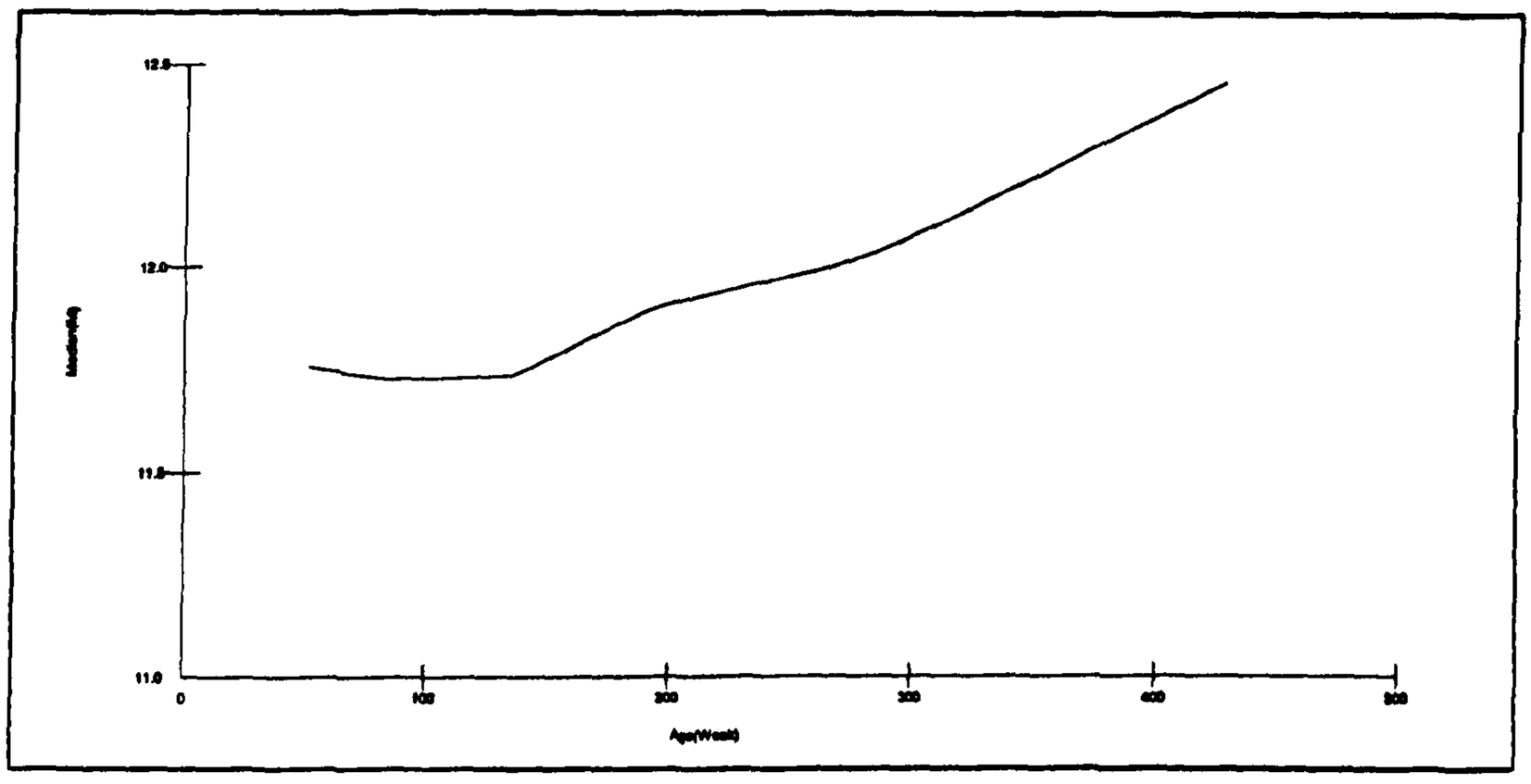


Figure.5.14: Conditional (one previous time) Median (M) fitted by a spline curve with 5 equivalent degrees of freedom after used the imputation method in Children in Focus.

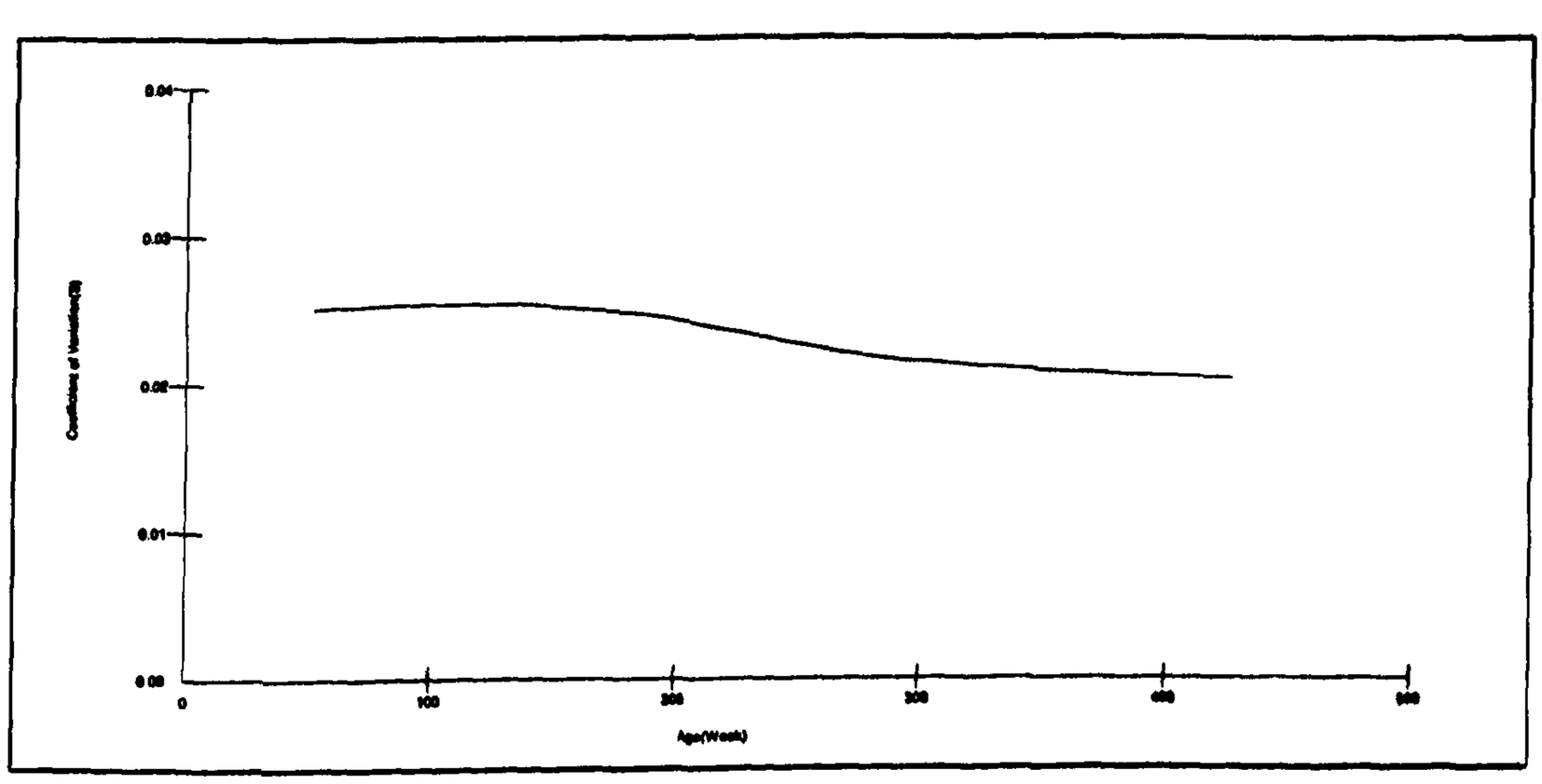


Figure.5.15: Conditional (one previous time) Coefficient of variation (S) fitted with by a spline curve with 3 equivalent degrees of freedom after used the imputation method in Children in Focus.

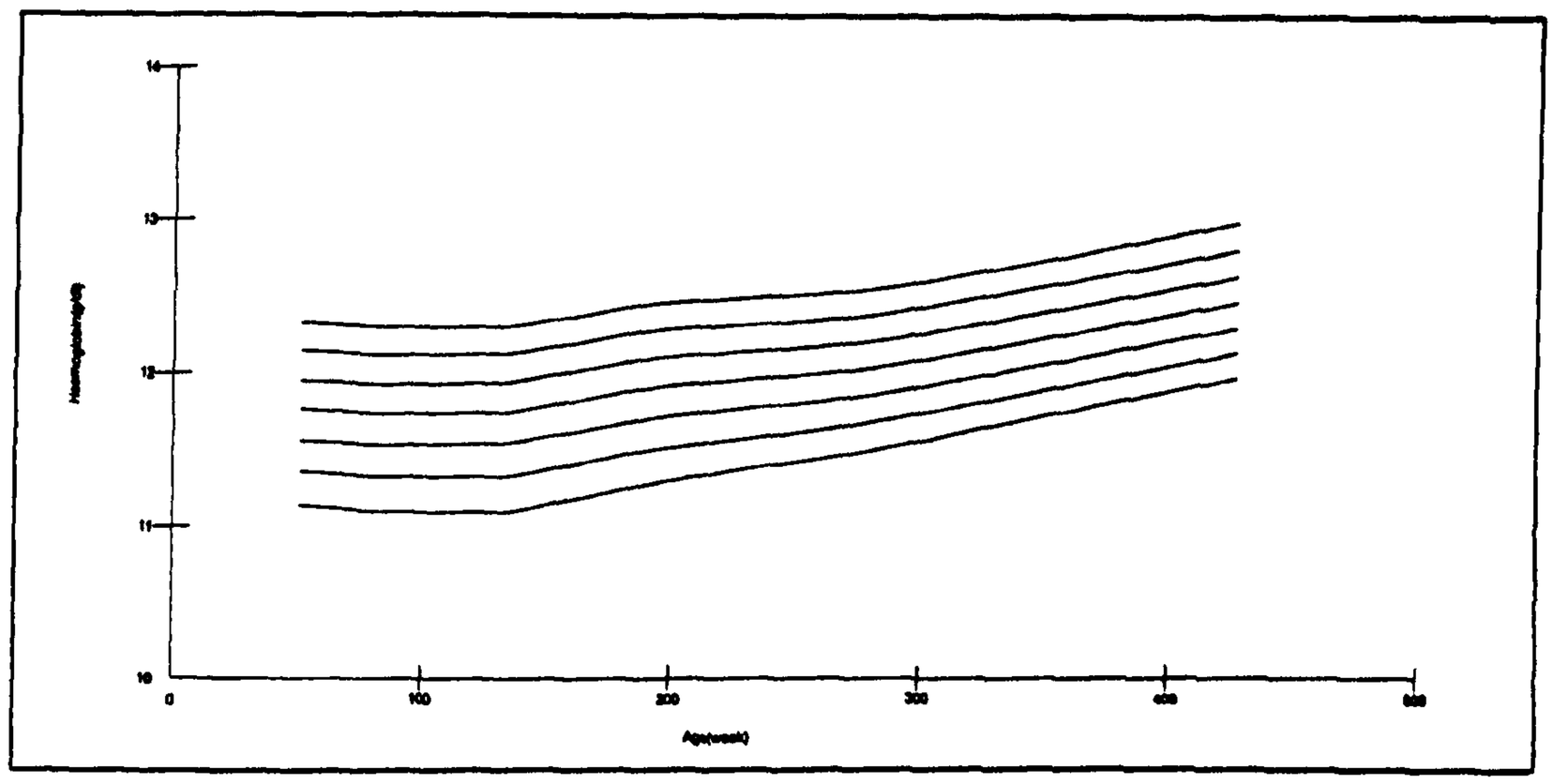


Figure.5.16: Conditional reference (one previous time) curve of haemoglobin concentration after used the imputation method in Children in Focus from 8 to 84 weeks of age based on the LMS curves within Figures 5.13, 5.14 and 5.15.The 3rd; 10th; 25th ; 50th; 75th; 90th and 97th.

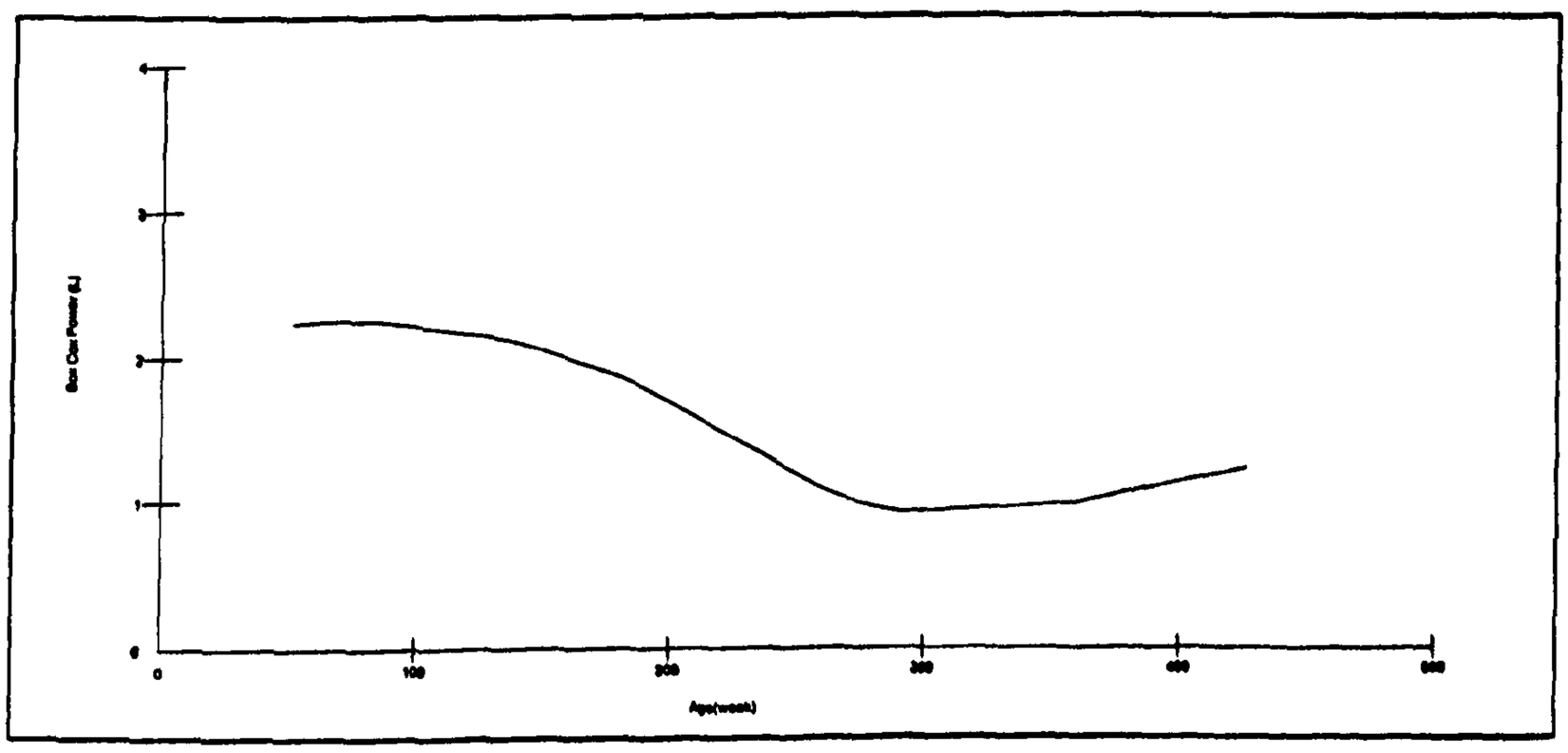


Figure.5.17: Conditional (all previous time) Box-Cox power (L) fitted by a spline curve with 3 equivalent degrees of freedom after used the imputation method in Children in Focus.

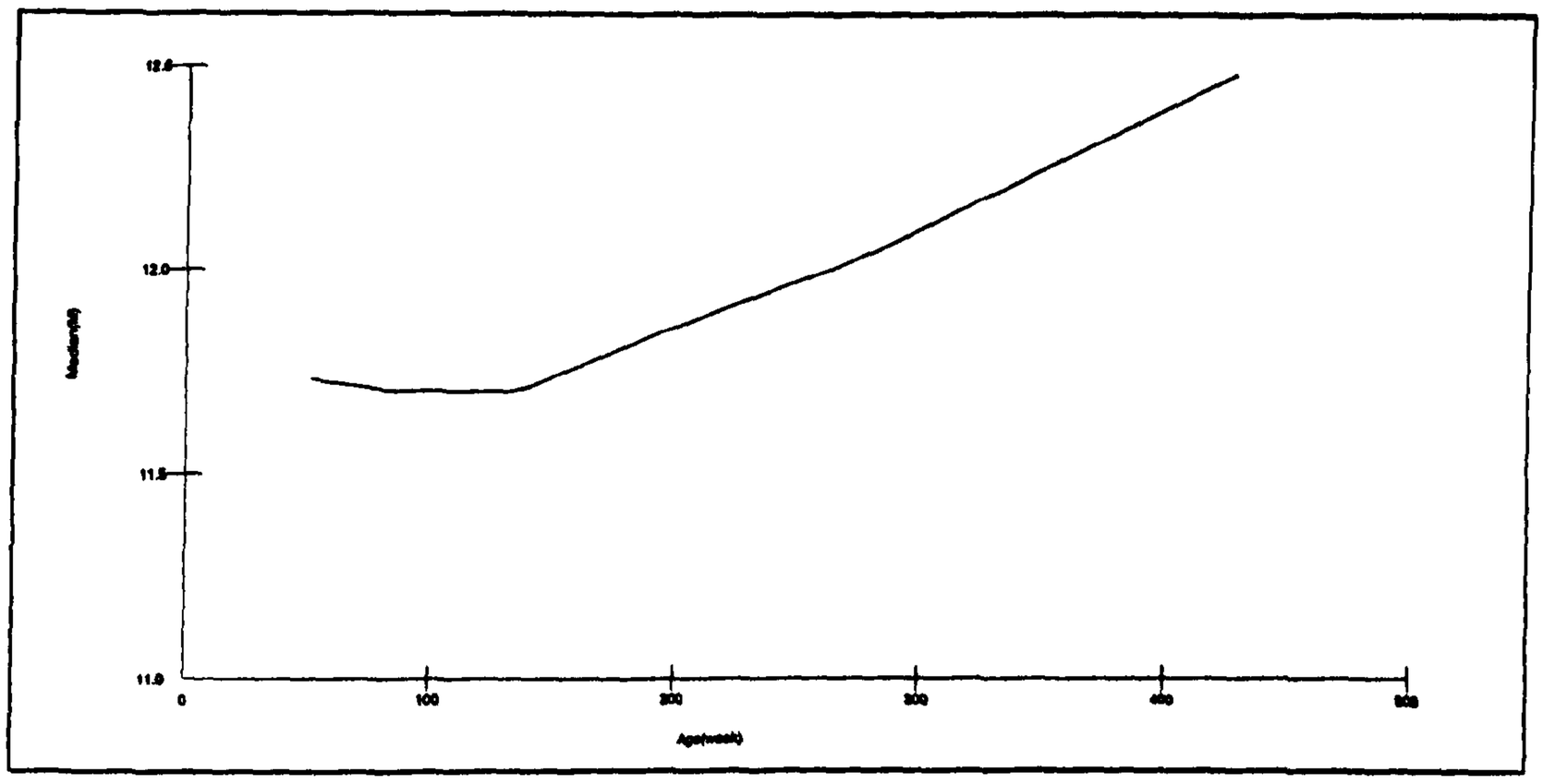


Figure .5.18: Conditional (all previous time) Median (M) fitted by a spline curve with 5 equivalent degrees of freedom after used the imputation method in Children in Focus.

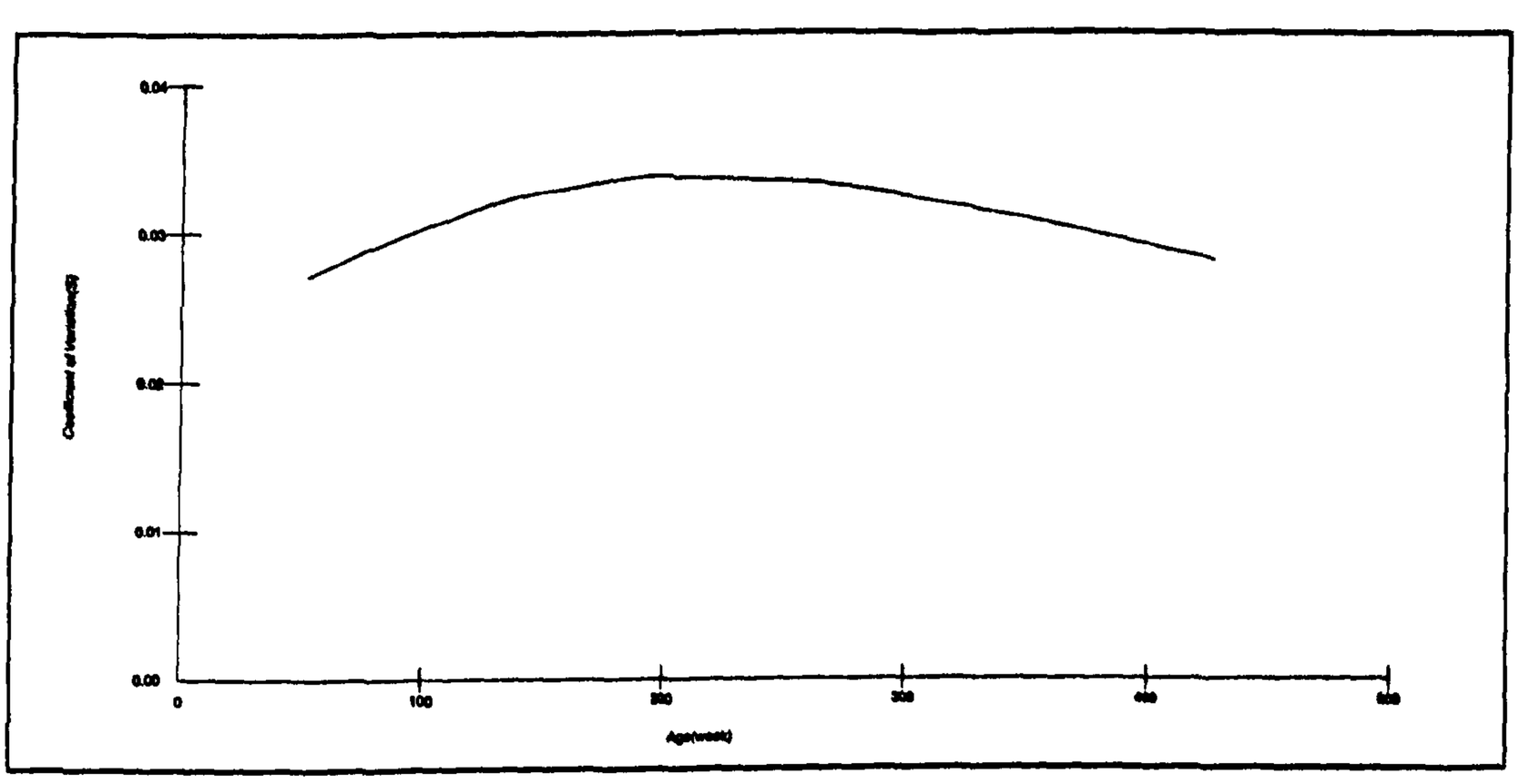


Figure.5.19: Conditional (all previous time) coefficient of variation (S) fitted with by a spline curve with 3 equivalent degrees of freedom after used the imputation method in Children in Focus.

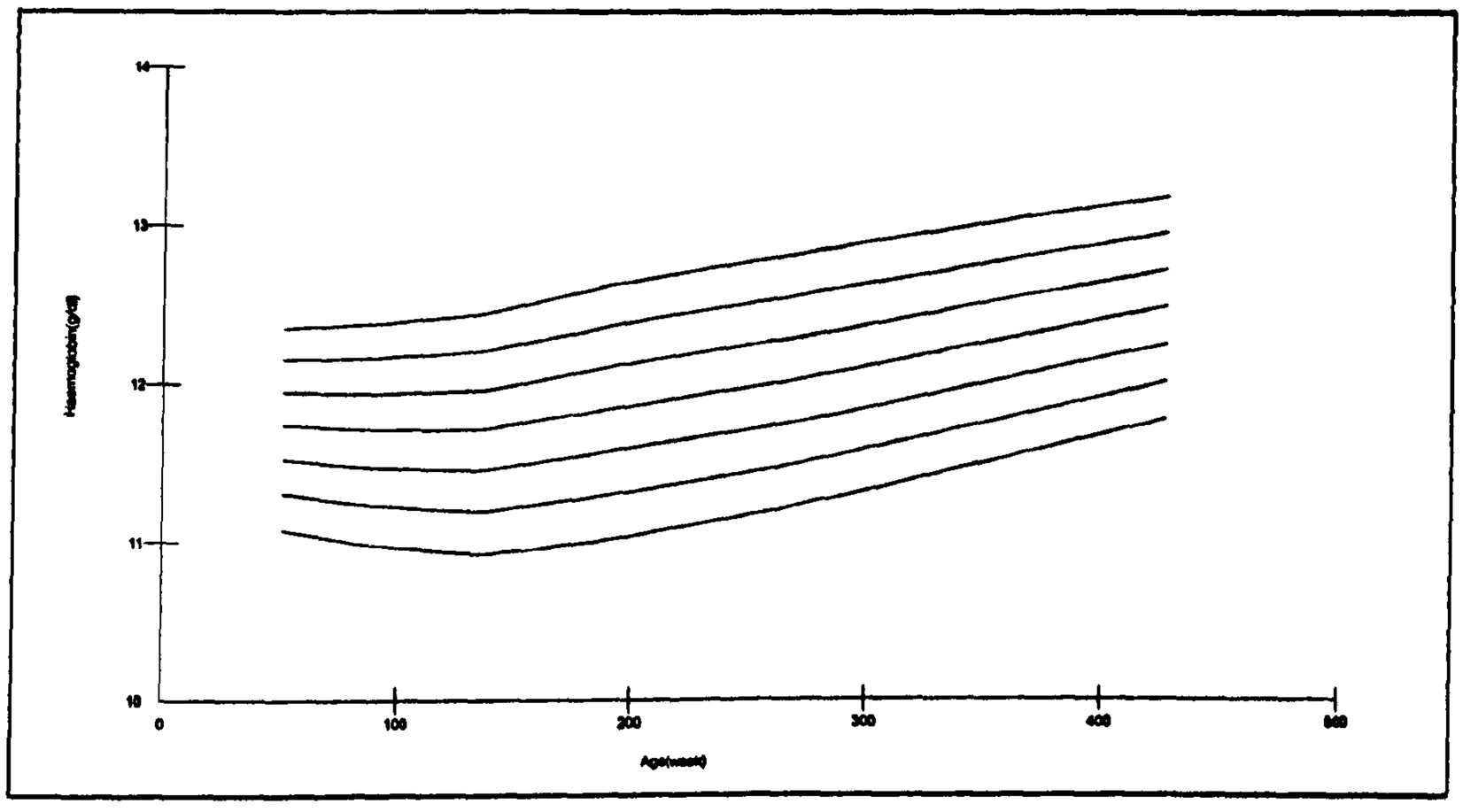


Figure.5.20: Seven conditional reference (all previous time) curves of haemoglobin concentration after used the imputation method in Children in Focus from 8 to 84 weeks of age based on the LMS curves within Figures 5.17,5.18 and 5.19.The 3^{rd}; 10^{th}; 25^{th}; 50^{th} ; 75^{th}; 90^{th} and 97^{th} .

		Z8	Z12	Z18	Z31	Z43	Z61	Z84
Z8	Pearson Correlation	.000	.273(**)	.212(**)	.265(**)	.143(**)	.179(**)	.213(**)
	Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.000
	N	1074	802	713	485	592	521	550
Z12	Pearson Correlation	.273(**)	1.000	.370(**)	.299(**)	.195(**)	.303(**)	.281(**)
	Sig. (2-tailed)	.000	.	.000	.000	.000	.000	.000
	N	802	920	691	445	551	490	489
Z18	Pearson Correlation	.212(**)	.370(**)	1.000	.400(**)	.279(**)	.276(**)	.313(**)
	Sig. (2-tailed)	.000	.000	.	.000	.000	.000	.000
	N	713	691	815	436	521	462	458
Z31	Pearson Correlation	.265(**)	.299(**)	.400(**)	1.000	.346(**)	.371(**)	.452(**)
	Sig. (2-tailed)	.000	.000	.000	.	.000	.000	.000
	N	485	445	436	584	423	367	359
Z43	Pearson Correlation	.143(**)	.195(**)	.279(**)	.346(**)	1.000	.286(**)	.339(**)
	Sig. (2-tailed)	.000	.000	.000	.000	.	.000	.000
	N	592	551	521	423	717	462	435
Z61	Pearson Correlation	.179(**)	.303(**)	.276(**)	.371(**)	.286(**)	1.000	.434(**)
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.	.000
	N	521	490	462	367	462	641	428
Z84	Pearson Correlation	.213(**)	.281(**)	.313(**)	.452(**)	.339(**)	.434(**)	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.
	N	550	489	458	359	435	428	677

** Correlation is significant at the 0.01 level (2-tailed).

Table.5.1: Correlation matrix for haemoglobin levels SD score in Children in Focus at 7 age groups from 8 month to 84 months.

Intercepts	Est.	SE	T-Value	P-Value
a_{12}	0.106	0.033	3.263	0.001
a_{43}	0.088	0.041	2.133	0.034
a_{84}	-0.243	0.071	-3.436	0.001

Table.5.2: The significant intercepts in multiple regressions.

CHAPTER 6

Discussion and Conclusion

The conclusions of the analyses done in previous chapters are drawn together in this Chapter. This leads into a discussion of the methods used, to fit models to mothers' haemoglobin level during pregnancy, children's haemoglobin levels, and reference curves for haemoglobin levels in childhood.

6.1. Haemoglobin Levels during pregnancy:

Iron deficiency is one of the most important nutrient deficiencies in the world. Iron deficiency anaemia is prevalent in women during pregnancy in every country though its prevalence changes amongst countries.

Generally, Iron deficiency is low in the first trimester increasing in the second trimester. About half of Iron deficiency anaemia happens after the 25th week of pregnancy (Guidozzi, Patel, & McPhail 1995;Milman, Agger, & Niesn 1994).

In the developed countries, estimates indicate that more than 30% of women during pregnancy will have depleted Iron stores by delivery time, and in some populations such as adolescents depleted Iron stores might happen in over 80% of the population (Morbidity and Mortality Weekly Report 1998).

These estimates are even higher in the developing world. For example, around half of pregnant women in Africa, 39% of pregnant women in Latin America, 80% of pregnant women in Southeast Asia, over 60% of pregnant women in the eastern Mediterranean, and 40% of pregnant women in the West Pacific are anaemic (WHO 1997).

The prevalence of anaemia data through out the world appears to indicate that Iron intake is inadequate in most diets. Thus, a significant proportion of pregnant women are unable to reach their required daily Iron levels (WHO 1992) .

In order to prevent the decline in haemoglobin concentration and the reduction in Iron stores correlated with pregnancy, interventions are designed (Beard 2000).

Once Iron deficiency anaemia is diagnosed, Iron supplements are recommended. Iron supplements might also be recommended for women at risk of Iron deficiency anaemia. To discover women at risk, screening for risk factors such as low total energy intake, or low dietary intake of meat or vitamin C are necessary.

The specific reason for the low hemoglobin levels found in mothers during pregnancy is still unknown in the vast majority of studies (Scanlon et al. 2000) .

Our model is summarized by the present study in order to effectively address repeated haemoglobin measures data obtained during pregnancy. The models provide a useful approach to study the role of some factors in haemoglobin levels. Due to the important role of Iron supplement in variation of haemoglobin concentrations during pregnancy, we have calculated curves for haemoglobin during pregnancy in two Iron groups and a non-Iron group of mothers separately.

Milman (Milman et al. 1999;Milman , Byg, & Agger 2000) has calculated a curve for haemoglobin levels in pregnant women with and without Iron supplement.

The Milman's haemoglobin curve with Iron supplement has a different shape compared with the curve obtained in our study.

Compared with this study curve, Milman reported lower haemoglobin levels in early pregnancy and higher levels in late pregnancy in non-iron treated. Also there is a progressive decrease in the haemoglobin levels, with the lowest value found in the third trimester in this group similar to mothers who never took Iron supplement in our study, Group 1 (Milman , Bergholt, Byg, Eriksen, & Graudal 1999).

It is impossible to compare Iron supplement group result between these studies, as Iron supplementation in Milman's study took place quite early on in pregnancy. While in our study, Iron supplementation took place at later stages.

It needs to be noted that in our study, genuine interpretation of results in relation to Groups 2 and 3, was hindered by the lack of more specific information about the exact week in which use of Iron supplements began. Notably, the mean haemoglobin levels on Milman's curves (Milman , Byg, & Agger 2000) are similar to our curve for the mothers who never took Iron supplements, and slightly higher than our curves for the Iron supplement groups.

Koller analysed haemoglobin levels from the beginning of pregnancy for mothers who took 100-200 mg of Iron (Koller et al. 1979) .

The various haemoglobin levels were modelled by a polynomial of second degree in Koller's study. As we already mentioned, in our study the polynomial models with quadratic, cubic or higher order terms were poor fits to the data.

The log-likelihood test is also used to test the difference between models that fitted by using an ANOVA procedure. But results illustrated that the biggest value was log-likelihood of cubic polynomial over quadratic polynomial and others.

However in the present study, the best model for the data was the cubic spline model, as can be seen in Tables 3.5 to 3.7.

Milman reports (Milman , Agger, & Nielsen 1991) that around 12% of the placebo treated mother during pregnancy had evidence of Iron deficiency anaemia (haemoglobin, 11g/dl).

Figure 3.5 reports haemoglobin levels for three distinct groups, firstly mothers who never took Iron supplement (Group 1), secondly mothers who took Iron supplement between early pregnancy and 18 week of pregnancy (Group 2) and thirdly, mothers who took Iron supplement between 18 weeks and 32 weeks (Group 3).

As shown in Figure 3.5, the lowest level of haemoglobin can be found in the third trimester for mothers without Iron supplement and in the second trimester for mothers with Iron supplement, but haemoglobin levels were higher in mothers in-group 2 compared with mothers in-group 3, at the same point during the second trimester. Therefore, taking iron supplement from early during pregnancy is suggested.

Consequently, pregnancy is a time in which the risk for developing Iron deficiency anaemia is high.

However, a modest drop in haemoglobin levels in pregnancy seems to be a normal physiological event; haemoglobin concentration reaches a low point in last part of the second trimester of pregnancy and then rises again slightly in the last part of third trimester in groups 2 and 3.

Worldwide, the reason for anaemia in at least half of the cases amongst pregnant women is due to nutritional Iron deficiency. It is unclear whether mothers during pregnancy can normally reach a normal Iron status without the need for Iron supplement. Some scientists agree that it is impossible to keep Iron levels at an acceptable level during pregnancy, by merely following a normal diet and therefore Iron supplement essential (Guidozzi, Patel, & McPhail 1995).

In ALSPAC study, women were asked about their use of Iron supplement in early pregnancy (started by 18 week into pregnancy) and in late pregnancy (18-32 weeks) but the exact time of starting to take supplements was not recorded.
We found that the mean hemoglobin concentration was lower, in Iron-group mothers during the 1st and 2nd trimester compared with non-Iron group.
Our data confirmed that Iron supplement should be administered from early in pregnancy, However we cannot give any suggestions in which week is the best time to start taking Iron supplement.
For final model in this study, initially all significant covariates in main effect models were included in the analysis in final model separately in each iron supplement group. However the final model only included covariates that significantly contributed to the model and so other covariates were excluded from final model.
The covariates that were most commonly found to affect haemoglobin levels during pregnancy in the final model were parity and vegetarianism of mothers.
Maternal education, singleton or multiple pregnancy, were included in the final model for just in the remaining groups, no effect was found of these covariates perhaps due to inadequate statistical power, but in general because of the small sample size, making it difficult to detect an effect.

6.2. Haemoglobin levels during childhood:

This report summarizes our model in order to efficiently address repeated haemoglobin measures data obtained in early life. The model provides a useful approach for studying the role of nutrition, and other factors affecting haemoglobin levels.
The main finding of this study, was that early haemoglobin levels could explain a major part of the variation in haemoglobin levels between children, by growth velocity, which had a positive influence and by different nutrition intake, which can have either a positive or a negative influence.

However, our finding shows that strong tracking, together with the significant influence of growth velocity and nutrition intakes on haemoglobin levels, supported the view that haemoglobin levels during early life are a useful indicator of Iron stores, and thereby of the risk of developing Iron deficiency later on in childhood.

Risk of Iron deficiency is high in early childhood due to the fact that on the one hand there is a high demand for Iron. At time of fast growth and on the other hand there might be poor level of Iron in the diet (Male et al. 2001).

Haemoglobin level of 10.5g/dl was indicated as the criterion for anaemia (Fuchs et al. 1993;Siimes, Salmenpera, & Perheentupa 1984) . In a study of routine screening of one year old infants in Norway, 37% had haemoglobin levels below 11 g/dl (Klem 1993), and in a study in Sweden where healthy, breastfeeding infants took part, the mean haemoglobin in 6 months old children was 10.9 g/dl (Lonnerdal & Hernell 1994).

Our data indicates that in childhood, the mean haemoglobin levels have a value of more than 11 g/dl.

In the analysis of *Children in Focus* data, the ethnicity of the children had no significant association with haemoglobin level.

The non-white children in this study were not a large sample, making it difficult to interpret the data (Noble, Emmett, & The ALSPAC Study Team 2001).The objective of the current study was to examine the influence of dietary factors, growth and some other covariates such as maternal education and the children's sex, on haemoglobin levels in children.

Educational status of the mother was used as an indicator in this study. We found a significant negative interaction between maternal education with vocational level at knot (,69) and Degree level at knot (228,324) and age of children.

These negative interactions, reported a decrease in mothers' haemoglobin level in Vocational and Degree levels with time at these knots.

Birth weight was found to have a significant negative correlation with haemoglobin concentration in our study, which may reflect more rapid utilization of Iron status due to greater Iron needs for growth in children with higher weight, though birth weight was not a significant covariate in our final model.

The other factor illustrating a significant negative effect on children's haemoglobin levels was parity. That suggests that children of mothers with a parity of one have higher haemoglobin concentration than children of mother with a parity of two or more.

The most important nutrient intakes were Vitamin C, Calcium and NSP. These intakes were factors, which clearly had a positive influence on haemoglobin level of children. The results of the present study illustrate the importance of these nutrition intakes compared to other nutrition intakes.

Other nutrient intakes such as haem Iron intake, fat intake, saturated fat, monounsaturated fat, energy intake and Iron intake did not explain any of the variations in haemoglobin levels, which is probably due to the relatively uniform and almost universal use of these intakes.

Conversely, there was a positive correlation between weight and haemoglobin levels, consistent with previous report(Sherriff, Emond, Hawkins, Golding, & the ALSPAC Children in Focus Study Team 1999).

The positive association between female gender and haemoglobin levels may partially be explained by the faster growth rate observed in boys. The increased risk of anaemia in boys however, cannot be explained by growth and may reflect physiological differences between genders.

In brief, the significant time stationary covariates include maternal education, sex of the child, parity and birth weight and significant time dependent covariates include vitamin C, Calcium, NSP intake and children's weight.

Although, in the analysis, all of these covariates were found statistically associated with haemoglobin levels as main effects, birth weight (time independent covariate) and Calcium and NSP intake (time dependent covariate) were not significant in the Final model.

For this study we obtained information on maternal haemoglobin level in Iron supplementation groups.

According to this information, there is a positive association between random slope in the model, which was fitted to maternal data in Group 1 as a covariate and children's haemoglobin level.

There was also a positive correlation between random intercept in the mothers in Groups 2 and 3 and children's haemoglobin level in 8-month-old. As a result, it was recommended that pregnant women should take Iron supplement.

The small number of subjects in some covariates does not permit the necessary statistical power to detect a significant difference, nor can a conclusion from the studied groups, be applied to a larger population.

6.3. Reference curves for haemoglobin during childhood:

In this study, we have provided reference curves, which describe the changes in haemoglobin levels in children across the first seven years of life. Haemoglobin levels increase in a curvilinear relationship across this age range.

The main purpose of this study is to apply existing methodology to construct longitudinal reference curves for haemoglobin levels in children. However a problem was found in that the reference curves obtained by usual procedures are basically cross-sectional and therefore, not automatically suitable to serve as reference charts for longitudinal curves. In this study, we paid particular attention to this problem by comparing the reference curves for a longitudinal data, with the cross-sectional references.

The most important advantage of longitudinal data is that it may be used to construct growth or different standards for the reference curve in general and, possibly important for clinical practice, for the individual infant and child (conditional reference curves). The conditional reference curve has a clear role in the assessment of individual haemoglobin levels and is generally more appropriate than the unconditional reference curve.

The introduction of conditional reference curves, allow one to adjust the centiles at each measurement point of a case, according to the previous measurement.

Producing both cross-sectional and conditional centile charts in the curves, allows one to monitor at the same time, both size and change in haemoglobin concentration from past measurements. The main question being whether cross-sectional reference curves were at all useful. The answer to this would be that the advantage of the cross-sectional reference curves appears to be that they are easier to tabulate and plot, than conditional reference curves.

In the literature many sources have reported some kind of centiles as the basis for assessing age-specific reference curves development. The majority of these sources used cross-sectional data and others have used longitudinal data to produce reference curves.

The reference curves in this study were constructed for the distributions of haemoglobin levels in children ages 8 months old to 7 years old in a large cohort study *Children in Focus.*

Haemoglobin levels tend to increase as age increases. There are clear differences between cross-sectional and conditional reference curves in haemoglobin concentration.

The cross-sectional reference curves become narrower with an increase in age. While the conditional reference curves for the haemoglobin concentration are narrower than the corresponding centiles in cross-sectional curves, regardless of age.

The large range in haemoglobin levels observed in the cross-sectional reference curves between the 3^{rd} and 97^{th} percentiles compares with conditional reference curves. This large range in haemoglobin levels could suggest that extreme values are due to a large S value.

The S curve for conditional curves are lower than the S curve in cross-sectional, but the amplitude of the difference depends on the kind of conditional methods which the S curves is being compared.

As you can see in Figure 5.12, the usefulness of reference curves is demonstrated by considering all previous time, as the Figures 5.3b, 5.7 and 5.11 show, the S values for this reference curve are the lowest amongst other curves, before the use of an imputation method.

For dealing with missing values, there are many imputation methods. In the present study, the conditional reference curves were reproduced by using the imputation method.

The expectation is that the estimated coefficient of variation (S) after the imputed data usually underestimates the true variation.

Hence the coefficient of variation, which was previously obtained by all previous times, should still be smaller than the coefficient of variation obtained by one previous time.

Whereas, after using the single imputation method, the smaller S value between conditional reference curves is produced by using one previous time point.

This might be related to some disadvantages in the single imputation method, such as depending on the observed values (Perez et al. 2002).

According to the different S values found between the beginning and the end of measurements in the coefficient of variation Figures, it might reasonably be expected that haemoglobin in children in early of life, would show more of a variation than in older children.

The shapes of the conditional reference curves are relatively similar to each other, throughout the whole age range.

Although as expected, the coefficient of variation of haemoglobin levels in reference curves with one previous time, was greater than the reference curves with all previous time.

The LMS method with penalized likelihood is highly flexible and practical. It is often easy to produce credible reference curves even when the data appear to have a complicated shape. The complication of the shape of parameter curves is reflected in the number of *e.d.f s*.

Regarding haemoglobin level during childhood, the aim of this study is to draw attention early on, to possible problems such as anaemia and specifically Iron deficiency anaemia, which are common in most parts of the world, especially so in the developing countries. The easy monitoring of variables such as the measurement of haemoglobin levels and ferritin levels in infants and children are useful screening tools, which allow us to promptly detect any problems.

By producing appropriate reference curves for our population we were able to identify children at the extremes of haemoglobin concentration.

It is therefore important, to have current haemoglobin reference curves to look at regional differences within the UK, being able to compare them with other countries.

We suggest that these reference curves may be used to assess abnormal haemoglobin concentration in children in disease states.

There are a number of areas where conditional reference curves may be useful, such as monitoring of ferritin levels and blood pressure during pregnancy.

6.4. Methodology:

In this study we have combined some modern techniques in order to produce a flexible model, for representing variations in haemoglobin levels in both mothers during pregnancy and in children.

Longitudinal studies are one of the most commonly selected study designs, in research on human health.

Approaches to the analysis of longitudinal data carry on being developed, and these contributions offer absorbing insights into some of the methods that are proving to be helpful.

Furthermore, they offer useful propositions as to how a data analyst can use these methods in their research. Some of the specific areas of longitudinal data analysis that are discussed here include the parametric methods of analysis such as Generalized Linear Mixed Models and Cubic Splines.

We showed how a General Linear Mixed Model could recover information in the data, which cannot be found by a Fixed Effect Model, when incomplete, and/or unbalanced data is at hand. This is an important benefit of the Mixed Models. The General Mixed Model approach offers a flexible way of modelling correlation in the data.

This is particularly relevant for longitudinal data or other kinds of repeated measures data. Therefore, the appropriate inference about fixed effects is obtained and the covariance structure itself provides further insight into the problems that are in front of us. Also in most cases, a Mixed Model is actually the only suitable model for the given data.

According to the data in this study, fitting a polynomial does not provide a meaningful estimation for the relationship between haemoglobin levels and time. A best fit the model with an exact order was difficult to determine.

Therefore, a meaningful fit should be provided by others methods, which solve the problems of polynomials.

However, in this study we have combined some modern techniques to produce a flexible model for representing haemoglobin variation. The tools we combined were Mixed Effects Models for longitudinal data, Cubic Splines for flexible function fitting.

When Cubic Splines are used, the polynomial will be better by forcing them through the data points and using a polynomial for a short segment to reduce untoward excursions.

Cubic spline is one of great flexibility for data analysis. They are also particularly useful as graphical tools for communicating complicated finding in a normal way.

The mean position curve for a large population may have a shape that is sufficiently smooth to be fitted by a simple parametric function, but individual curves at subjects level show a huge variation in shape.

Thus, successful modelling of particular curves is likely to need a model such as spline models. Splines have the advantage of fitting into the Mixed Model framework, allowing unbalanced data and environmental effects from a normal part of the model to be dealt with easily.

A specific advantage of the spline approach is the additional flexibility regarding the shape of the fitted curves. Whereas, it is not possible to produce such a reasonable fit, even when using high degree polynomials. As an example, a cubic curve has two turning points and one point of inflection, but the inflection point is inevitably halfway between the turning points. When the knots' number has been determined, prior knowledge of the curves can be used to guide knots' location.

Finally, an extremely potentially useful method is presented in this study, in which the method was used to link General Linear Mixed Model and cubic spline.

Fitted a function to longitudinal data in Chapters 3 and 4 using the Linear Mixed Model is suitable. But any meaningful relation between haemoglobin levels and time were not given by either a quadratic, cubic or higher order polynomial according to log-likelihood and AIC criteria.

It would be difficult to provide an explicit polynomial function because the data are sparse in some parts, differences in time lags, incomplete and unbalanced structure in mothers'haemoglobin levels during prognancy and their children.

Hence we focus in finding an appropriate model , concluding that the cubic spline model , which applies a piece-wise polynomial fitting technique is better than the polynomial fitting procedure, because it has more parameters than a polynomial with the same order.

This was confirmed by the values of log-likelihood and AIC. Measurements of haemoglobin concentrations in children are one of the information required when evaluating children's health. They are also useful, for identifying children who are at risk of developing some problem such as anaemia and Iron deficiency anaemia. As reference curves are used in medical practice to assess abnormality, this study provides reference curves for haemoglobin levels during childhood in this population. Most previous reference curves estimation have assumed that the data is independent even if it is apparently correlated, but this is not appropriate for most longitudinal data.

The conditional curves' estimation is based on longitudinal data, which can have a correlation between measurements at different times.

The present study is to consider two methodologies, for constructing cross-sectional and conditional reference curves for haemoglobin levels during childhood.

In the conditional reference curve approach, we define reference curves by using one previous and all previous time points as well as by specifying a correlation structure for haemoglobin levels in various ages by using stepwise linear regressions.

Therefore, the reference curves illustrated in this study allow longitudinal assessment of haemoglobin levels in children, and cautious monitoring of variation in haemoglobin levels, early on should draw attention to any possible problems.

Amongst the methods that have been used for reference curves, the LMS method with based on spline (Cole & Green 1992) has good flexibility. The LMS statistical function is applied to the smoothed curves generated for easiness of interpolation between percentiles, a normal transformation of the curves is useful. A normal transformation makes it possible to estimate any percentile and allows the calculation of standard deviation and z-scores.

The LMS method (Cole & Green 1992) has proved to be a flexible way of modelling the distribution of children's haemoglobin at different ages and can present a valid index for haemoglobin level during childhood.

For each set of percentile curves, the initial smoothing methods were applied to the seven percentiles (3^{rd}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} and 97^{th}) for each age group.

Reference curves for haemoglobin levels during childhood have not been previously shown in any great detail.

We propose that the reference curves of haemoglobin illustrated in Chapter 5 are a useful way of expressing the possible problems during childhood such as anaemia, and that conditional reference curves are more appropriate than cross-sectional for longitudinal data in normal children.

The methodology that has been developed for producing a conditional reference curve allows the use of both, one previous time in children's haemoglobin as well as measurements in all previous time in an individual.

The present study is in addition, a comparison between cross-sectional and conditional reference curves. Providing some evidence of a coefficient of variation difference, in these reference curves.

The results show that the conditional reference curves approach with one and all previous time is more efficient than the cross-sectional reference curves.

For creating an appropriate conditional reference curve for children, we would need to compare conditional reference curves and their coefficient of variations.

Our results show, that conditional reference curves produced by one previous time and all previous times, are very close together in almost every aspect and there is no clear reason to use one over another. Therefore, the use of more complex methods (all previous times) is not recommended.

6.5. Future work:

The following list shows some possible ways in which the work of this thesis may be extended:

- Further research with an adequate number of subjects is needed to illuminate the effect of maternal nutrition intake on the mothers during pregnancy and children's subsequent health to avoid the risk of them developing anaemia.

- Further studies examining ways to ensure adequate nutrition intake or vitamin supplement along with Iron should be carried out to try to improve the Iron status of pregnant women and children.

- Further studies with more focus to mothers' haemoglobin concentration and their children and on some covariates such as singleton and twin are needed to fully elucidate the effects of them on haemoglobin levels.

- Investigating further the association between haemoglobin levels and development cognitive, motor development and school achievement in children.

- A limitation of this study relates to the lack of information regarding the time of taking iron supplementation during pregnancy and regarding nutrition intake for children states. Collecting such data during a long-term follow-up is difficult, but additional studies are needed to confirm haemoglobin levels trends to associated with more and adequate information about these covariates.

- Further research is needed to investigate whether varying Iron status in older children is due to aging or some other factors.

- Increased knowledge and increased information about factors may be important tools in the prevention of Iron deficiency in mother during pregnancy and in children.

- A limited sample size is a problem in creating appropriate conditional reference curves.

- With reference curves it may be possible to identify patterns of childhood haemoglobin levels that lead to adult haemoglobin concentration and risk factors for some diseases related to anaemia.

- Recommend that different reference curves are constructed for haemoglobin levels for boys and girls separately and for white and non-white children separately.

Bibliography:

Aggett, P. J., Barclay, S., & Whitley, J. E. 1989, "Iron for the suckling", *Acta Paediatr Scand Suppl*, vol. 361, pp. 96-102.

Allen, L. H. 1997, "Pregnancy and iron deficiency: unresolved issues", *Nutr Rev*, vol. 55, no. 4, pp. 91-101.

Beard, J. L. 2000, "Effectiveness and strategies of iron supplementation during pregnancy", *American Journal of Clinical Nutrition*, vol. 71, pp. 1288S-1294S.

Box, G. E. P. & Cox, D. R. 1964, "An analysis of transformation(with discussion)", *Royal Statistical Society*, vol. 26, no. B, pp. 211-252.

Breslow, N. E. & Clayton, D. G. 1993, "Approximate inference in Generalized Linear Mixed Models", *Journal of the American Statistics Association*, vol. 88, pp. 9-25.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. 1994, *HLM TM 2/3: Hierarchical Linear Modelling with the HLM/2L and HLM/3L Programs* Scientific Software International, Chicago.

Cheney, W. & Kincaid, D. 1985, *Numerical Mathematics and Computing*. Brooks / Cole Publishing Company, California.

Cnaan, A., Laird, N. M., & Slasor, P. 1997, "Using the General Linear Mixed Model to analysis unbalanced repeated measure and longitudinal data", *Statistics in Medicine*, vol. 16, pp. 2349-2380.

Cole, T. J. 1988, "Fitting Smoothed Centile Curves to Reference Data", *Royal Statistical Society*, vol. 151, no. 3, pp. 385-418.

Cole, T. J. 1994, "Growth Chart For Both Cross-Sectional and Longitudinal Data", *Statistics in Medicine*, vol. 13, pp. 2477-2492.

Cole, T. J., Freeman, J. V., & Preece, M. A. 1998, "British 1990 growth reference Centile for weight, height, body mass index and head circumference fitted by maximum penalized likelihood", *Statistics in Medicine*, vol. 17, pp. 407-429.

Cole, T. J. & Green , J. P. 1992, "Smoothing Reference Centile Curves :The LMS Method and Penalized Likelihood", *Statistics in Medicine*, vol. 11, pp. 1305-1319.

Cook, J. D., Skikne, B. S., & Baynes, R. D. 1994, "Circulating transferrin receptors and assessment of iron status", *Journal of Nutritional Biochemistry*, vol. 5, no. 7, pp. 322-330.

Cowin, I., Emond , A., Emmett, P., & The ALSPAC Study Team 2001, "Association between composition of the diet and haemoglobin and ferritin levels in 18-month-old children", *European Journal of Clinical Nutrition*, vol. 55, no. 278, p. 286.

Dawson, E. & McGanity, W. 1987, "Protection of maternal iron stores in pregnancy", *J Reprod Med*, vol. 32, pp. 478-487.

Diggle, P. J., Liang, K. Y., & Zeger, S. L. 1994, *Analysis of Longitudinal Data.* Oxford University Press, New York.

Dixon, W. J. & Chief , E. 1990, *BMDP Statistical Software Manual* University of California Press, Berkeley, CA, USA.

Duggan, M., Steel , G., Elwys , G., & Noble , C. 1991, "Iron status, energy intake, and nutritional status of healthy young Asian children", *Arch Dis Child*, vol. 66, no. 1386, p. 1389.

Emmett, P., North, K., & Noble, S. 2000, "Types of drinks consumed by infants at 4 and 8 months of age: a descriptive study. The ALSPAC Study Team", *Public Health Nutr*, vol. 3, no. 2, pp. 211-217.

Emond , A., Hawkins, N., Pennock, C., & Golding, J. 1996, "Haemoglobin and ferritin concentration in infants at 8 moths of age", *Arch Dis Child*, vol. 74, no. 1, pp. 36-39.

Fatti, L. P., Senaoana, E. M., & Thompson, M. L. 1998, "Bayesian Updating in Reference Centile Charts", *Royal Statistical Society*, vol. 161, no. 1, pp. 103-115.

Fitzmaurice, G. M., Laird, N. M., & Rotnitzky, A. G. 1993, "Regression Models for Discrete Longitudinal Responses", *Statistical Sciences*, vol. 8, pp. 284-309.

Fuchs, G. J., Farris, R. P., DeWier, M., Hutchinson, S. W., Warrier, R., & Doucet, H. 1993, "Iron status and intake of older infants fed formula vs cow milk with cereal", *American Journal of Clinical Nutrition*, vol. 58, pp. 343-348.

Gennaro, S., Fehder, W. P., Cnaan, A., York, R., Campbell, D. E., Gallagher, P. R., & Douglas, S. D. 1997, "Immune responses in mothers of term and preterm very-low-birth-weight infants", *Clinical and Diagnostic Laboratory Immunology*, vol. 4, no. 5, pp. 565-571.

Godfrey, K. M., Redman, C. W., Barker, D. J., & Osmond, C. 1991, "The effect of maternal anaemia and iron deficiency on the ratio of fatal weight to placental weight", *Br Journal Obstet Gynecol*, vol. 98, no. 9, pp. 886-891.

Golding, J. 1989, "European longitudinal study of pregnancy and childhood (ELSPAC)", *Paediatr Perinat.Epidemiol.*, vol. 3, no. 4, pp. 460-469.

Golding, J., Pembrey, M., Jones, R., & The ALSPAC Study Team 2001, "The Avon Longitudinal study of parents and children, study methodology", *Paediatric and peinatal Epidemiology*, vol. 15, pp. 74-87.

Golding, J., Pembrey, M., & Jones, R. 2001, "ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology", *Paediatr Perinat.Epidemiol.*, vol. 15, no. 1, pp. 74-87.

Gregoire, T. G., Brillinger, D. R., Diggle, P. J., Russek , C. E., Warren, W. G., & Wolfinger, R. D. 1997, *Modelling longitudinal and Spatially Correlated Data.* Springer-Verlag, New York.

Guidozzi, F., Patel, R., & McPhail, A. 1995, "A prospective study of iron status in white and black pregnant women an urban hospital", *The South African Medical Journal (SAMJ)*, vol. 85, pp. 171-173.

Hallberg , L. 1995, "Results of surveys to assess iron status in Europe", *Nutr Rev*, vol. 53, no. 11, pp. 314-322.

Hemminki, E. & Rimpela, U. 1991, "A randomized comparison of routine versus selective iron supplementation during pregnancy", *J Am Coll.Nutr*, vol. 10, no. 1, pp. 3-10.

Hercberg, S., Galan, P., & Preziosi, P. 2000, "Consequences of iron deficiency in pregnant woman", *Clinical Drug Investigation*, vol. 19, pp. 1-7.

International Food Policy Research Institute 2000, *Effects of Diet in Improving Iron Status of Women: What Role for Food-Based Interventions?*

Jong, N., Romano, A. B., & Gibson, R. 2002, "Zinc and iron status during pregnancy of Filipino women", *Asia Pacific J Clin Nutr*, vol. 11, no. 3, pp. 186-193.

Klem, K. 1993, "Anemia in one year old infants-still a problem?", *Tidsskr Nor Laegeforen*, vol. 113, pp. 327-328.

Koller, O., Sagen, N., Ulstein, M., & Vaula, D. 1979, "Fetal growth retardation associated with inadequate haemodilution in otherwise uncomplicated pregnancy", *Acta Obstet Gynecol Scand*, vol. 58, pp. 9-13.

Laird, N. M. & Ware, J. H. 1982, "Random effects models for longitudinal data", *Biometrics*, vol. 38, pp. 963-974.

Lao, T., Loong, E., Chin, R., Lam, C., & Lam, Y. 1991, "Relationship between newborn and maternal iron status and haematological indices", *Biology of the Neonate*, vol. 60, no. 5, pp. 303-307.

Liang, K. Y. & Zeger, S. L. 1986, "Longitudinal data analysis using Generalized Linear Models", *Biometrika*, vol. 73, pp. 13-22.

Little, R. J. A. & Rubin, D. B. 2002, *Statistical analysis with Missing Data.*, 2nd edn, John Wiley and Sons, New York.

Lonnerdal, B. & Hernell, O. 1994, "Iron, Zinc,Copper and Selenium status of breast fed infants and infants fed trace element fortified milk based infant formula", *Acta Paediatrca*, vol. 83, pp. 367-373.

Looker, A. C., Dallman, P. R., & Carroll, M. B. 1997, "Prevalence of iron deficiency in the United States", *The journal of the American Medical Association*, vol. 277, pp. 973-976.

Ma, A., Chen, X., Zheng, M., Wang, Y., Xu, R., & Li, J. 2002, "Iron status and dietary intake of Chinese pregnant women with anaemia in the third trimester", *Asia Pac.J Clin Nutr*, vol. 11, no. 3, pp. 171-175.

Male, C., Persson , L., Freeman, V., & Haschke, F. 2001, "Prevalence of iron deficiency in 12-mo-old infants from 11 European areas and influence of dietary factors on iron status (Euro-Growth study)", *Acta Paediatr*, vol. 90, pp. 492-498.

Manor, O. & Kark, J. D. 1996, "A comparative study of four methods for analysis of repeated measures data", *Statistics in Medicine*, vol. 15, pp. 1143-1159.

McCullagh, P. & Nelder, J. A. 1989, *Generalized Linear Models* Chapman & Hall , 2-6 Boundary Row, London , Cambridge.

McCulloch, C. E. & Searle, S. R. 2001, *Generalized , Linear and Mixed models.* John Wiley and Sons, Inc.

Michaelsen, K., Milman, N., & Amuelson, G. 1995, "A longitudinal study of iron status in healthy Danish infants;effects of early iron status, growth velocity and dietary factors", *Acta Paediatr*, vol. 84, pp. 1035-1044.

Milman , N., Agger, A. O., & Nielsen, O. J. 1991, "Iron supplementation during pregnancy. Effect on iron statue markers, serum erythropoietin and human placental lactogen. A placebo controlled study in 207 Danish women", *Dan Med Bull*, vol. 38, pp. 471-476.

Milman , N., Bergholt, T., Byg, K. E., Eriksen, L., & Graudal, N. 1999, "Iron status and iron balance during pregnancy. A critical reappraisal of iron supplementation", *Acta Obstet Gynecol Scand*, vol. 78, no. 9, pp. 749-757.

Milman , N., Byg, K. E., & Agger, A. O. 2000, "Hemoglobin and erythrocyte indices during normal pregnancy and postpartum in 206 women with and without iron supplementation", *Acta Obstet Gynecol Scand*, vol. 79, no. 2, pp. 89-98.

Milman, N., Agger, A., & Niesn, O. 1994, "Iron status markers and serum erythropoietin in 120 mothers and new born infants ;effect of iron supplementation in normal pregnancy", *Acta Obstet Gynecol Scand*, vol. 73, pp. 200-204.

Molenberghs, G. & Verbeke, G. 2002, *Linear mixed models for longitudinal data* Springer-Verlag, New York.

Morbidity and Mortality Weekly Report 1998, *Recommendations to prevent and control iron deficiency in the United States*, Centres for Disease Control and Prevention, USA, 47.

Mumford, S. E. 1999, "Children of the 90s: ethical guidance for a longitudinal study", *Arch Dis Child Fetal Neonatal Ed*, vol. 81, no. 2, p. F146-F151.

Murphy, J., O'Riordan, J., Newcombe, R., & Coles, E. 1986, "Relation of haemoglobin levels in first and second trimester to outcome of pregnancy", *Lancet* pp. 992-995.

Nestel, P. 1993, *Food fortification in developing countries* US Agency for International Development, Washington, DC.

Nicholas, J. H. & Stuart, R. L. 1999, "Review of Software to Fit Generalized Estimating Equation Regression Models", *American Statistical Association*, vol. 53, pp. 160-169.

Noble, S., Emmett, P., & The ALSPAC Study Team 2001, "Food and nutrient intake in a cohort of 8-month-old infants in the south-west of England in 1993", *European Journal of Clinical Nutrition*, vol. 55, pp. 698-707.

North, K., Emmett, P., & The ALSPAC Study Team 2000, "Multivariate analysis of diet among three-year-old children and association with social-demographic characteristic", *European Journal of Clinical Nutrition*, vol. 54, pp. 73-80.

Oski , F. A. 1989, *The causes of iron deficiency in infancy* Raven Press, New York.

Perez , A., Dennis, R. J., Gil, J., Rondon , M., & Lopez, A. 2002, "Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia", *Statistics in Medicine*, vol. 21, pp. 3885-3896.

Pinheiro, J. C. & Bates, M. D. 2000, *Mixed effects models in S and S-Plus*. Springer, New York.

Pisacane, A., De Vizia, B., Veccaro, F., Valiante, F., Russo, M., & Grillo, G. 1995, "Iron status in breast-fed infants", *Pediatrics*, vol. 127, pp. 429-431.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, *Numerical Recipes in FORTRAN; The art of scientific computing*. Cambridge University Press, Cambridge.

Preziosi, P., Prual, A., Galan, P., & Daouda, H. 1997, "Effect of Iron Supplementation on the Iron Status of Pregnant Women: Consequences for Newborns", *Am J Clin Nutr*, vol. 66, no. 5, pp. 1178-1182.

Rogers, I., Emmett, P., & The ALSPAC Study Team 1998, "Diet during pregnancy in a population of pregnant women in South West England", *European Journal of Clinical Nutrition*, vol. 52, no. 4, pp. 246-250.

SAS Institute 1992, *SAS/STAT Software: Changes and Enhancements for Release 6.07*, SAS institute Inc, Cary,NC, SAS Technical Report P-229.

Scanlon, K., Yip, R., Schieve, L., & Cogswell, M. 2000, "High and low haemoglobin levels during pregnancy: differential risks for preterm birth and small for gestational age", *Obstet Gynecol*, vol. 96, pp. 741-748.

Schafer, J. L. 1997, *Analysis of incomplete multivariate data.* Chapman & Hall, London.

Scheffer, J. 2002, "Dealing with Missing Data", *Ref.Lett.Inf.Math.Sci*, vol. 3, pp. 153-160.

Scheike, T. H., Zhang, M. J., & Juul, A. 1999, "Comparing Reference Charts for Cross-Section and Longitudinal Data", *Biometrical Journal*, vol. 41, pp. 679-687.

Searle, S. R. 1971, *Linear Models John Wiley & Sons, Inc, New York.*

Searle, S. R., Casella, G., & McCulloch, C. E. 1992, *Variance Components.* John Wiley & Sons, New York.

Shah, B. V., Barnwell, B. G., & Bieler, G. S. 1997, *SUDAAN User's Manual*, Research Triangle Institute, Research Triangle Park, NC.

Sherriff, A., Emond, A., Bell, J. C., Golding, J., & Team, t. A. S. 2001, "Should infants be screened for anaemia? A prospective study investigating the relation between haemoglobin at 8, 12, and 18 months and development at 18 months", *Archives of Disease in Childhood*, vol. 84, no. 6, pp. 480-485.

Sherriff, A., Emond, A., Hawkins, N., Golding, J., & the ALSPAC Children in Focus Study Team 1999, "Haemoglobin and ferritin concentrations in children aged 12 and 18 months", *Archives of Disease in Childhood*, vol. 80, no. 2, pp. 153-157.

Sidebotham, P., Golding, J., & The ALSPAC Study Team 2001, "Child malteatment in the "Children of the Nineties" A longitudinal study of parental risk factores", *Child Abuse and Neglect*, vol. 25, pp. 1177-1200.

Sidebotham, P. & Golding, J. 2001, "Child maltreatment in the "children of the nineties" a longitudinal study of parental risk factors", *Child Abuse Negl.*, vol. 25, no. 9, pp. 1177-1200.

Siimes, M. A., Salmenpera, L., & Perheentupa, J. 1984, "Exclusive breast-feeding for 9 months: risk of iron deficiency", *Journal of Pediatr*, vol. 104, pp. 196-199.

Tapiero, H., Gate, L., & Tew, K. D. 2001, "Iron: deficiencies and requirements", *Biomedicine & Pharmacotherapy*, vol. 55, no. 6, pp. 324-332.

Twisk, J. & de Vente, W. 2002, "Attrition in longitudinal studies. How to deal with missing data", *J Clin Epidemiol.*, vol. 55, no. 4, pp. 329-337.

Verbeke , G., Spiessens, B., & Lesaffer, E. 2001, "Conditional Linear Mixed Models", *American Statistician*, vol. 55, no. 1, pp. 25-34.

Walravens, P. A. 1989, "Nutritional Sources of Iron in Infants and Toddlers.," in *Dietary Iron :Birth to Two Years*, Raven Press, New York.

Walter , T., Dallman , P. R., Pizarro , F., Velozo , L., Pena , G., & Bartholmey , S. J. 1993, "Effectiveness of iron fortified cereal in prevention of iron deficiency anaemia", *Pediatrics*, vol. 92, pp. 976-982.

Ware, J. H. 1985, "Linear models for the analysis of longitudinal studies", *American Statistician*, vol. 39, pp. 95-101.

Wharf, S. G., Fox, T. E., Fairweather, S. J., & Cook, J. D. 1997, "Factors affecting iron stores in infants 4-18 months of age", *European Journal of Clinical Nutrition*, vol. 51, pp. 504-509.

WHO 1992, *The prevalence of anemia in women: a tabulation of available information*, World Health Organization, Geneva.

WHO 1997, *The third report on the world nutrition situation*, ACC/SCN Secretariat,World Health Organization, Geneva,Switzerland.

WHO 1998, *Complementary Feeding of Young Children in Developing Countries; A Review of Current Scientific Knowledge.* Geneva,World Health Organization, WHO/NUT/98.1.

Williams, M. D. & Wheby, M. S. 1992, "Anemia in pregnancy", *Med Clin North Am*, vol. 76, no. 3, pp. 631-647.

Woodhouse, G. 1995, *A guide to MLn for new users.*, Institute of Education, University of London, London.

Wright, E. M. & Royston, P. 1997, "A Comparison of Statistical Methods for Age-related Reference Intervals", *Royal Statistical Society*, vol. 160, no. 1, pp. 47-69.

Wright, E. M. & Royston, P. 1999, "Calculating reference intervals for laboratory measurements", *Statistical Methods In Medical Research*, vol. 8, pp. 93-112.

Zeger, S. L. & Liang, K. Y. 1986, "Longitudinal data analysis for discrete and continuous outcomes", *Biometrics*, vol. 42, pp. 121-130.

Zeger, S. L. & Liang, K. Y. 1992, "An overview of Methods for the Analysis of Longitudinal Data", *Statistics in Medicine*, vol. 11, no. 14-15, pp. 1825-1839.

Ziaei, S., Hatefian, E., & Togeh, G. 2002, "Iron Status in Newborns Born to Iron-Deficient Mothers", *Iranian Journal of Medical Sciences*, vol. 28, no. 2, pp. 62-64.

Zorn, C. J. W. 2001, "Generalized Estimating Equation Models for Correlated Data :A Review with Applications", *American Journal of political Sciences*, vol. 45, no. 2, pp. 470-490.