Statistical Methods of Constructing Growth Charts

Elizabeth Rose Irwin

A Dissertation Submitted to the

University of Glasgow

for the degree of

Master of Science

School of Mathematics & Statistics

November 2013

Abstract

People are interested in monitoring growth in many fields. Growth charts provide an approach for doing this, illustrating how the distribution of a growth measurement changes according to some time covariate, for a particular population. The general form of a growth chart is a series of smooth centile curves showing how selected centiles of the growth measurement change when plotted against the time covariate. These curves are based on a representative sample from a reference population. Different modelling approaches are available for producing such growth charts, including the LMS method and quantile regression approaches. These approaches are explored in this thesis using data from the Growth and Development Study data, which allows construction of gender-specific weight growth charts for full-term infants.

Acknowledgements

I am heartily thankful to my supervisor, Dr Tereza Neocleous, whose enthusiasm, support and guidance throughout my Masters has allowed me to develop a real understanding of this subject. I would also like to thank Professor Charlotte Wright, who not only provided the data which made this thesis possible, but also some very helpful insights. I would also like to thank the Information Service Division(ISD) for funding my research and my family and friends for their continuing encouragement throughout my Master's year.

Declaration

I have prepared this thesis myself; no section of it has been submitted previously as part of any application for a degree. I carried out the work reported in it, except where otherwise stated.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

People are interested in monitoring growth in many fields. Growth charts provide an approach for doing this, illustrating typical growth patterns, describing how a growth measurement changes according to some time covariate, often age, for a particular population. They are constructed on a reference population which contains a representative sample from this population, whose measurement may have been observed at multiple points (ages) during growth. The general form of a growth chart is a series of smoothed centile curves, showing how selected centiles of the growth measurement change when plotted against the time covariate. They typically illustrate reference centile curves for a symmetric subset of the 5th, 10th, 25th, 50th, 75th, 90th and 95th, with the 50th centile representing the median (Cole, 1988). These reference centile curves separate the reference population into parts, with for example the 5th centile curve representing that five percent of the reference populations growth measurements are less than or equal to the estimated 5th centile curve value at each value of the time covariate(each age) and 95 percent above. The reference centile curves therefore give an impression of the rate of change in all parts of the growth measurements distribution.

My research primarily focuses on growth charts constructed for infants' weight measurements, which depict reference centile curves illustrating how infants' weights change between birth and roughly two years of age.

Such reference curves are used to monitor infants during the early ages of development, by determining which centile any given infant lies on at a particular age given their recorded weight measurement.

Infants whose values move between the centiles with passing age, as well as those with values that lie outside the reference range are viewed as potentially having a concerning growth pattern, which should be further investigated. These reference growth charts are therefore widely used in medical practice as a screening tool (Cole and Green, 1992).

It is crucial that gender-specific growth charts are constructed, as there are likely to be differences in how weight changes with age between female and male infants.

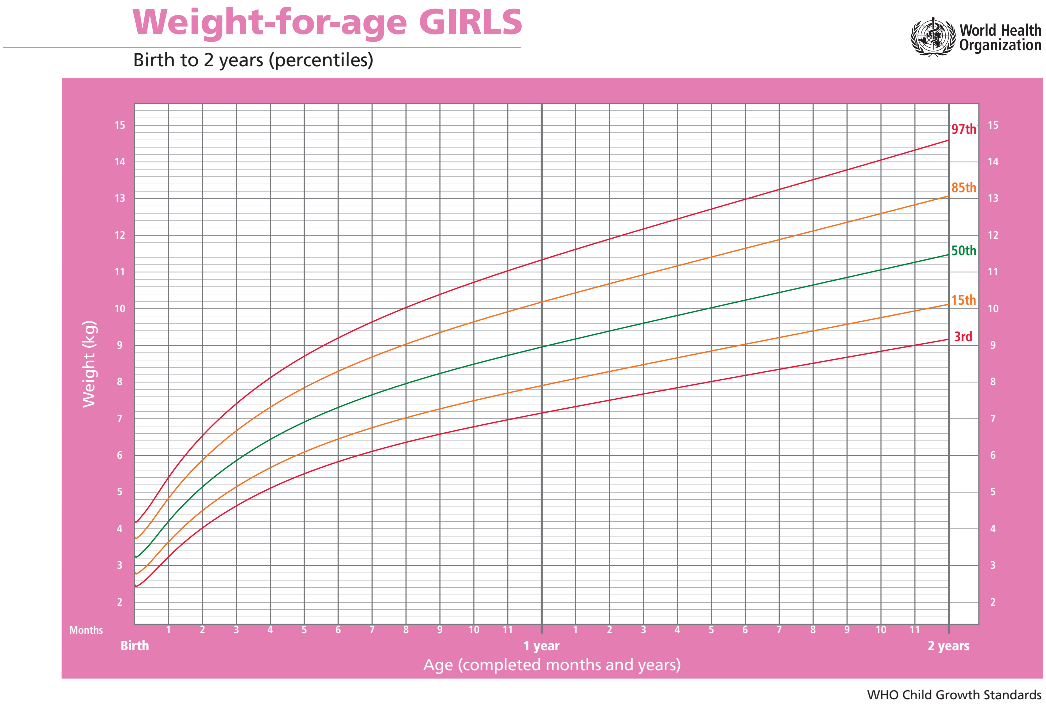
Reference growth curves, which condition on age, only provide a valuable snapshot of the dispersion of growth measurements at various ages, whereas reference growth curves which condition on age, as well as prior growth history and other crucial additional information such as parental heights, can be more informative. They allow for a more insightful explanation into an individual's current growth measurement.

The World Health Organisation (WHO) weight-for-age child growth standard seen in Figure 1.1 is used internationally to monitor growth in infants and children from birth to two years of age. This standard, which was updated in 2006, is based on WHO Multicentre Growth Reference Study (MGRS) designed explicitly for creating growth charts (de Onis et al., 2006). The MGRS, which was implemented between 1997 and 2003, collected growth data and related information from 8440 healthy breastfed infants and young children from diverse ethnic backgrounds and cultural settings de Onis et al. (2006). The purpose of using such a diverse reference population was to allow the construction of growth charts which could be used internationally. Although a variety of methods for studying growth have been proposed (e.g.Pan and Goldstein (1997)), the LMS method is the most commonly applied technique for constructing growth charts. The LMS methodology has been widely applied among other methods (e.g. GAMLSS with the Box-Cox power exponential distribution, Rigby and Stasinopoulos (2004)) for constructing the WHO growth standards (de Onis et al., 2006). My research aims to explore the LMS method, an approach discussed in detail in chapter 3, and several other approaches of constructing growth charts.

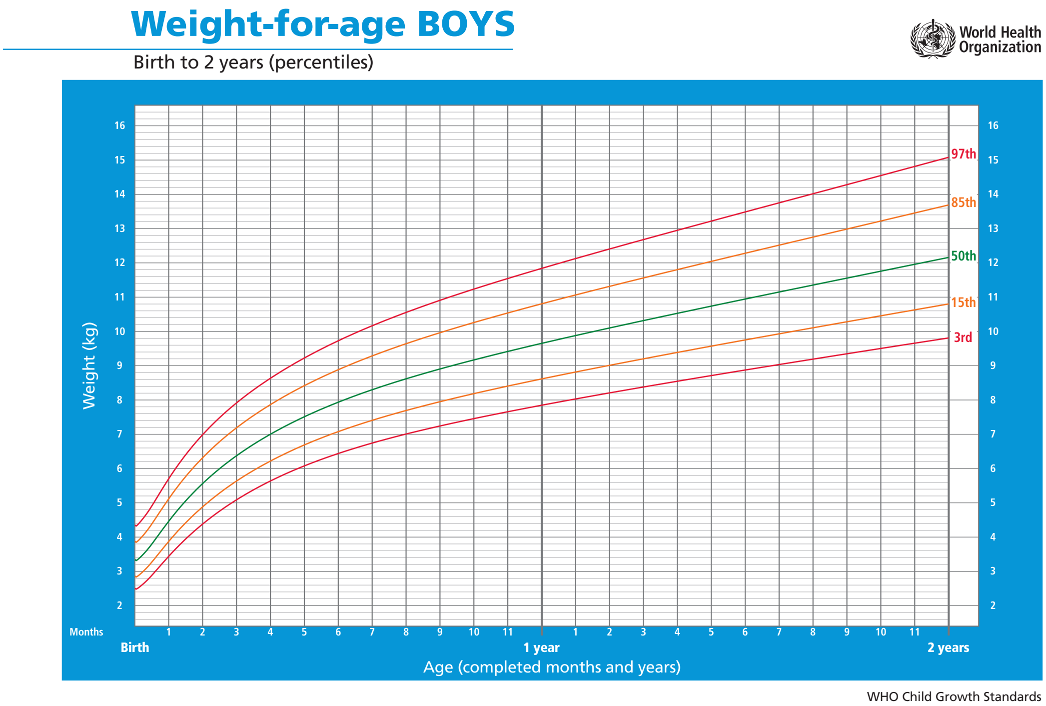
1.1 Growth And Development Study Data

The different statistical approaches to growth chart modelling examined in my research are primarily applied to data from a Growth and Development Study from 1994 which investigated growth in infancy in Newcastle upon Tyne (Wright et al., 1994). This data was kindly provided by Charlotte M Wright, Professor of Community Child Health at the University of Glasgow.

This cohort study contains 3658 infants who were identified using the Child Health Computer system as being aged between 18-30 months and living in Newcastle upon Tyne in November 1989. The Child Health Computer system covers a range of functionalities, which includes registration of infants at birth and documentation of demographical information (Wales National Health Service, 2013). The infants' health records were then reviewed to collect their birth weight (kg) and up to ten subsequent weights (kg) between birth and 1132 days of age, together with some other limited medical information. The ten subsequent weights which may have been documented in these records, were the infant's weights observed at around one, two, three, four, five, six, eight, ten and twelve months after birth as well as their last available weight after 12 months.



(a) Female Full-Term Infants



(b) Male Full-Term Infants

Figure 1.1: WHO weight-for-age child growth standards

The focus of this research was the subset of 3418 full-term infants excluding the 235 infants born before 37 weeks gestation. This is because these pre-term infants are likely to be less healthy, weighing less at birth and will therefore tend to grow differently in their early weeks of development. The general practice is for separate growth charts, formerly called *Low Birth Weight Charts*, to be used to plot growth patterns of such pre-term infants and those with significant early health problems (Royal College of Paediatrics and Child Health, 2013). It therefore seems inappropriate for the study data on pre-term infants to be considered when trying to construct growth charts modelling typical infants growth patterns.

In this study there are an almost even proportion of full-term infants of both genders, with 1712 males and 1706 females. This is a positive quality to the data as it allows suitable growth charts to be modelled for both genders.

Five years after the study was first established, when the infants were aged 8-9 years, a 20% systematic sample was taken of the 2812 full-term infants for whom at least three weights had been retrieved (Wright and Cheetham, 1999). The infants in this 20% sample were then traced and a letter and consent form was sent to the family, which included a request for both parents' heights. Infants were then measured in school by a research nurse over an eight month period. Heights were recorded to 0.1mm using the Leicester height measurer and this additional data is also available for our analysis.

1.2 Exploratory Analysis of Growth and Development Study Data

Figure 1.2 shows how the weight of full-term infants gradually increases with age. However the rate of increase appears to steadily reduce with age,

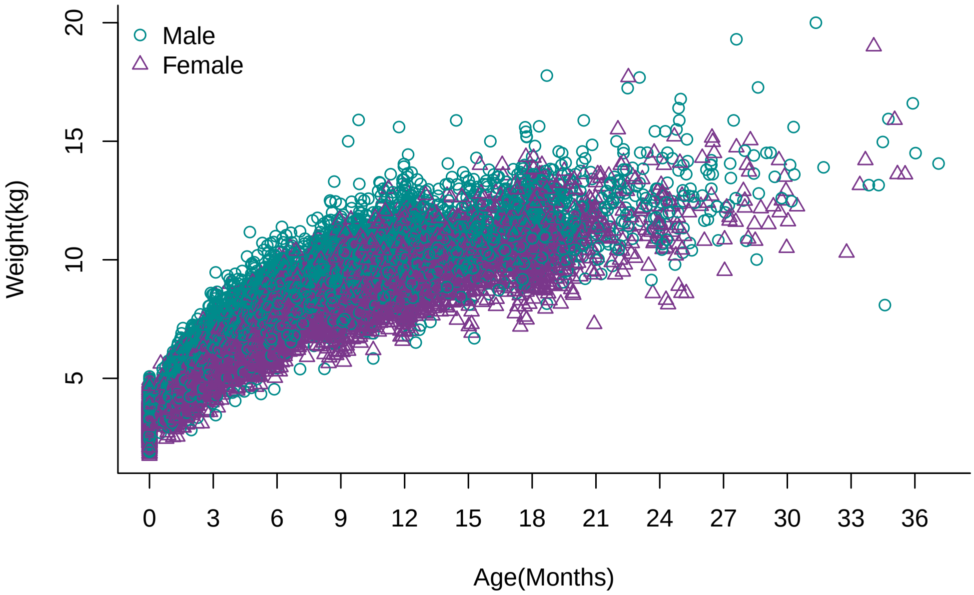


Figure 1.2: Plot of weight measurements of full-term infants in the Growth and Development Study data by gender, between birth and 37 months of age.

reaching a near-plateau by the end of the first year, and continues to taper off gently from this point onwards. This is the expected overall growth pattern under conditions of adequate nutrition and psychosocial care with no chronic infections or unusual rates and/or severity of acute infections (de Onis et al., 2009). The number of weight measurements recorded for full-term infants in this study becomes more limited with age, so the trend in the tail of this distribution is not as clear. This trend in growth is almost identical between full-term female and male infants with a substantial overlap in records between infants of both genders. However in some cases the recorded weight measurements for male infants are slightly higher than those for female infants of the same age. This overall trend observed is clearly non linear and thus the approaches considered in my research allow the curved nature of the trend to be incorporated into the modelling of the growth charts.

Table 1.1: Summary statistics for birth weights of full-term infants in the Growth and Development Study data by gender

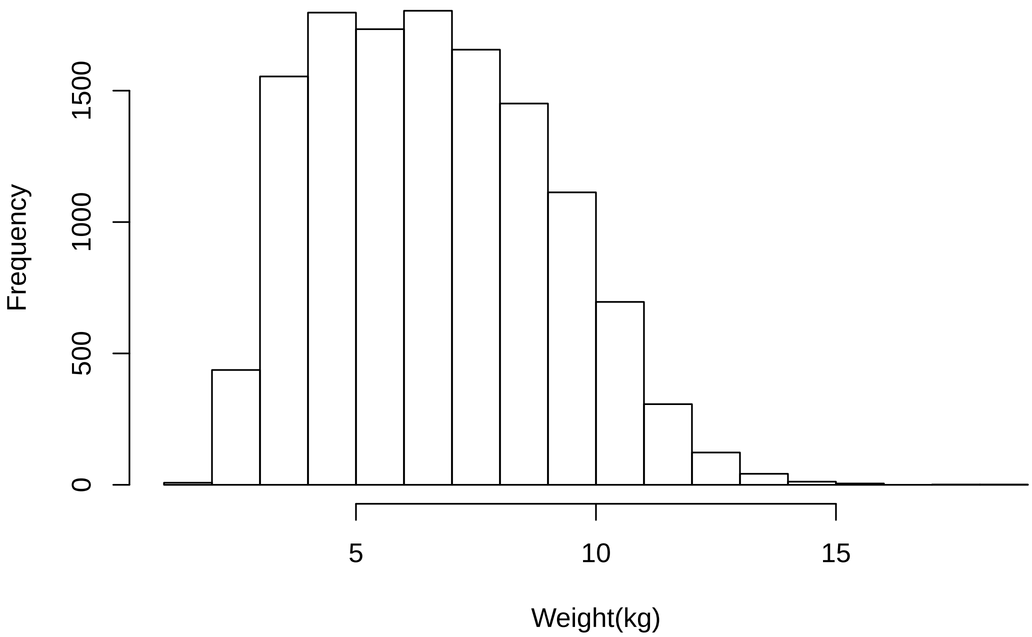
Gender	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	Standard Deviation
Female	1.730	2.980	3.290	3.290	3.600	4.920	0.488
Male	1.900	3.090	3.430	3.431	3.750	5.080	0.494

Table 1.1 and Figure 1.3 show that there is a substantial overlap in the recorded weights of full-term male and female infants. However as indicated from Figure 1.2, the distribution of male infants' birth weight is slightly shifted to the right obtaining a median birth weight of 3.431 kg in comparison to 3.290 kg for female infants. Furthermore the mean birth weight for male infants is 3.431 kg, 0.141 kg higher than the female infant mean birth weight.

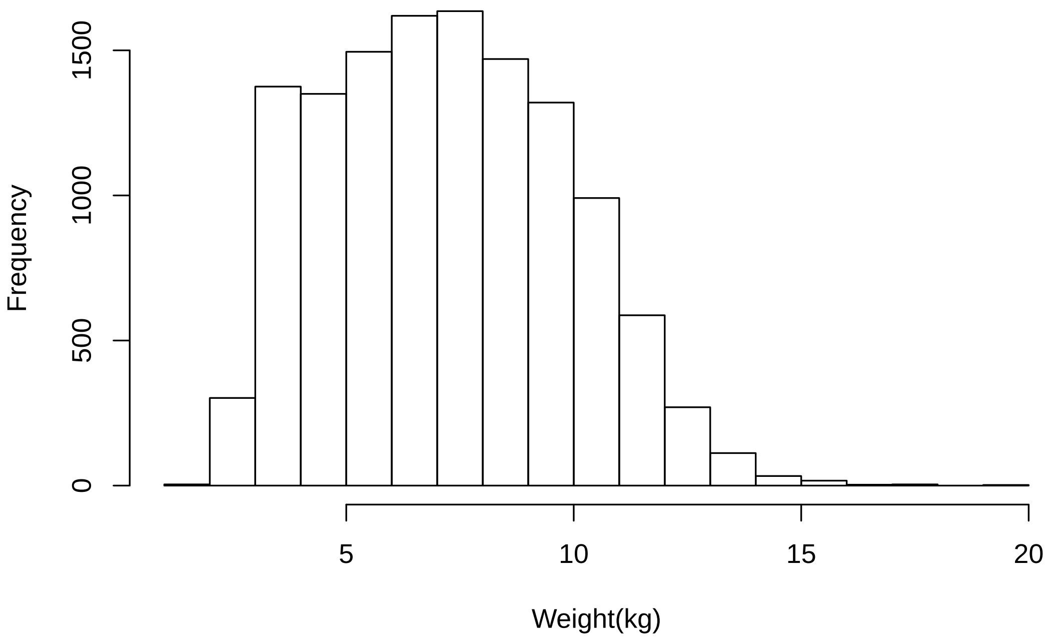
1.3 Case Infants

For illustration purposes, throughout this thesis, screening based on the growth charts constructed by each of the considered statistical methods is performed on four selected case infants from the Growth and Development Study, who were identified as experiencing unusual growth patterns.

Figure 1.4 illustrates the growth patterns of the two female case infants, showing that Subject 1500 had considerably lower weight measurements than most of her peers. However her rate of growth appears to follow the typical trend identified from Figure 1.2. Subject 146's birth weight of 2.92 kg was relatively low, 0.37 kg lower than the average birth weight for full-term female infants, however she then displays rapid growth till roughly 8 months of age. After this point a sudden drop in growth rate was observed. Table 1.2 gives more precise details on these measurements, indicating that at the age of roughly 12 months, the age at which the screening decision is considered, almost 4 months since her last measurement, subject 146 is reported to have



(a) Female full-term infants



(b) Male full-term infants

Figure 1.3: Histograms of weights (Kg) of full-term infants by gender

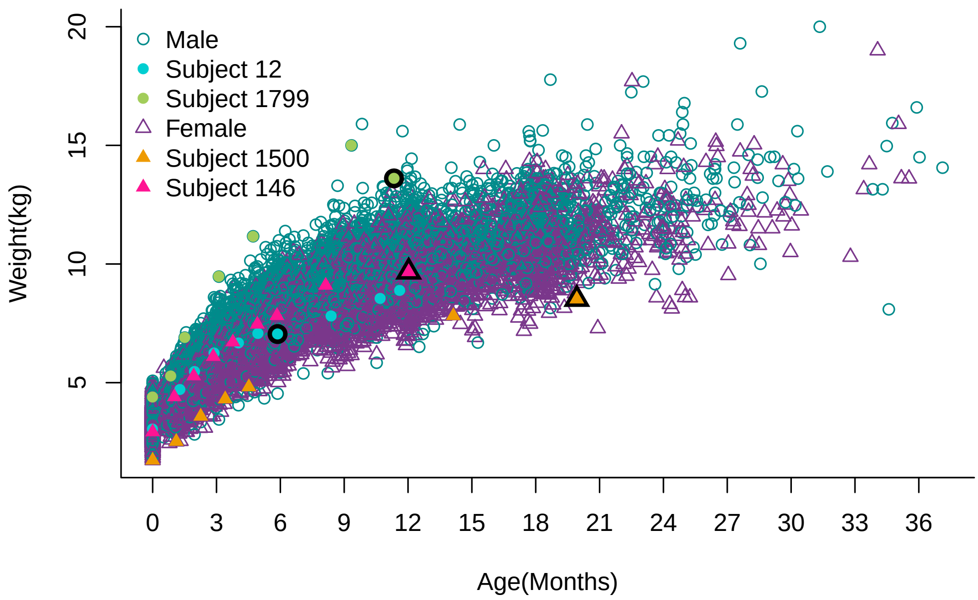


Figure 1.4: Plot of weight measurements of full-term infants in the Growth and Development Study by gender. Highlighted are the weight measurements observed for each of the four case infants, with the point bordered in black in each case denoting the observation at which the screening decision is considered.

gained only 0.57 kg.

Figure 1.4 demonstrates the growth patterns of the two male case infants, showing that subject 12 was only slightly below the median weight at birth. He then showed an unusual growth pattern up to the age of 4.95 months, with weight continually increasing but with the rate of growth reducing as age increased. A small fall in weight of 0.03 kg was then observed between the age of 4.95 and 5.87 months; the age at which the screening decision is considered, after which he grew steadily. Subject 1799 was heavier than most of his peers at birth. He then continually showed a increase in weight up to the age of 9.64 months, placing his weight well above the typical weight observed for full-term male infants of his age. However after this point a

		Measurements								
	Subject	1	2	3	4	5	6	7	8	9
Age (Months)	1500	Birth	1.11	2.26	3.41	4.52	14.13	19.93		
Weight (Kg)	1500	1.73	2.52	3.58	4.31	4.82	7.84	8.52		
Age (Months)	146	Birth	1.02	1.93	2.85	3.77	4.92	5.84	8.13	12.03
Weight (Kg)	146	2.92	4.4	5.28	6.1	6.71	7.46	7.82	9.1	9.67

Table 1.2: Weight measurements of the two full-term female case infants.

substantial drop in weight was observed, with him reported to have lost 1.39 kg by the age of 11.34 months; the age at which the screening decision is considered. Table 1.3 indicates that at the age of 11.34 months, 2 months since his last measurement, subject 1799 is reported to have lost 1.39 kg.

		Measurements									
	Subject	1	2	3	4	5	6	7	8	9	10
Age(Months)	12	Birth	1.28	1.97	2.89	4.03	4.95	5.87	8.39	10.69	11.61
Weight(Kg)	12	3.05	4.71	5.48	6.25	6.68	7.08	7.05	7.82	8.54	8.88
Age(Months)	1799	Birth	0.85	1.51	3.11	4.72	9.34	11.34			
Weight(Kg)	1799	4.4	5.29	6.9	9.47	11.17	15	13.61			

Table 1.3: Weight measurements of the two full-term male case infants.

1.4 Other Datasets

The data from the Fourth Dutch Growth Study, (Fredriks et al., 2000a) (Fredriks et al., 2000b), are also used to illustrate several smoothing methods for curve estimation, identifying how changing particular properties of smoothing approaches influence the curves produced. This was a nationwide cross-sectional study of growth and development of the Dutch population between birth and 21 years of age. The data was collected by trained health care professionals and measured, among other variables, the height and weight of participants, using these measurements to calculate their body mass index (BMI). The BMI data from this study for 7482 of the male participants are provided in the **R** package **GAMLSS**(Stasinopoulos and Rigby, 2007).

Furthermore, the Abdominal Circumference Data available in the **R** package **GAMLSS**, was used to illustrate the linear quantile regression model approach. This study, also discussed in Stasinopoulos and Rigby (2007), recorded the abdominal circumference taken from fetuses during ultrasound scans at Kings College Hospital, London, at gestational ages ranging between 12 and 42 weeks. The data available in the **GAMLSS** package includes the abdominal circumference of 610 fetuses.

1.5 Overview of Thesis

Chapter 2 discusses smoothing techniques, which will be required for producing growth charts under some of the studied modelling approaches. This includes detailed descriptions of natural cubic splines, B-splines, P-splines and monotonically constrained splines.

Chapter 3 gives a detailed description of the LMS model approach, which produces reference growth curves that allow for conditionality on a time covariate, often age, and assumes the data follows a normal distribution once a suitable power transformation has been performed. This is the statistical method used to construct the WHO weight-for-age child growth standards. The Growth and Development Study Data, described previously, is used to illustrate the LMS method for composing gender-specific weight growth charts for full-term infants, firstly using the **lmsqreg** package and then the **lms** function in the **GAMLSS** package, which are both available in **R**. Visual comparison of the curves produced via these packages and screening of the four case infants based on their gender-specific reference weight growth chart is performed.

Chapter 4 describes the quantile regression model approach, a non-parametric approach which also composes reference growth curves that condition on a time covariate using both unpenalised B-splines and P-splines, the latter fitted using the package **quantreggrowth** (Muggeo et al., 2012) in **R**. The Growth and Development Study Data is used to illustrate the suitability of the quantile regression model for composing gender-specific weight growth charts for full-term infants. Visual comparison of these gender-specific growth weight charts to those composed using the LMS approaches is also performed, as well as comparison of the LMS method and quantile regression approach in terms of centile estimates deduced for the four case infants.

Chapter 5 discusses an extension of the quantile regression approach, which allows conditionality on age as well as prior growth history and additional relevant data. The Growth and Development Study data is used to consider models that allow conditionality on age and a prior weight measurement; conditionality on age and two prior weight measurements and those which additionally incorporate average parental height. Screening based on the resulting growth charts is then performed on the four case infants and comparisons to the conclusions drawn from the previous approaches is made.

Chapter 6 describes a user friendly interactive web application which was designed using the **R** package **shiny**, and allows monitoring of new infant weight measurements based on reference growth charts modelled on the Growth and Development Study data, composed via several of the modelling approaches discussed in the previous chapters.

Chapter 7 describes the conditional gain SD score approach which is an alternative approach to constructing growth gain references that allows conditionality on a time covariate as well as a prior growth measurement, by looking at the change in SD scores. The World Health Organisations' (WHO)

Child Growth Velocity Standards for weight, uses this proposed approach for construction of these standards. The conditional gain SD score approach is applied to the Growth and Development Study data. Comparison of the four case infants centile estimates at their screening age is made directly to those obtained when modelling using the longitudinal model approach.
This chapter then concludes the effectiveness of the different statistical methods of constructing growth charts, discusses the limitations associated with each modelling approach and details further work which could be performed.

The Appendix contains a table, detailing the models labelled and referred to throughout the thesis.

Chapter 2

Smoothing Methods for Growth Curve Estimation

A nonlinear trend is generally exhibited in growth charts, so smoothing techniques are required for modelling the relationship between the growth measurement and the time covariate.

Smoothing techniques can be used to model the relationship between the response growth variable and the time covariate without specifying any particular form for the underlying regression function $f(x)$, which describes their relationship. This function $f(x)$ can be estimated by a smooth function $\hat{f}(x)$. This process is often called nonparametric regression and fits the model

$$Y_i = f(x_i) + \epsilon_i, \qquad i = 1, \cdots, m \tag{2.1}$$

in the case of one covariate where Y denotes the response growth variable, x the time covariate and ϵ_i is usually assumed to be an independent error term with mean 0 and variance σ^2 (Fox, 2002).

Smoothers have two main purposes. Firstly they provide a way of exploring and presenting the relationship between the covariate and response variable, which consequently allows predictions of observations to be made without reference to a fixed parametric model (Silverman, 1985). Secondly they estimate interesting properties of the curve that describe the dependence between the mean of Y and the predictor x.

Smoothing methods that are well established include moving averages, kernel and local polynomial regression, smoothing splines, regression splines, and penalised regression splines (Meyer, 2012). The methods which smooth using splines are nonparametric regression curve fitting approaches, which represent the fit as a piecewise regression. They are able to provide a natural and flexible approach to curve function estimation, which copes well whether or not observations are observed at regular intervals (Silverman, 1985). A spline is defined as a function that is constructed piece-wise from polynomial functions, which are joined together smoothly at pre-defined subintervals of x. These connection points are referred to as knots.

The main difference between smoothing splines methodology with regression and penalised regression splines methodology is that smoothing splines explicitly penalise roughness and use the data points themselves as potential knots, whereas regression splines can place knots at any point, usually at equidistant/equiquantile points (Nie and Racine, 2012).

2.1 Smoothing Splines

The most widely used approach to curve fitting is least squares. The residual sum of squares $\sum_{i=1}^{m}(y_i - f(x_i))^2$ can be minimised by choosing $\hat{f}(x)$ to be any function that interpolates the data provided that x_i are all distinct, but such interpolation would not be satisfactory (Silverman, 1985), because it is almost certainly too rough. Therefore, to avoid this, a second term is added to the expression which is a measure of the local curvature of the function. This term, referred to as a *roughness penalty* is the integrated squared derivative of the regression function and will be large when $f(x)$ is rough with a rapidly changing slope (Fox, 2002). The modified sum of squares is then given by

$$\sum_{i=1}^{m}(y_i - f(x_i))^2 + \lambda \int_{x_{min}}^{x_{max}} f''(x)^2 dx$$

where λ is a smoothing parameter. Increasing λ penalises fluctuations, and so produces a smoother curve. For this choice of roughness penalty, the function $\hat{f}(x)$ which minimises this modified least squares criterion is a natural cubic spline with knots at the distinct observed values of x, with λ used to control the smoothing (Fox, 2002). This means that $\hat{f}(x)$ is a piecewise cubic polynomial in each interval (x_i, x_{i+1}). Therefore when the data is divided up into m' intervals by $m'+1$ knots, the curve estimator can be produced by fitting a piecewise function of the form

$$f(x) = \begin{cases} f_1(x) & \text{if } x_1 \leq x < x_2; \\ f_2(x) & \text{if } x_2 \leq x < x_3; \\ & \vdots \\ f_{m'}(x) & \text{if } x_{m'} \leq x < x_{m'+1}. \end{cases} \tag{2.2}$$

where f_i is a third degree polynomial defined by

$$f_i(x) = \alpha_i(x - x_i)^3 + \beta_i(x - x_i)^2 + \theta_i(x - x_i) + \gamma_i$$

for i=1,$\cdots$,m'+1. (Mathews and Fink, 2004).

The functions $\hat{f}(x)$, $\hat{f}'(x)$ and $\hat{f}''(x)$ are continuous. This function $\hat{f}(x)$ should give the best compromise between the smoothness and goodness of fit for the function, for the given value of λ. Natural cubic splines require that the value of the second and third derivative at the minimum and maximum values of x are both equal to zero. This implies that the function is linear beyond the boundary knots. The complexity of the curve can alternatively be controlled by adjusting the equivalent number of degrees of freedom (e.d.f) instead of defining the λ value directly. The effective degree of freedom is the trace of the smoother matrix, ie $tr(S)$, where the smoother matrix S is defined as the linear operator that acts on the data to produce the estimate, such that

$$\hat{f}(x) = Sy$$

where $f(x)$ is the vector of fitted values of each of the explanatory values from the fitted model and y is the original vector of responses. A full discussion of the smoother matrix is given in Wood (2006). The e.d.f controls how rough or flexible the curve will be, and it is quite common for the smoothness of the fitted curve to be controlled by varying the e.d.f.

Cubic smoothing splines are among the most commonly used splines for practical and computational reasons and can be fitted using the **smooth.spline** function in **R**.

Figures 2.1a and 2.1b illustrates smooth curves fitted by natural cubic splines to the Fourth Dutch Growth study data, which is detailed in section 1.4, showing the effect of differing the value of the smoothing parameter λ or e.d.f value. This smoothing method performs well, capturing the discernible trend in BMI with age, even when a small value of λ or e.d.f is given. The curves evidently become less flexible and more smooth as λ increases, whereas conversely they become more flexible and less smooth as the e.d.f value increases.

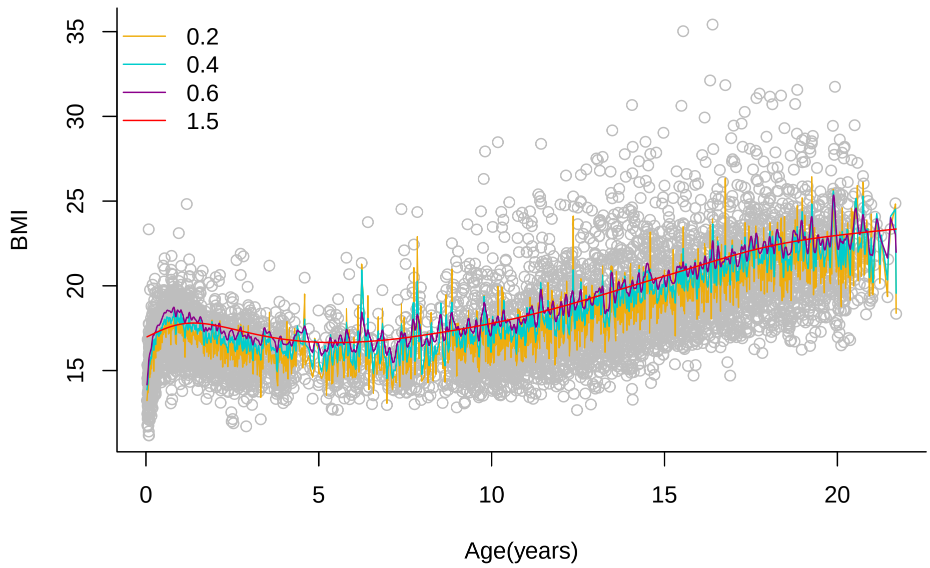
2.2 Regression Splines

B-splines are also attractive for nonparametric modelling. These, as well as other spline approaches, are underpinned by a set of known functions called basis functions, which are a common way to build a smooth function. Smooth functions can be approximated using weighted sums of the individual functions. While there are a wide variety of basis systems available, the choice of basis system is often dependent on the data to which the smooth function are to be fitted. For a general model of the form

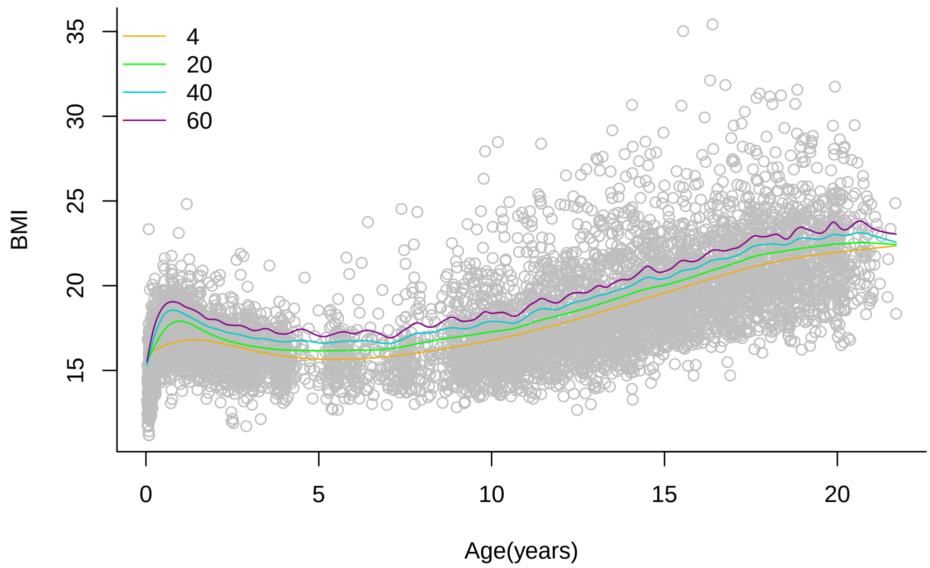
$$Y_i = f(x_i) + \epsilon_i \quad i = 1, \cdots, m,$$

a curve estimator can be produced by fitting the regression

$$Y_i = \alpha_0 B_0(x_i) + \alpha_1 B_1(x_i) + \alpha_2 B_2(x_i) + \cdots + \alpha_p B_p(x_i) + \epsilon_i$$



(a) Smoothing parameter λ specificied



(b) e.d.f specified

Figure 2.1: Plots of BMI of the 7482 male participants in the Fourth Dutch Growth Study, between birth and 21 years of ages. Superimposed are smooth curves fitted by natural cubic splines with smoothing parameter λ values between 0.2 and 1.5(corresponding e.d.f values between 4 and 60). For clarity the curves are offset from each other by 0.5 BMI units.

where the B_j are referred to as a basis functions. Therefore,

$$f(x) = \sum_{j=0}^{p} \alpha_j B_j(x).$$

An example of the basis function matrix for a polynomial of degree p would be

$$\begin{pmatrix} 1 & x_1 & \cdots & x_1^p \\ 1 & x_2 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & \cdots & x_m^p \end{pmatrix}$$

where the basis functions are $B_0(x) = 1,\ B_1(x) = x, \cdots, B_p(x) = x^p$.

This idea can then be extended to polynomial B-spline basis functions which are particularly flexible and computationally efficient for model fitting and are amongst the most commonly used basis systems. One of their key attributes is the compact support property which means that the basis functions are strictly local, with each basis function being strictly only non-zero over the interval between a small number of adjacent knots. This property results in a relatively sparse design matrix which makes B-splines computationally efficient. Polynomial B-spline basis functions are the most commonly used basis system and are composed of known spline functions, polynomial segments, which are joined together smoothly at pre-defined subintervals of x. Linear combinations of these spline functions can provide simple and quite flexible curve. In other words, the range x_{min} to x_{max} is divided into n' equal intervals by $n' + 1$ knots and each intervals is covered by $q + 1$ B-splines of degree q (Eilers and Marx, 1996). Hence the total number of knots required to form the B-spline is $n' + 2q + 1$ and the number of B-splines in the regression is $n = n' + q$.

The choice of the number of knots is critical when modelling with B-splines and has been a subject of much research, with too many knots leading to overfitting of the data and too few leading to underfitting (Eilers and Marx, 1996). In addition it must be decided if it is appropriate to have knots at equally spaced intervals or if more knots are needed in intervals of higher variability in the response variable y. Equally spaced knots are where knots are positioned at evenly spaced intervals of the covariate (age). Quantile knots are usually unequally spaced and if for example two quantile knots are implemented then one third of the data would fall below the first knot and two thirds below the second knot. Once the knots are given, the B-splines are computed recursively for any desired degree of the polynomial (Eilers and Marx, 1996). Typically natural quadratic ($q = 1$) or cubic B-splines ($q = 2$) are implemented which consist of connecting linear and quadratic pieces, respectively.

Let $B_j(x; q)$ denote the value of the jth B-spline of degree q at x for a given set of equally spaced knots. Then a fitted curve $\hat{f}$ to data (x_i, y_i) is the linear combination

$$\hat{f}(x) = \sum_{j=1}^{n} \hat{\alpha}_j B_j(x; q). \tag{2.3}$$

This creates a matrix of the B-spline basis function which describes how each of the n basis functions change with x. This method therefore takes a linear combination of the weighted averages of the basis functions at intervals of x as the covariates in the regression. The main disadvantage with this technique is that the regression coefficient estimates have no direct interpretation, however the plotted curves are generally able to fully capture the relationship between the explanatory and response variables.

The **splines** package in **R** can be used to implement regression splines, and in particular the **bs** function is used for fitting curves using a B-spline basis for a polynomial of any order.

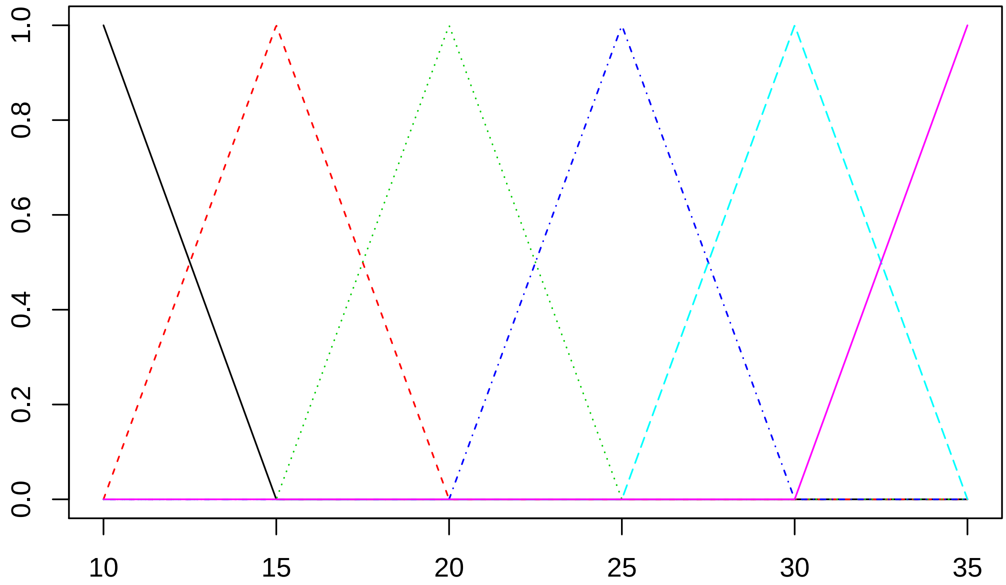
Figures 2.2a and 2.2b illustrate the differences between quadratic and cubic B-spline basis functions with six equally spaced inner knots. Figure 2.2a shows six B-splines of degree 1, each one based on three knots and Figure 2.2b illustrates five B-splines of degree 2, each based on four adjacent knots.

Figures 2.3 and 2.3b illustrate smooth curves fitted by B-splines of quadratic degree composed with varying numbers of quantile and equally spaced knots, respectively, applied to the Fourth Dutch Study data described in section 1.4. This regression spline smoothing method appears to perform well on the data, with indications that when there is a smooth pattern in the data, as illustrated here, low numbers of knots are adequate. This is because more knots lead to a larger amount of flexibility in the curves fitted which can lead to overfitting if the true pattern in the data is relatively smooth. The curves composed from equally spaced knots show more fluctuations in the curves, particularly in age intervals with fewer observations.

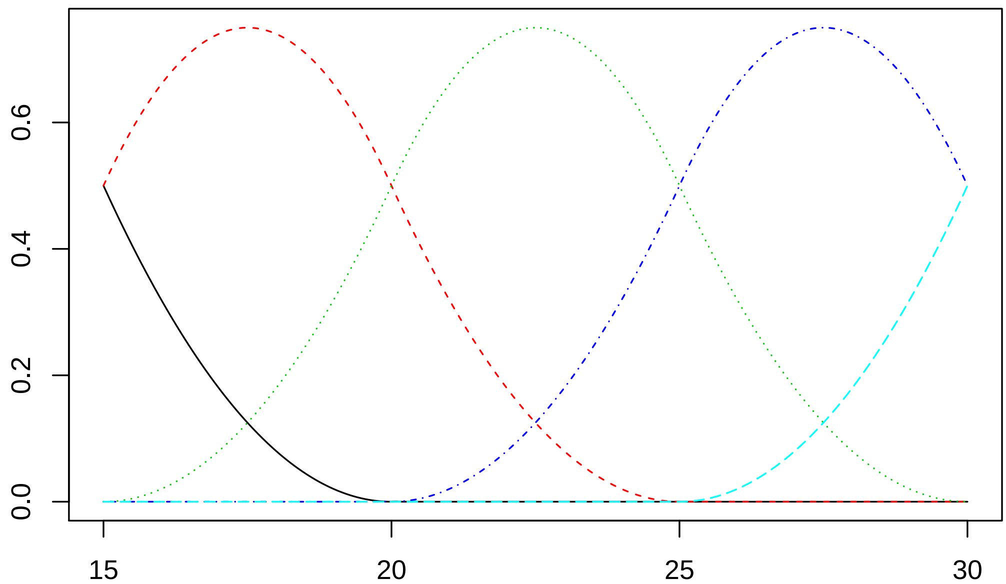
Increasing the degree of the B-spline as shown in Figure 2.4 improves the flexibility of the curve, with only minor differences visible between the curves produced by quadratic and cubic B-splines, the most common degrees of B-splines.

2.3 Penalised Regression Splines

An alternative to regression splines is to control the smoothness by using a relatively large number of knots but to prevent overfitting of the function by adding a penalty to the least square objective function which restricts the flexibility of the fitted curve. This is achieved by P-splines (Eilers and Marx, 1996). P-splines uses a B-spline basis, usually defined on evenly spaced knots, with a different penalty applied directly to the estimated coefficients of the fitted curve, with B_j controlling the smoothness of the function. P-splines have no boundary effects, are a straightforward extension of linear regression models, conserve moments of the data and have polynomial curve fits as limits (Eilers and Marx, 1996). Consider the regression of m data points (x_i, y_i) on a set of n B-splines $B_j(.)$. The least square objective function to

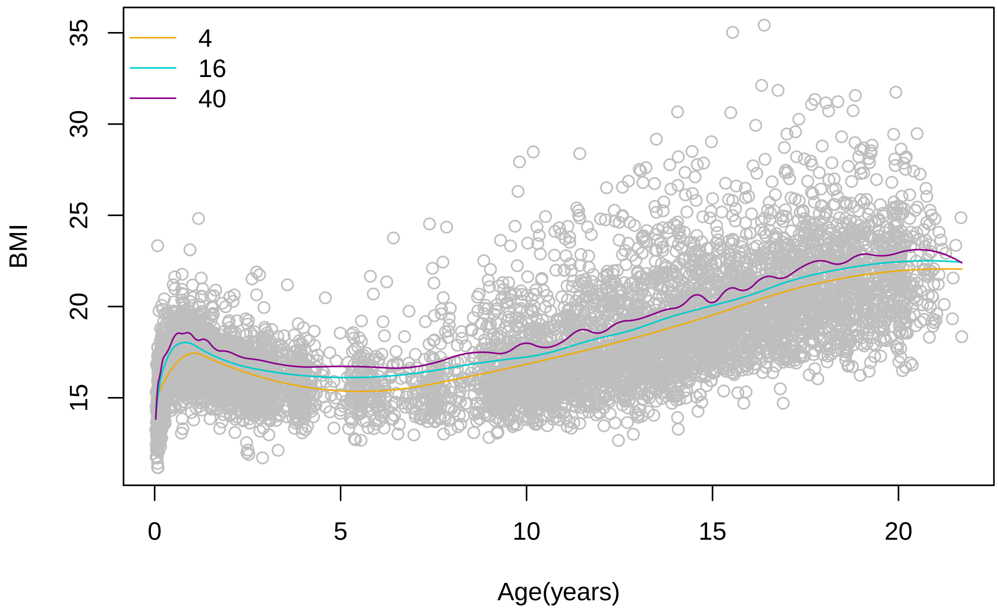


(a) Quadratic B-spline Basis function

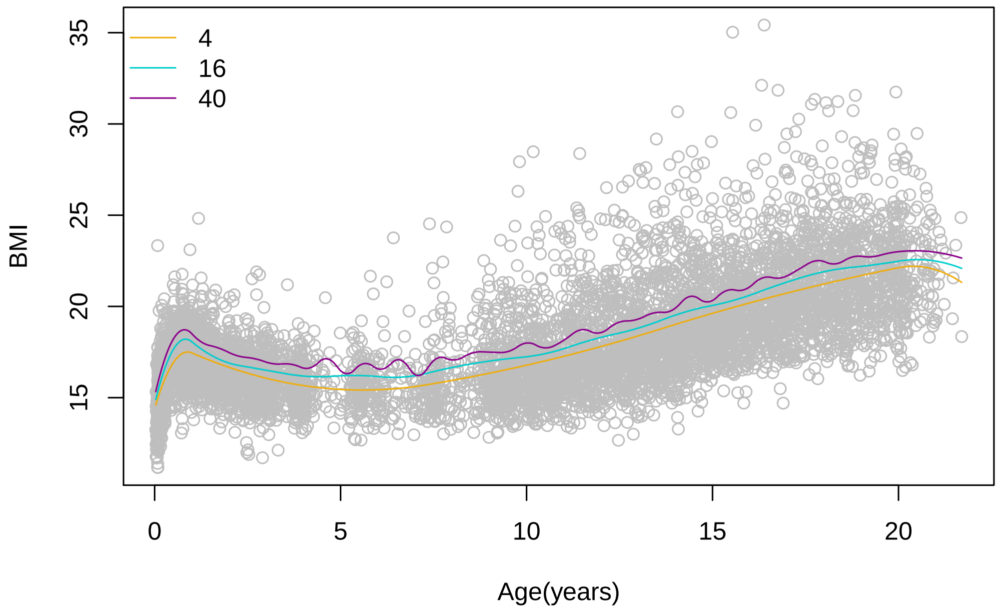


(b) Cubic B-spline Basis function

Figure 2.2: Quadratic and cubic B-spline basis functions shown for the interior knot sequence {10, 15, 20, 25, 30, 35}



(a) Quadratic B-splines with quantile knots



(b) Quadratic B-splines with equally spaced knots

Figure 2.3: BMI of the 7482 male participants in the Fourth Dutch Growth Study between birth and 21 years of age. Superimposed are smooth curves fitted by quadratic B-splines with varying number of quantile and equally spaced knots. For clarity the curves are offset from each other other by 0.5 BMI units.

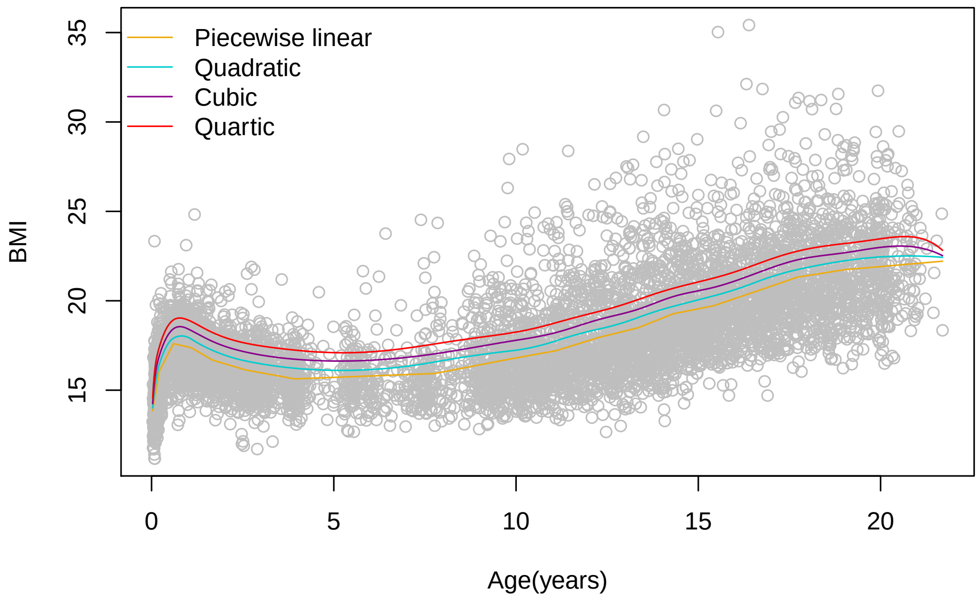


Figure 2.4: BMI of the 7482 male participants in the Fourth Dutch Growth Study between birth and 21 years of age. Superimposed are smooth curves fitted by B-splines of degree linear, quadratic and cubic, each with 16 quantile knots. For clarity the curves are offset from each other other by 0.5 BMI units.

minimise is

$$S = \sum_{i=1}^{m} \{y_i - \sum_{j=1}^{n} \alpha_j B_j(x_i)\}^2. \tag{2.4}$$

Suppose that the number of knots is relatively large, such that the fitted curves show more variation than is justified by the data. In order to make the resulting curve less flexible, Eilers and Marx (1996) proposed to base a penalty on finite differences of the coefficient of adjacent B-splines:

$$S = \sum_{i=1}^{m} \{y_i - \sum_{j=1}^{n} \alpha_j B_j(x_i)\}^2 + \lambda \sum_{j=k+1}^{n} \{\Delta^k \alpha_j\}^2 \tag{2.5}$$

where $\Delta\alpha_j = \alpha_j - \alpha_{j-1}$ and therefore for example $\Delta^2\alpha_j = \Delta\Delta\alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$. This reduces the dimensionality of the problem to n, the number of B-splines, instead of m, the number of observations. There is still the smoothing parameter λ which allows for continuous control over the smoothness of the fitted curve. For one approach on choosing the smoothness parameter see Green (1987). In practice the e.d.f is used to adjust the smoothness of the curves. The difference penalty is a good discrete approximation to the integrated square of the k-th derivative. P-splines allow a great deal of flexibility in that any order of penalty can be combined with any order of B-spline basis. The penalties are less easily interpreted than the usual spline penalties and if uneven knot spacing is required then the advantage of B-splines is lost (Wood, 2006).

The **gam** function in Wood's (2006) **mgcv** package, which is available in **R**, is one of many **R** packages that can be used to implement penalised regression splines.

Figure 2.5 illustrates the use of penalised regression splines on the Fourth Dutch Growth Study data. Adjusting the difference penalty and the degree of the P-spline only has a minor influence on the curves produced. These minor differences, also reflected when smoothing using natural cubic splines and B-splines with different smoothing choices i.e. changing the number of knots or degree of B-spline, are primarily due to the large sample size of this

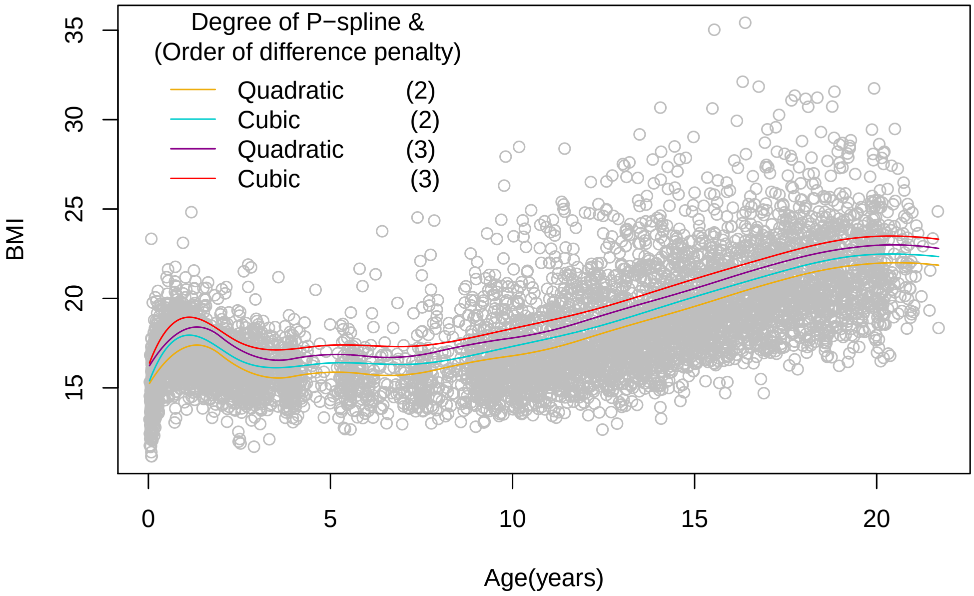


Figure 2.5: BMI of the 7482 male participants in the Fourth Dutch Growth Study between birth and 21 years of age. Superimposed are smooth curves fitted by P-splines of quadratic and cubic degree with second and third order difference penalties, each with 16 quantile knots and a smoothing parameter value of 1.2. For clarity the curves are offset from each other other by 0.5 BMI units.

study. In cases where a smaller sample size is available, the smooth curves will become a lot more sensitive to changes in these smoothing choices.

2.4 Monotonicity Constraints on Splines

Monotonicity constraints on splines as discussed in detail in Meyer (2012) and Bollaerts et al. (2006), constrain the function $f(x)$ to be either monotone increasing or decreasing, as well as smooth. In the case of growth measurements, such as height, it is common to assume that $f(x)$ is monotone increasing. For monotonically constrained P-spline regression, an asymmetric discrete penalty on the k-th order difference is used. This penalty is asymmetric since it differentially penalizes positive and negative kth-order differences, in order to restrict the sign of the k-th order differences and as such restrict the sign of the kth-order derivative of the fitted function. A penalty reflecting the constraints of a positive kth-order derivative within a range as defined by indicator variable v_j is

$$\sum_{j=k+1}^{n} v_j w(\alpha)_j (\Delta^k \alpha_j)^2 \tag{2.6}$$

with

$$v_j = \begin{cases} 1, & \text{if the constraint on } \partial^k f(x)/\partial x^k \text{ is to hold on at least part of the support of } B_j; \\ 0, & \text{otherwise,} \end{cases} \tag{2.7}$$

and with

$$w(\alpha)_j = \begin{cases} 0, & \text{if } \Delta^k_{\alpha_j} \geq 0,; \\ 1, & \text{otherwise},. \end{cases} \tag{2.8}$$

being asymmetric weights. Then, with κ being a user-defined constraint parameter, the overall loss function is:

$$S = \sum_{i=1}^{m} \{y_i - \sum_{j=1}^{n} \alpha_j B_j(x_i)\}^2 + \lambda \sum_{j=k+1}^{n} \{\Delta^k \alpha_j\}^2 + \kappa \sum_{k+1}^{n} v_j w(\alpha)_j (\Delta^n \alpha_j)^2. \tag{2.9}$$

Monotonicity constraints can also be applied to B-splines. A sufficient condition to ensure that the fitted curves are strictly non decreasing is to guarantee that the first-order difference of adjacent coefficients are non-negative (Muggeo et al., 2012). This means ensuring that $\hat{\alpha}_{k,j+1} - \hat{\alpha}_{k,j} \leq 0$ for $j = 1, \cdots, n-1$. There are standard linear inequality constraints which must be considered and discussions of these are found in Muggeo et al. (2012).

The **gcrq** function in the **quantregGrowth** package (Muggeo et al., 2012), which is available in **R**, can be used to implement monotonicity constrained regression and penalised regression splines.

Although applying P-splines with a monotonicity constraint to the Fourth

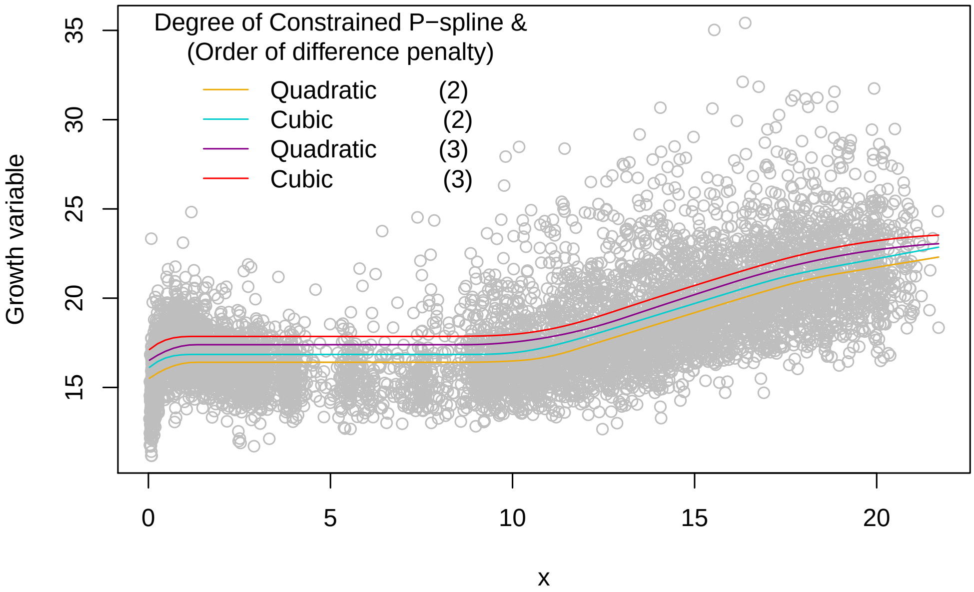


Figure 2.6: BMI of the 7482 male participants in the Fourth Dutch Growth Study between birth and 21 years of age. Superimposed are smooth curves fitted by non-decreasing P-splines of quadratic and cubic degree with second and third order difference penalties, each with 16 quantile knots and a smoothing parameter λ value of 1.2. For clarity the curves are offset from each other other by 0.5 BMI units.

Dutch Growth Study is clearly unsuitable, Figure 2.6 illustrates the effectiveness of such a constraint on the curves fitted.

The methods discussed in this chapter are used in Chapters 3, 4 and 5 to estimate growth curves.

Chapter 3

The LMS Method for Growth Curve Estimation

The first method of growth curve estimation to be explored in this thesis is the LMS method.

The LMS method provides an approach which conditions on a time covariate, often age, to obtain reference centile curves for growth data, showing how the distribution of a growth measurement changes according to this time covariate. This method is more appropriate than a simple reference range, as the measurements of growth are typically highly dependent on a time covariate (Cole and Green, 1992). This approach assumes that the growth measurement has an underlying skew distribution which can be normalised by applying a suitable power transformation.

The LMS method describes how the distribution changes by three uncorrelated curves, the L, M and S curves. These three curves respectively illustrate the skewness of the distribution and the trends in the median and standard deviation at each value of the time covariate.

3.1 LMS Model Methodology

The response growth variable (for example weight), denoted by y, is assumed positive. The LMS method by Cole and Green (1992) applies a Box-Cox power transformation (Box and Cox, 1964) to the data to obtain normality, determining the optimal power estimate of λ to do so. This is because most commonly used measures of growth, with the exception of height, have data that are usually more skew than the normal distribution (Cole, 1988). The data is then standardised so that the percentiles can be read off the normal distribution. This is done by firstly dividing y by the median (y/μ) and then by dividing this Box-Cox transformed variable by the standard deviation σ, so that the data follows a standard normal distribution $N(0, 1)$.
Suppose that y has median μ, and that y^λ or in the case $\lambda = 0, \log_e(y)$ is normally distributed. It is then appropriate to consider the transformed variable,

$$x = \frac{(y/\mu)^\lambda - 1}{\lambda}, \qquad \lambda \neq 0 \tag{3.1}$$

or

$$x = \log_e(y/\mu), \qquad \lambda = 0$$

based on the family of transformations proposed by Box and Cox (1964). This transformation means that x is zero when the value of y and the median μ are equal, and is continuous at λ=0. For λ=1 the standard deviation (SD) of x is exactly the coefficient of variation (CV) of y, and this remains approximately true for all moderate λ (Cole and Green, 1992). The optimal value of λ is that which minimizes the SD of x.
Denoting the SD of x (and the CV of y) by σ, the *SD score* of x and hence of y is given by

$$z = x/\sigma$$

$$z = \frac{(y/\mu)^\lambda - 1}{\lambda\sigma}, \qquad \lambda \neq 0 \tag{3.2}$$

or

$$z = \frac{\log(y/\mu)}{\sigma}, \qquad \lambda = 0$$

and it is assumed that z has a standard normal distribution.

Assume now that the distribution of y varies with time covariate t, and that λ, μ and σ at t are read off the smooth curves $L(t)$, $M(t)$ and $S(t)$. It follows that

$$z = \frac{(y/M(t))^{L(t)} - 1}{L(t)S(t)}, \qquad \lambda \neq 0 \tag{3.3}$$

or

$$z = \frac{\log(y/M(t))}{S(t)}, \qquad \lambda = 0$$

In general if y has a cumulative distribution function $F_Y(y) = P(Y \leq y)$. The 100τth centile $C_Y(\tau)$ or τth quantile $Q_Y(\tau)$ of Y is defined as

$$Q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$$

where $\tau \in [0, 1]$.

The τth quantile of y at x, denoted by $Q_Y(\tau|X)$, separates the reference population into parts. For example, for $\tau = 0.05$, $Q_Y(0.05|X)$ is the 5th centile of the distribution of y conditional on the value of X. In other words, 5 percent of the values of y are less than or equal to the specified function of X.

Rearranging Equation (3.3) shows that the 100τth centile of y at t is given by

$$C_Y(\tau|t) = M(t)(1 + L(t)S(t)Z_\tau)^{\frac{1}{L(t)}}, \qquad \lambda \neq 0 \tag{3.4}$$

Or

$$C_Y(\tau|t) = M(t)\exp[S(t)Z_\tau], \qquad \lambda = 0$$

for the LMS method, where Z_τ is the normal equivalent deviate (NED) of size τ. This shows that if the L, M and S curves are smooth, then so are the centile curves (Cole and Green, 1992).

The initial LMS method developed by Cole (1988) requires the data to be divided into distinct groups by the time covariate, which should ideally be as narrow as practicable, given the available data to obtain the L, M and S curves. The optimal power transformation λ to obtain normality is calculated for each of the series of distinct groups. For each group, λ is estimated by initially finding the standard deviation of the natural log, inverse and original growth measurement which correspond to λ values of 1, 0 and -1 in $(y^\lambda$ (or $\lambda = 0, \log_e(y))$. The geometric mean of the growth measurement is also computed as

$$\mu_G = \exp(\frac{1}{n}\Sigma(\log(y_i))). \tag{3.5}$$

The standard deviation of the original growth measurement is divided by the geometric mean to give a form of the arithmetic coefficient of variation (S_a). The standard deviation of the inverse growth measurement is then multiplied by the geometric mean to give the harmonic coefficient of variation (S_h). The standard deviation of the natural log transformation gives the geometric coefficient of variation (S_g). The estimate of λ is given by

$$\lambda = \frac{\log(S_h/S_a)}{2\log(S_h S_a/S_g^2)}. \tag{3.6}$$

This process is repeated for each distinct group and the resulting λ values are plotted against the time covariate value, with the trend then smoothed across the time covariate to give the L curve.

The median and coefficient of variation have to be similarly estimated for each group. Firstly the growth measurements are raised to the optimal power λ, found for that distinct group. The median and standard deviation are then determined for the λ transformed growth measurement. The standard deviation is divided both by λ and the geometric mean raised to the power λ and if this result is negative, the absolute value is taken. The median of the transformed growth measurement is back-transformed by raising it to the power $1/\lambda$ (Cole, 1988). The median and coefficient of variation for each distinct group are plotted on separate plots and then the group values are smoothed across the time covariate to produce the M and S curves, respectively.

Initially Cole (1988) suggested drawing the smooth curves L,M and S by eye, but subsequent researchers proposed smoothing methods such as natural cubic splines. See chapter 2 for details of such smoothing methods. Green (1988) highlighted the subjective and complex nature of Cole's fitted algorithm for the LMS method to deduce the L, M and S curves, in particular the requirement of separating the data into distinct groups which is highly arbitrary. He alternatively suggested the method of maximum penalised likelihood (Green, 1987) which can be used to provide smooth estimates of the L, M and S curves directly, requiring only the choice of smoothing constants for the three curves, as detailed in Cole and Green (1992).

For the case of n independent observations y_i at corresponding time covariate values t_i, the log-likelihood function ℓ derived from Equation (3.3) is given (apart from the constant) by

$$\ell = \ell(L, M, S) = \sum_{i=1}^{n} (L(t_i) \log \frac{y_i}{M(t_i)} - \log S(t_i) - \frac{1}{2} z_i^2) \tag{3.7}$$

where z_i are the SD scores corresponding to y_i. The curves $L(t)$, $M(t)$, $S(t)$ are estimated by maximizing the penalised likelihood

$$\ell - \frac{1}{2}\tau_\lambda \int \{L''(t)\}^2 dt - \frac{1}{2}\tau_\mu \int \{M''(t)\}^2 dt - \frac{1}{2}\tau_\sigma \int \{S''(t)\}^2 dt. \tag{3.8}$$

where τ_λ , τ_μ and τ_σ are smoothing parameters. The three integrals provide roughness according to the squared second derivatives of the L, M and S curves, so that maximizing Equation (3.9) strikes a balance between the fit to the data and smoothness of the L, M and S curves. As discussed in detail in section 2.1, it can be shown that these forms of penalty lead to natural cubic splines with knots at each distinct value of the covariate t. Thus only the smoothing parameters τ_λ , τ_μ and τ_σ need to be chosen in order to fit the model, as each of the distinct time covariate values will be taken as knots.

Once the L, M and S curves have been constructed they can then be used to calculate the required normalised growth chart. Firstly it must be decided which particular centiles are required. For growth curves these are commonly the 5th, 10th, 25th, 50th, 75th, 90th, 95th centiles, and the value of the corresponding Z_τ can be read off the standard normal distribution, as explained earlier on page 32. Also the number of time points at which the smooth estimates of the L, M and S curves should be extracted for the centile curve plotting should be decided. These time points will then be evenly dispersed across the entire time covariate range. Then for each time covariate value (t) and SD score in turn, values for $L(t)$, $M(t)$ and $S(t)$ can be read off their corresponding curves and substituted into Equation (3.4) which gives the 100τth centile of the growth measurement at time covariate value t. This is used to construct an entire growth chart, with more time points required if the reference centile curves are jagged in appearance.

The **lmsqreg.fit** function in Carey's (2002) **lmsqreg** package in **R** can be used to implement this LMS model approach of fitting growth charts, with the complexity of each of the three fitted cubic spline curves, the L, M and S curves, determined by its specified equivalent degrees of freedom (e.d.f). This package uses a default of 50 age points for the centile curve plotting.

An alternative approach for obtaining growth curves using the LMS model is to use the **lms** function in Stasinopoulos and Rigby's (2007) **GAMLSS** package in **R**. This approach uses a generalized additive model for location, scale and shape **GAMLSS** to construct the growth charts.Rigby and Stasinopoulos (2005) defined the original formulation of the **GAMLSS** model which was later expanded to overcome some of the limitations associated with the popular generalised linear models (GLMs) (McCullagh and Nelder, 1989), (Hardin and Hilbe, 2007) and generalised additive models (GAMs) (Hastie and Tibshirani, 1990). GAMLSS model methodology extends GLMs by allowing for generalised additive model methodology to be used. The GAM approach assumes the response variable y follows an exponential family distribution however this assumption is relaxed and replaced by a general distribution family when the GAMLSS approach is used. In GAMLSS the systematic part of the model is expanded to allow modelling not only of the mean (or location), as in the case when using the GAM approach, but of other parameters of the distribution of y also (Stasinopoulos and Rigby, 2007).

A generalised additive model(GAM) is a generalised linear model in which part of the linear predictor is specified in terms of a sum of smooth functions of explanatory variables (Wood, 2006). In general the model has a structure something like

$$g(\mu_i) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \cdots . \tag{3.9}$$

where $\mu_i = E(Y_i)$ and Y_i are i.i.d response variables which follow some exponential family distribution. X_i^* is a row of the model matrix for any strictly parametric model components, θ is the corresponding parameter vector, and the f_j are smooth functions of the explanatory variables x_k. Detailed discussion of how to represent the smoothing functions as well as determining the appropriate degree of smoothness for these functions of the covariate variable is given in Wood (2006).

GAMLSS are semi-parametric univariate regression models, where a parametric distribution assumption is required for the response variable. They are therefore a generalization of the previous LMS modelling approach where the response variable has a specified frequency distribution D(μ,σ,λ,α), where the parameters represent the first four moments of the distribution (Stanojevic et al., 2009). A wide variety of distributional forms are available, of which the Box-Cox Cole and Green (BCCG) distribution is typically selected when modelling growth reference curves. This distribution has location (μ), scale (σ) and shape (λ) distribution parameters. Therefore, essentially when choosing the BCCG distribution for the response variable, the Cole and Green (1992) method is being used for centile estimation. Other distributions which are commonly selected when modelling such data include the Box-Cox exponential power (BCPE) and Box-Cox t (BCT) distribution, with the latter distribution being appropriate when there are heavy tails in the response variable. These distributions are an extension of the BCCG distribution and include an additional shape parameter α.

The GAMLSS model assumes independent observations y_i for $i = 1, 2, \dots, m$ with probability (density) function $f(y_i|\Theta^i)$ conditional on $\Theta^i = (\Theta_{1i}, \Theta_{2i}, \Theta_{3i}, \Theta_{4i}) = (\mu_i, \sigma_i, \lambda_i, \alpha_i)$, a vector of four distribution parameters, each of which can be a function of the covariate and/or random effects terms. Let $y^T = (y_1, y_2, y_3, \dots, y_m)$ be the m length vector of the response variables. For $k = 1, 2, 3, 4$, let $g_k(\cdot)$ be known monotonic link functions relating the distribution parameters to explanatory variables by,

$$g_k(\Theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} h_{jk}(x_{jk}),$$

i.e.

$$g_1(\mu) = \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} h_{j1}(x_{j1})$$

$$g_2(\sigma) = \eta_2 = X_2\beta_2 + \sum_{j=1}^{J_2} h_{j2}(x_{j2})$$

$$g_3(\lambda) = \eta_3 = X_3\beta_3 + \sum_{j=1}^{J_3} h_{j3}(x_{j3})$$

$$g_4(\alpha) = \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} h_{j4}(x_{j4})$$

where μ, σ, λ, α and η_k are vectors of length m. $\beta_k^T = (\beta_{1K}, \beta_{2K}, \cdots, \beta_{J_k^{'}K})$ is a parameter vector of length $J_K^{'}$. X_k is a fixed known design matrix of order $n \times J_K^{'}$ and h_{jk} is a smooth non-parametric function of explanatory variable x_{jk}, $j = 1, \cdots, J_K$ and $k = 1, \cdots, 4$ (Stasinopoulos and Rigby, 2007).

The **lms** function in the **GAMLSS** package, models the distribution parameters λ, μ, σ and α as additive non-parametric smoothing functions of the time covariate x, using P-splines for smoothing. This is a smoothing method discussed in detail in section 2.3. The **lms** function requires the e.d.f of each of the smooth non-parametric P-spline functions to be specified, which adjusts the smoothness of the curves. In the case where the response variable has a specified Box-Cox Cole and Green (BCCG) distribution this means that only three e.d.f's have to be chosen. This is effectively similar to specifying the e.d.f in the **lmsqreg.fit** function, which as previously discussed controls the smoothness of the L, M and S curves, respectively. The **lms** function gives more flexibility in the smoothing options than the **lmsqreg.fit** function, by fitting using P-splines. It by default adopts P-splines of cubic degree with 2nd order difference penalty with 20 equally spaced knots.

The LMS growth curves, produced by both modelling approaches, can be used to assess a measurement of growth for any individual by expressing the measurements in centile or SD score terms (Cole, 1990). If G is the individual's growth measurement at time covariate value t then their SD score is defined as

$$SDS = \frac{(G/M(t))^{L(t)} - 1}{L(t)S(t)} \tag{3.10}$$

where $M(t)$, $L(t)$ and $S(t)$ are the values read off the curves at age t.

3.2 LMS Model for the Growth and Development Study Data

The Growth and Development Study Data, described in section 1.1, was used to illustrate the LMS method for composing gender-specific weight growth charts, firstly using the **lmsqreg.fit** function in the **lmsqreg** package and then the **lms** function in the **GAMLSS** package, which are both available in **R**. Similar plots to those in Cole and Green (1992) were used to explore the effect of varying the e.d.f values on the smoothness of the L, M and S curves when the **lmsqreg** package was implemented.

Figure 3.1 shows the L, M and S curves for the weight of full-term female infants from birth to roughly 36 months of age, over a range of fitted e.d.f values between 2.5 and 8. In each case the smoothest curves with 2.5 e.d.f, appear to provide a poor fit to the data. As the e.d.f increases the curves become more complex, and generally those with between 3 and 4 e.d.f seem to produce sensible curves, with a large similarity between there curves. Conversely with 8 e.d.f the curves appear undersmoothed. L, M and S curves smoothed by natural cubic splines with between 3 and 4 e.d.f therefore appear to be a sensible choice.

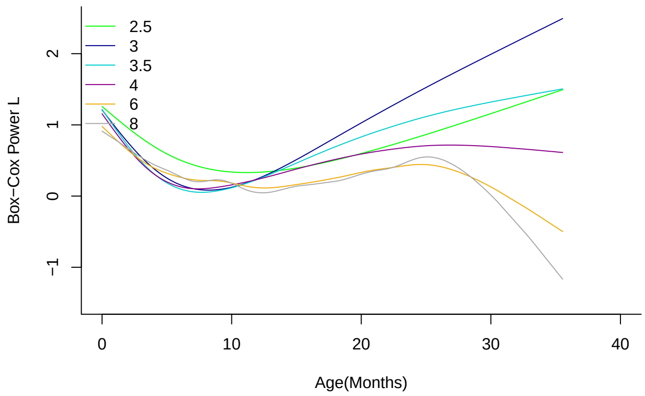
The L curves (Figure 3.1a) fitted with these e.d.f values all illustrate a rapid reduction in the Box-Cox power of the distribution from birth till around 8 months of age, indicating an increase in skewness to the right. The Box-Cox power then begins to steadily increase after this dip, in each case. The M curves (Figure 3.1b) all demonstrate that the median weight of full-term female infants continually rises as age increases, however the rate of growth reduces as age increase. The S curves (Figure 3.1c) fitted with between 3 and 4 e.d.f show a substantial dip in the coefficient of variation between birth and the age of 36 months, with the smallest level of variability occurring at around 10 months of age.

Test	P-value
Kolmogorov-Smirnov Test	0.252
Student T-test	0.695
Chi-Squared Test	0.895

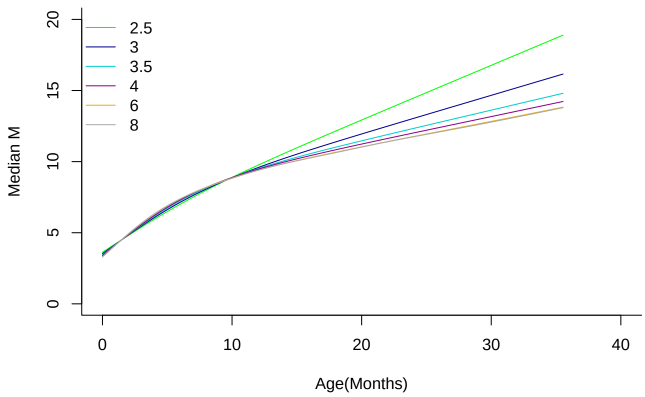
Table 3.1: Goodness of fit tests P-values, showing the performance of the weight growth chart for full-term females infants, constructed by LMS Model 1.

The growth charts shown in Figures 3.2a and 3.2b show the seven reference centile curves, for the 5th, 10th, 25th, 50th, 75th ,90th and 95th centile, for the weight measurements of full-term female infants, obtained from the L, M and S curves using Equation (3.4). In Figure 3.2a the curves have been smoothed by natural cubic spline each with 3 e.d.f, while in Figure 3.2b the e.d.f was equal to 4. The growth chart shown in Figure 3.2b is well smoothed over the whole age range and is able to capture the trend in weights of full-term female infants from birth till roughly 36 months. In contrast, the growth chart in Figure 3.2a exhibits a considerable fall in the growth rate in the lower centiles which may not reflect reality as curves would be expected to be non-decreasing. Thus fitting a weight growth chart for full-term female infants based on the **lmsqreg** package with L, M and S curves smoothed by natural cubic spline curves each with 4 e.d.f, which will be referred to as LMS Model 1 for the remainder of this thesis, seems sensible.

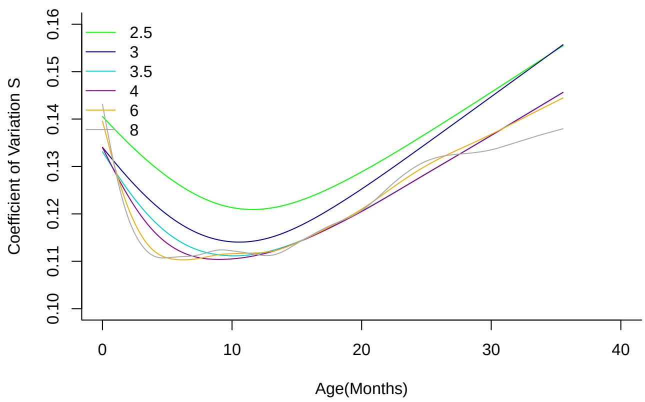
The results in Table 3.1 demonstrate that LMS Model 1 is a good fit to the full-term female infant data, with relatively high p-value of 0.252, 0.695 and 0.895 being obtained for the Kolmogorov-Smirnov, Student T and Chi-Squared goodness-of fit tests, respectively. This indicates, as desired, an overall good agreement between the SD scores from the fitted LMS model and a standard normal distribution.



(a) L Curve

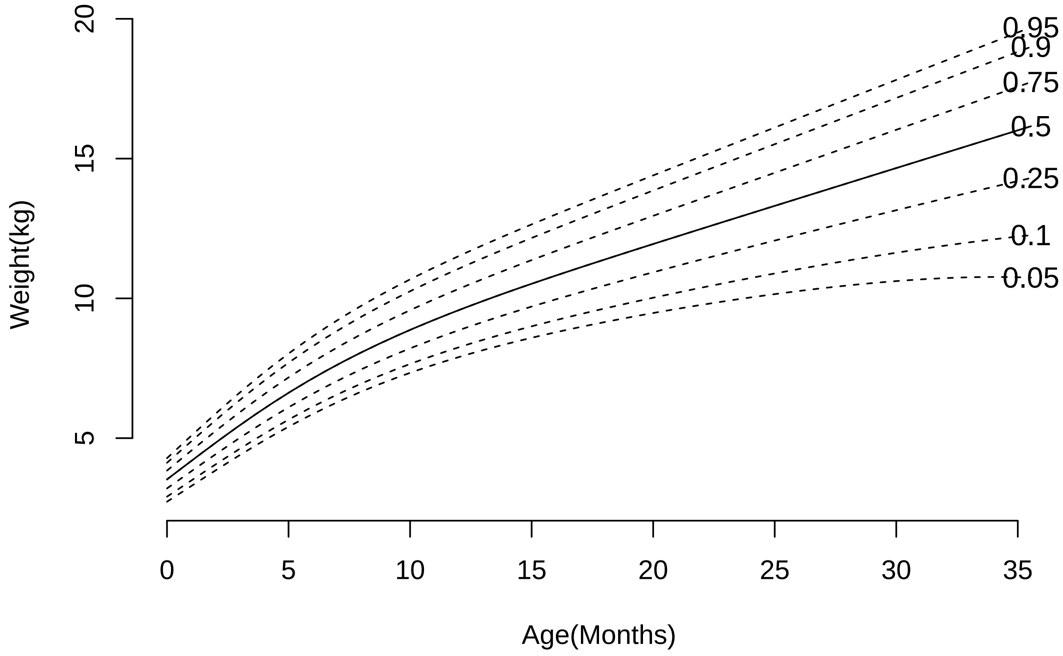


(b) M Curve

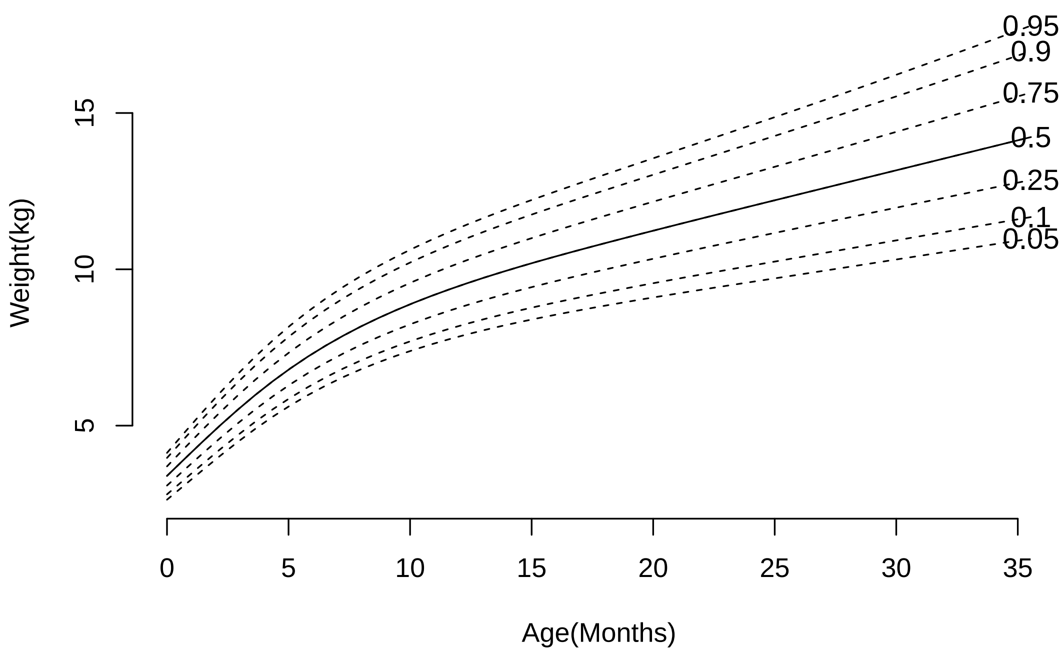


(c) S Curve

Figure 3.1: L, M and S curves for weight in full-term female infants from birth to roughly 36 months of age, fitted by a series of natural cubic spline curves with between 2.5 and 8 e.d.f.



(a) 3 e.d.f



(b) 4 e.d.f

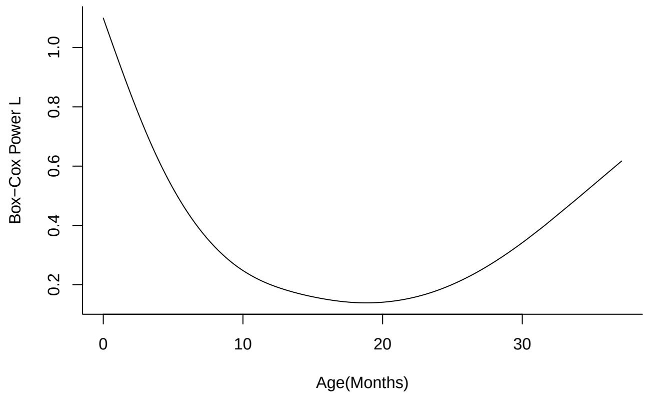
Figure 3.2: Weight growth chart for full-term female infants from birth to 36 months of age, based on the L, M and S curves in Figure 3.1.

A similar selection process was used to identify a suitable choice of equivalent degrees of freedom for the smoothness of the L, M and S curves for the full-term male infant data.

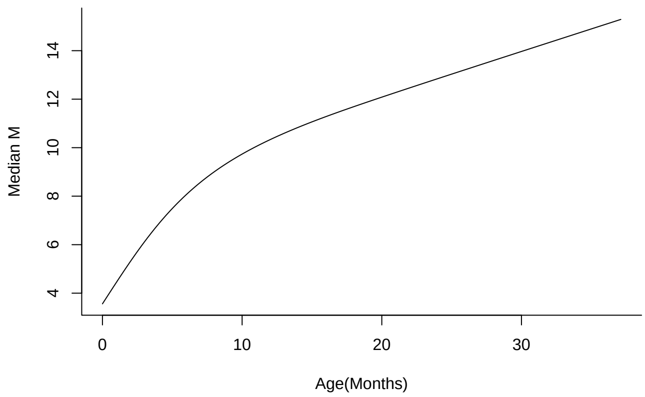
Figure 3.4 shows the fitted L, M and S curves for full-term male infants' weights between birth and roughly 37 months of age, obtained by setting the e.d.f to 3, 4 and 5, respectively. In general the trends are consistent with those observed for the full-term infants with for instance the L curve (Figure 3.3a) illustrating a progressive reduction in the Box-Cox power distribution, between birth till around the age of 20 months, before it steadily increases after the dip. This is compatible with the trend seen for full-term female infants and again indicates an increase in skewness to the right of the data. The observed trend for median weight measurements for full-term male infants, as seen in the M curve (Figure 3.3b) is also similar to that observed for the full-term females infants, however the initial rate of growth is slightly higher. The S curve (Figure 3.3c) for the full-term male infants, also exhibits a dip in the coefficient of variation between birth and around 37 months of age, with it being constantly low between the ages of 7 and 17 months.

Figure 3.4 gives the corresponding set of seven reference centile curves, from the 5th to the 95th, obtained from the L, M and S curves in Figure 3.3. This illustrates that a sensible weight growth curve estimate for full-term males infants is composed using the **lmsqreg** package with the L, M and S curves smoothed by natural cubic spline curves with respective e.d.f values 3, 4 and 5. This will be referred to as LMS Model 2 for the remainder of the thesis.

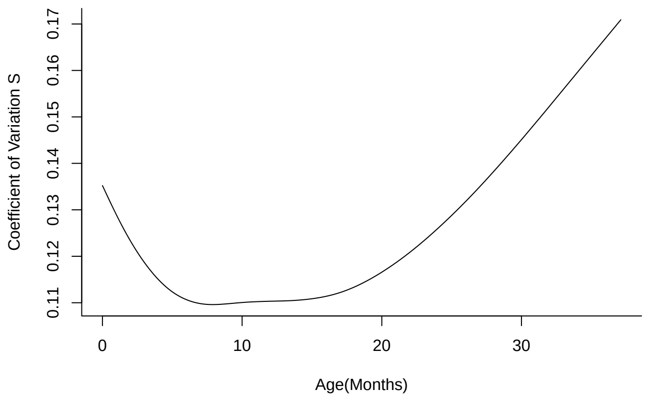
The results in Table 3.2 demonstrate that LMS Model 2 is a relatively good fit to the full-term male infant data, indicating an overall good agreement between the SD scores from the fitted LMS model and a standard normal distribution, according to a majority of the three tests. The Kolmogorov-Smirnov Test result is the exception, as it obtains small p-values for a major-



(a) Box-Cox power (L curve) fitted by a spline curve with 3 e.d.f.



(b) Median (M curve) fitted by a spline curve with 4 e.d.f.



(c) Coefficient of variation (S Curve) fitted by a spline curve with 5 e.d.f.

Figure 3.3: L, M and S curves for weight in full-term male infants from birth to roughly 37 months of age, fitted by natural cubic splines curves with 3,4,5 e.d.f's, respectively.

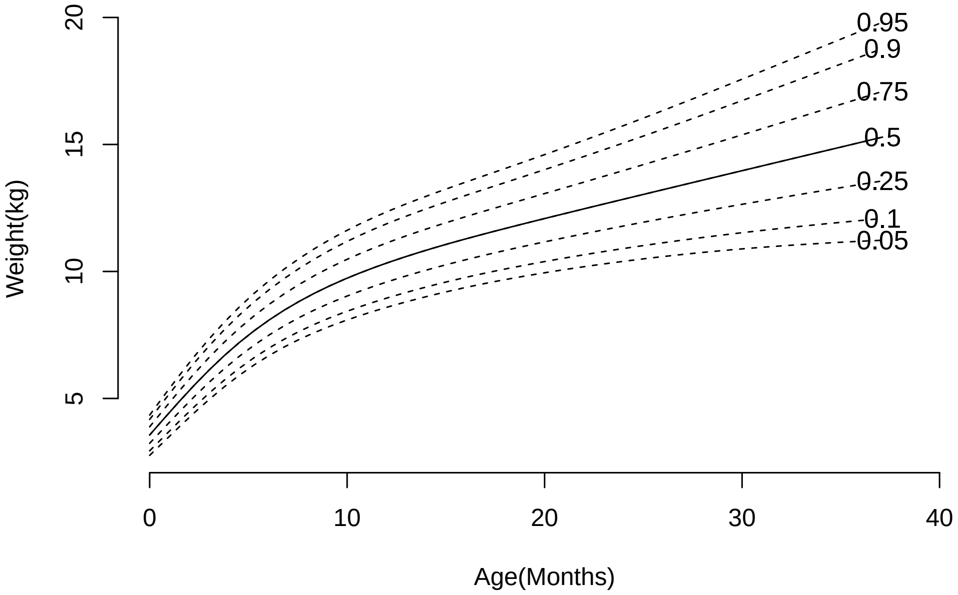


Figure 3.4: Weight growth chart for full-term male infants from birth to 37 months of age, based on the L, M and S curves in Figure 3.3.

ity of the age intervals as detailed in Table 3.3 and consequently an overall small p-value of 0.034 for the whole growth chart, which indicates lack of fit. However there is a strong agreement between the two other goodness of fit test approaches that the growth chart is a good fit, with overall p-values of 0.410 and 0.941 obtained for the Student T-test and Chi-Squared Tests respectively.

Test	P-value
Kolmogorov-Smirnov Test	0.034
Student T-test	0.410
Chi-Squared Test	0.941

Table 3.2: Goodness of fit tests for the growth charts for full-term males infants, constructed by LMS Model 2.

	Intervals in Age(Months)					
	0-1.31	1.31-3.08	3.08-5.53	5.53-10.1	10.1-37.1	Overall
Kolmogorov-Smirnov Test	<0.005	<0.005	<0.005	0.333	<0.005	0.034

Table 3.3: Kolmogorov-Smirnov Goodness of fit test for LMS Model 2, in different age intervals.

The LMS method was also used to compose gender-specific weight growth charts from the Growth and Development Study data using the **lms** function in the **GAMLSS** package. Given $X = x$, Y is modelled by a Box-Cox Cole and Green distribution, BCCG(μ, σ, λ), where the parameters μ, σ and λ are modelled as smooth non-parametric functions of x. The e.d.f's for the non-parametric P-spline functions must be specified. Identity link functions were chosen for μ and λ, while log link functions were chosen for the σ to ensure that $\sigma > 0$. The Akaike information criterion (AIC) was proposed as a modelling selection criterion to determine the most desirable choice of the e.d.f's. The AIC (Akaike, 1974) is a measure of the relative quality of a statistical model, where a small AIC value is optimal. The basic idea of AIC is to correct the log-likelihood of a fitted model for the effective number of parameters. The technique chosen to determine suitable e.d.f values was to initially find the most suitable e.d.f value for the M curve, by increasing and decreasing its value by 1 until the change in AIC is small. Next the e.d.f values for the S and L curves curves can be obtained by similar techniques, where in general smaller e.d.f values are preferred when the difference in AIC is not large.

As detailed in Table 3.4, although initially increasing the e.d.f value for the the M curve for the full-term female infant data reduces the AIC considerably, once it exceeds the value of 5 no big improvement is visible so thus this e.d.f value is fixed at 5. The e.d.f value for the S curves is also similarly fixed at 5. A larger e.d.f value for the L curve leads to lack of fit whereas alternatively reducing its value improves the AIC, so it seems sensible to choose a small e.d.f value of 1. Therefore a sensible weight growth chart for full-term female infants is produced using the LMS method with cubic P-splines with e.d.f values of 1, 5 and 5 for the L, M and S curves respectively. This will be referred to as LMS Model 3.

Similar techniques were applied to the full-term male infant data to choose the e.d.f's, as seen in Table 3.5. Although increasing the e.d.f values considerably reduces the AIC there is overwhelming evidence of overfitting in the growth charts produced. For example modelling with e.d.f values equal to 3, 4 and 3, respectively, as shown in Figure 3.5 exhibits a fall in the growth rate in the lower centiles, which is not consistent with the anticipated trend. Hence the LMS method fitted using cubic P-splines with e.d.f's 1, 3 and 2 for the L, M and S curves respectively, referred to as LMS Model 4, is found to be sensible.

Figures 3.6a and 3.6b show the weight growth charts for full-term female and male infants obtained from LMS Models 3 and 4, respectively. These growth charts both appear to have relatively smooth reference centile curves, capturing the key trends in weight of full-term infants of each gender.

Comparing the gender-specific weight growth charts for full-term infants produced by LMS Models 1 and 2 directly to those produced by LMS Models 3 and 4, Figure 3.7, shows that in fact the two approaches of smoothing for the L, M and S curves, natural cubic splines and P-splines, produce almost identical reference centile curves with only slight differences being visible at later months.

These growth curves can then be used to assess individual infants' weight measurements. As there is a large amount of agreement between the gender-

	e.d.f		
AIC	L	M	S
28364.3	3	3	3
28341.7	3	4	3
28337.4	3	5	3
28337.0	3	6	3
28327.6	3	5	4
28322.5	3	5	5
28320.0	3	5	6
28324.0	4	5	5
28320.9	2	5	5
28286.1	1	5	5

Table 3.4: AIC values for LMS models fitted with a series of P-spline curves with different e.d.f's for the L, M and S curves for weight in full-term female infants from birth to roughly 36 months of age.

	e.d.f		
AIC	L	M	S
29757.1	3	3	3
29712.1	3	4	3
29750.9	3	3	4
29772.4	3	3	2
29784.3	3	3	1.5
29770.8	2	3	2
29769.9	1.5	3	2
29768.9	1	3	2

Table 3.5: AIC values for LMS models fitted with a series of P-spline curves with different e.d.f's for the L, M and S curves for weight in full-term male infants from birth to roughly 37 months of age.

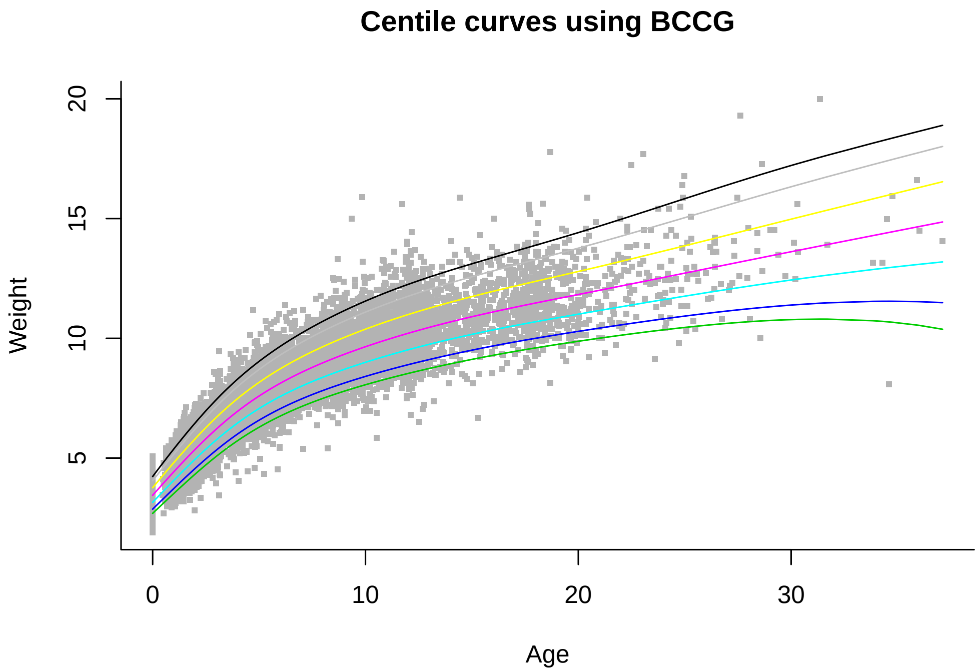
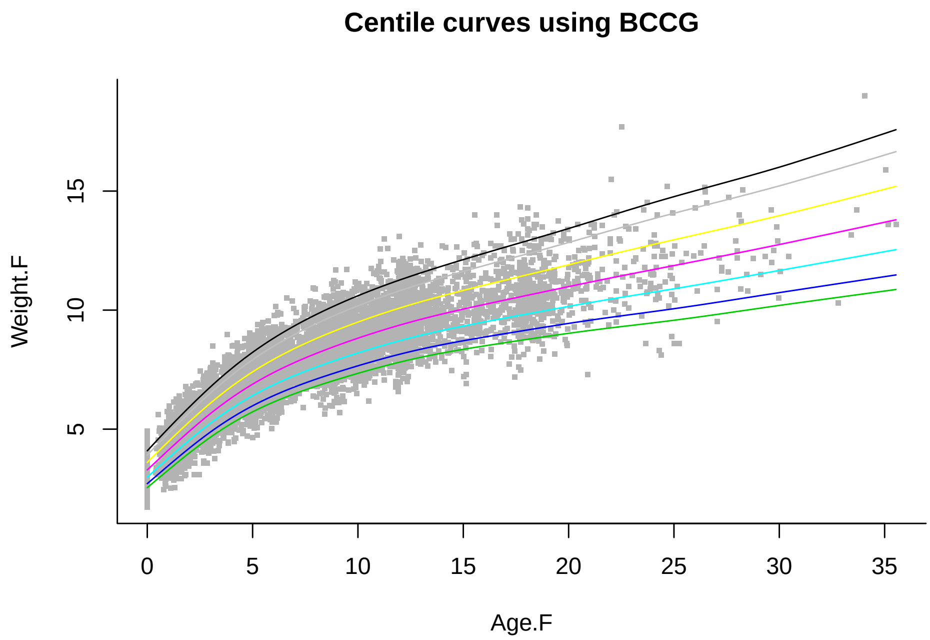
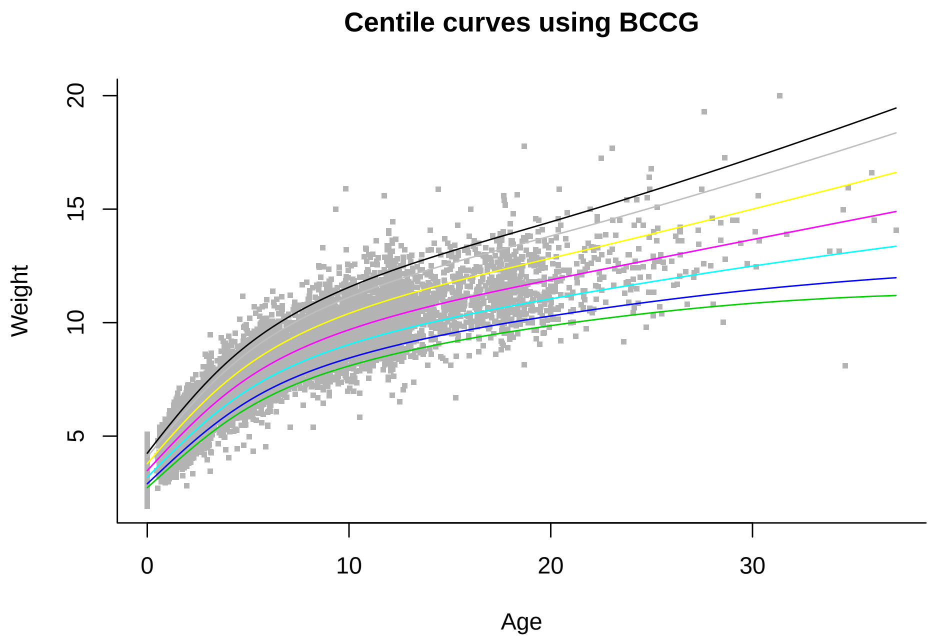


Figure 3.5: Weight growth chart for full-term male infants based on the L, M and S curves fitted with P-spline curves with 3, 4 and 3 e.d.f's ,respectively.

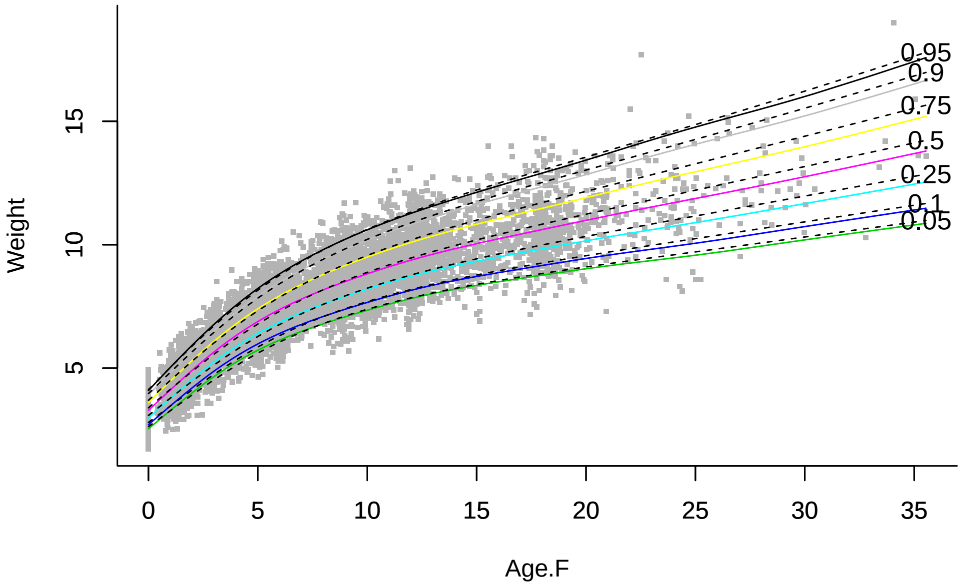


(a) Full-term female infants

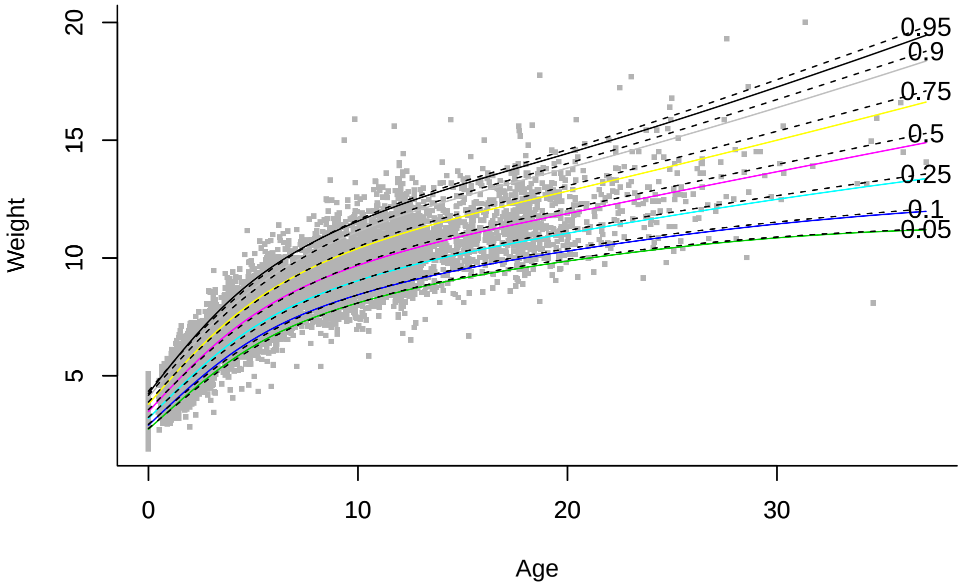


(b) Full-term male infants

Figure 3.6: Gender specific weight growth charts for full-term female and male infants, obtained from LMS Models 3 and 4, respectively



(a) Full-term female infants



(b) Full-term male infants

Figure 3.7: Comparison of gender-specific weight growth charts for full-term infants, fitted by LMS Models 1 and 2 to LMS Models 3 and 4. The solid lines represent the reference centile curves produced by the latter models.

specific weight growth charts for full-term infants produced by the two model approaches, screening will focus on the growth charts produced by LMS Models 1 and 2, as the results ascertained by the other model approach are likely to be very similar. Screening based on these growth charts was performed on the four case infants detailed in section 1.3, at the specified screening ages.

Figure 3.8a illustrates that Subject 1500 had considerably lower weight measurements than most of her peers, falling below the 5th centile curve at all ages. However her rate of growth appeared to remain consistent with the typical growth pattern. At roughly 20 months of age, the age at which the screening decision is considered, her weight measurement of 8.52 kg corresponds to an SD score of -2.111, as detailed in Table 3.6. This indicates that her weight value is below the median weight of full-term female infants of that age.

Subject 146's birth weight places her just above the 10th centile curve however she then displays rapid growth thereafter, with her observed weight at the age of roughly 8 months placing her just below the 75th centile curve. A sudden drop in growth rate is then observed after this point, with her falling just above the median weight by her final weight measurement; the age at which the screening decisions is considered. Her SD score at this age of 0.187, indicates that her weight value is above the median weight of full-term female infants of that age.

Figure 3.8b demonstrates that Subject 12 has a low birth weight falling between the 10th and 25 centile; he then showed an unusual growth pattern up to the age of 4.95 months, with weight continually increasing but with the rate of growth reducing dramatically as age increased. This led to his weight moving between the reference centile curves. A small fall in weight of 0.03 kg was then observed between the age of 4.95 and 5.87 months, the age at which the screening is considered, with him falling just above the 10th

Subject	Gender	Screening Age(Months)	SD Score
1500	Female	19.93	-2.111
146	Female	12.03	0.187
12	Male	5.87	-1.100
1799	Male	11.34	2.748

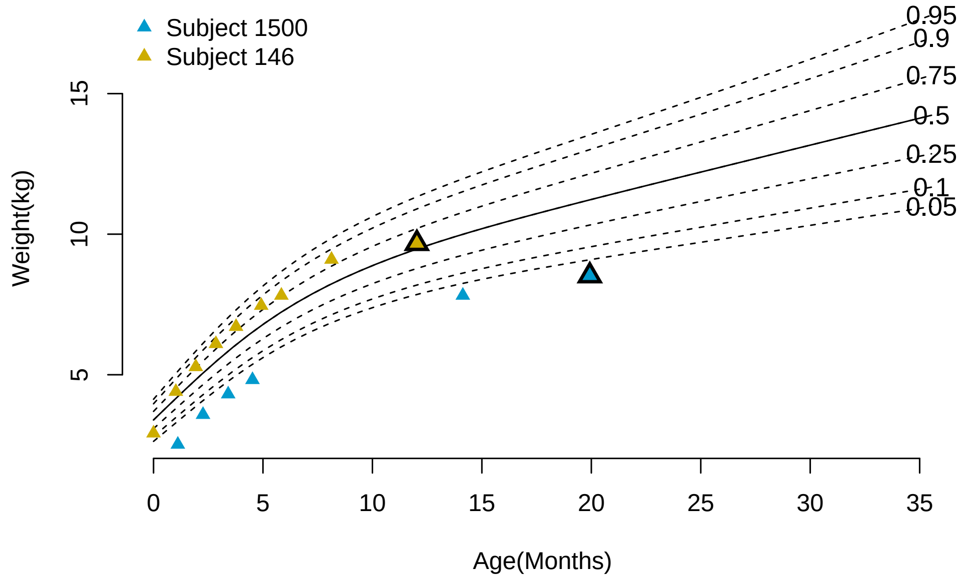
Table 3.6: LMS SD score estimates for the case infants at the specified screening ages, estimated from the gender-specific weight growth charts for full-term female and male infants produced by LMS Models 1 and 2, respectively.

reference centile curve at this age, after which he grew steadily. At roughly 6 months, his weight measurement of 7.05 kg is below the median weight of full-term male infants of that age, with an SD score of -1.100.

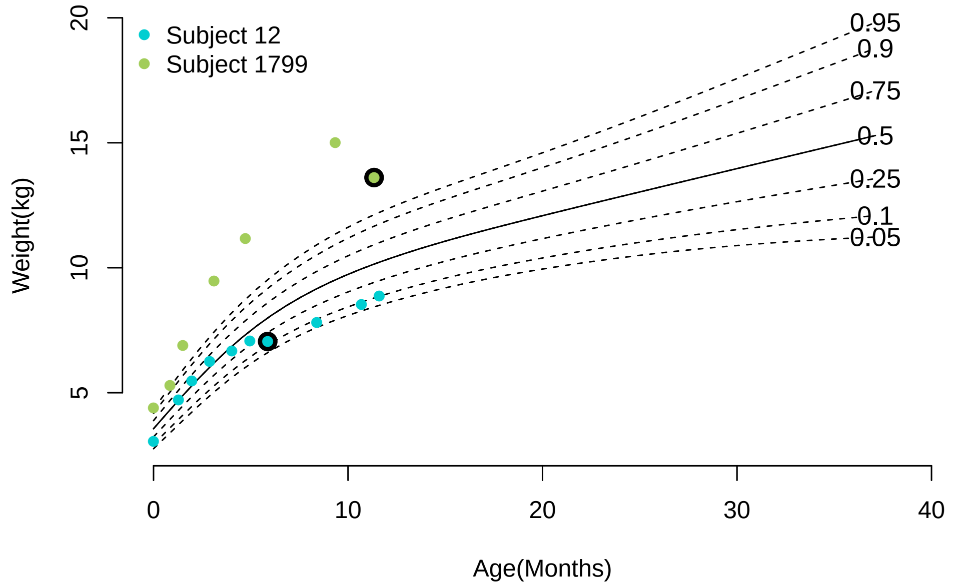
Subject 1799 was heavier than most of his peers at birth, falling on the 95th centile reference curve. He then continually showed a vast increase in weight up to the age of 9.64 months, placing his weight well above the typical weight observed for full-term male infants of his age. However after this point a substantial drop in weight to 13.61 kg was observed by his last weight measurement at 11.34 months of age. His SD score at this age of 2.748, indicating that his weight value is above the median weight of full-term male infants of that age.

3.3 Summary

The LMS method appears to perform well on the Growth and Development Study data, and it is used extensively to produce conventional growth charts such as the WHO weight-for-age child growth standard which are used widely by health professionals for monitoring growth of infants. Although only the e.d.f value of the three curves has been specified when modelling



(a) Full-term female infants



(b) Full-term male infants

Figure 3.8: Gender specific weight growth charts for full-term female and male infants produced by LMS Models 1 and 2, respectively. Superimposed are the observed weight measurements of the two female case infants, with the point bordered in black in each case denoting the observation at which the screening decision is considered.

using either the **lmsqreg** package or the **lms** function in the **GAMLSS** package, with what appears to be quite robust approaches for determining their most desirable values, this modelling approach clearly has its limitations. These include its inability to allow additional informative covariates to be included in the model as well as assuming that the data has been normalised once a suitable power transformation has been performed. In addition, the LMS method does not take into consideration the longitudinal nature of data, such as the Growth and Development Study data, treating the measurements as if they are observed independently on different individuals. These conventional growth charts developed for cross-sectional data are useful when examining an individual's measurement at a specific screening age, as done previously. However if we wish to study an individual's growth pattern rather than a single measurement, tracking them on a conventional growth chart may give us an incomplete impression of their growth. When longitudinal data are available it would be more sensible to incorporate this information in the model. This is further pursued in chapters 5 and briefly discussed again in section 7.1.

Chapter 4

Quantile Regression for Growth Curve Estimation

Quantile regression (Koenker and Bassett, 1978) is another useful tool for constructing conditional growth charts. This approach is a non-parametric method, having the ability to estimate quantiles given a time covariate without the usual distribution assumptions. This deals with the limitation of assuming normality, as stated when using the LMS model, as if the model is distribution free then there are no assumptions to violate. Quantile regression models can be used to obtain selected centiles, such as the 5th, 10th, 25th, 50th (the median), 75th, 90th and 95th centiles, helping to give an impression of the rates of change in all parts of the growth measurement's distribution.

4.1 Linear Quantile Regression Model Methodology

Consider the linear model

$$E(Y_i) = b_0 + b_1 x_i + \epsilon_i \quad \text{for } i = 1, \cdots, m \tag{4.1}$$

for which the least square estimators of b_0 and b_1 are determined by minimising the sum of the squared residuals, ie by minimising $\sum_{i=1}^{m}(y_i-(b_0+b_1x_i))^2$. From solving this it is found that the conditional mean of y given x is the least square estimator of a set of values.

In a similar way the median is the solution to the problem of minimising a sum of absolute deviations,

$$\sum_{i=1}^{m} |y_i - (b_0 + b_1 x_i)|.$$

This can be applied in linear quantile regression to determine the median regression, ie when τ is 0.5, which guarantees that there is a balance between the number of observations lying above and below the conditional median line. Koenker and Bassett (1978) proposed further extending this idea to the estimation of linear parametric models for conditional quantile functions, allowing estimations for different values of τ. Minimising

$$\sum_{i=1}^{m} \rho_\tau(Y_i - x_i^T \beta) \tag{4.2}$$

with respect to p-dimensional parameter β yields an estimate of the τth quantile function of Y given the covariate x, where $\rho_\tau(.)$ is the tilted absolute value function which takes the form

$$\rho_\tau(u) = u(\tau - I(u \leq 0)) = \begin{cases} u\tau & \text{if } u > 0; \\ u(\tau - 1) & \text{if } u < 0. \end{cases} \tag{4.3}$$

This guarantees that 100τ percent of the observations lie below the τ percentile line and $1 - (100\tau)$ percent above. In the simplest case where there is a clear linear relationship between the response y and covariate x, a linear quantile regression model approach can be used to derive the conditional quantiles(or centiles) of y given x. For model (4.1), the reference centile lines can be determined for the τth quantile by minimising

$$\sum_{i=1}^{m} \rho_\tau(y_i - (b_0 + b_1 x_i)).$$

This standard form of the τth quantile gives a non smooth function to minimize, that cannot be differentiated. Instead it can be solved efficiently using linear programming methods as described at the beginning of the chapter. Inference for quantile estimates can be challenging because it involves non-parametric density estimation. In practice, resampling methods such as the bootstrap are employed to produce standard errors and confidence intervals. For more information on estimation and inference in quantile regression, see Koenker (2005).

For the special case τ=0.5, the estimate of the absolute value function takes the form

$$\rho_\tau(0.5) = \begin{cases} 0.5u & \text{if } u > 0; \\ -0.5u & \text{if } u < 0. \end{cases}$$

so the median centile estimate, derived for $y_1, \cdots, y_m$, would be equal to

$$\text{argmin}_{b_0} \sum_{i=1}^{m} \rho_{0.5}(y_i - 0.5b_0)$$

$$\text{argmin}_{b_0} \sum_{i=1}^{m} (y_i - 0.5b_0) I(y_i - 0.5b_0 > 0) - \sum_{i=1}^{m} (y_i - 0.5b_0) I(y_i - 0.5b_0 < 0)$$

The **rq** function which is available in the **quantreg** package in **R** can be used to implement the linear quantile regression model approach.

4.2 Linear Quantile Regression Model for the Abdominal Circumference Data

The Abdominal Circumference Data, as discussed in section 1.4, was used to illustrate the concept of the linear quantile model, which is fitted using the **R** package **quantreg**. Figure 4.1 which shows the change in abdominal circumference with gestational age across the range of centiles, illustrates the similarity in the lines produced by the ordinary least squares regression and the 0.50th quantile (median) linear regression models. This figure also illustrates the difference in modelling using the ordinary least square regression,

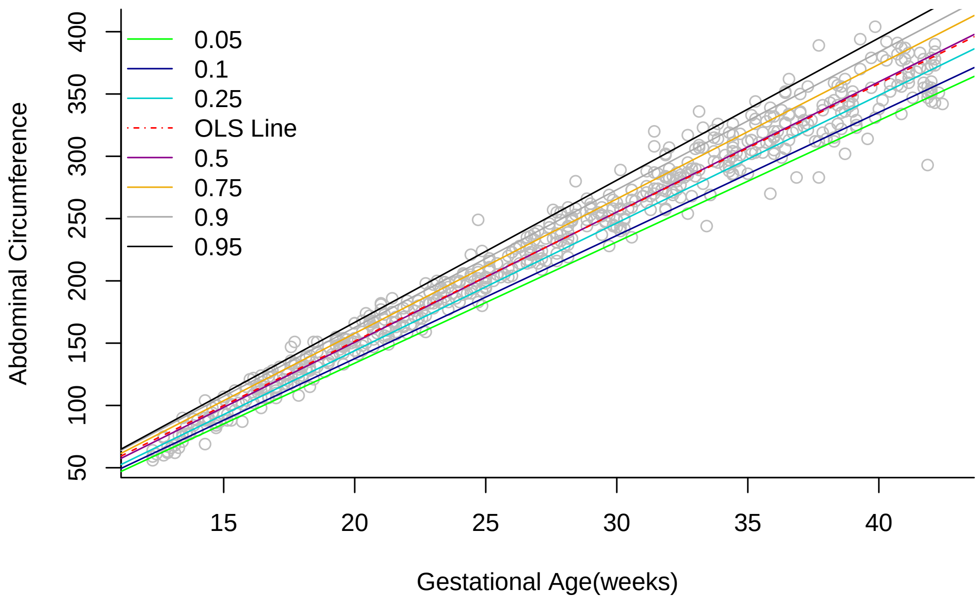


Figure 4.1: Abdominal Circumference Data with estimated quantiles(solid lines) constructed by the linear quantile regression model as well as the ordinary least squares regression line (dashed).

which draws a single line through the data points to linear quantile regression which gives a different line for each specified centile.

Linear quantile regression is not usually appropriate for growth chart modelling, because as with the Growth and Development Study Data, the patterns observed in the growth variable across the covariate are most commonly non-linear.

4.3 Quantile Regression Model Methodology for Growth Data

More generally, for reference growth charts it is convenient to parameterise the conditional quantile functions as linear combinations of a few fixed basis functions, which allow flexible reference centile curves to be fitted. B-splines, discussed in detail in section 2.2, are particulary convenient for this purpose. Given a choice of knots for the B-splines, estimation of the growth curves is equivalent to fitting a parametric linear quantile regression with $x_{ij} = B_j(x_i)$ where B_j is the jth function in the B-spline basis. For details see Wei et al. (2006).

For the conditional τth quantile of Y at x, a general form for the quantile regression model is

$$Q_Y(\tau|x_i) = f_\tau(x_i)$$

where $f_\tau(x_i)$ is a smooth but unspecified function relating the quantiles to the covariate x, in the general case for growth charts, age. $f_\tau(\cdot)$ describes the age-specific growth charts at the selected τ which is expressed when implementing B-splines of degree q as $f_\tau(\cdot) = \sum_{j=1}^{J} \alpha_j B_j(x; q)$ which is a linear combination of the J basis functions and the corresponding coefficients $\hat{\alpha}_j$ to be estimated (Muggeo et al., 2012).

This modelling method, which adopts a B-spline term for the covariate, can be implemented in **R** using the **quantreg** package. As detailed in section 2.2, when fitting a B-spline model it is necessary to choose the order of the polynomials as well as the number and placement of knots. Explanatory approaches are adopted to determine sensible choices for the remaining of this chapter.

4.4 Quantile Regression Model for the Growth and Development Study Data

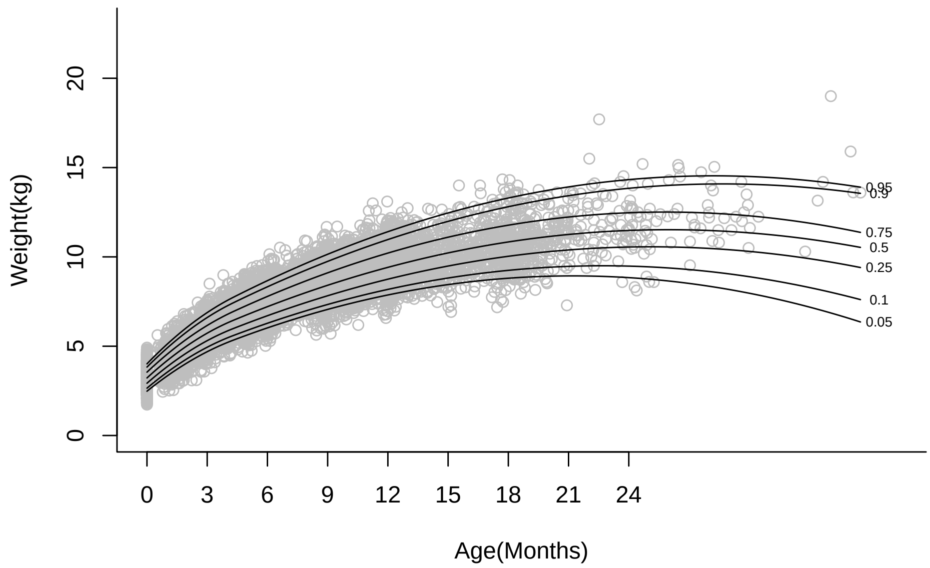
The Growth and Development Study Data was used to illustrate the quantile regression model for composing gender-specific weight growth charts for full-term infants, using the **quantreg** and **quantreggrowth R** packages

The first approach to choosing the degree of the B-spline as well as the most favourable number and positioning of knots, was to use a small number of quantile knots with B-splines of quadratic and cubic degree. The reference centile curves produced via these approaches were examined visually, as well via a diagnostic plot and the number of knots were increased if this appeared necessary. If sensible curves were not deduced when quantile knots were considered, the process was repeated with unequally spaced knots until a more satisfactory curve was obtained.

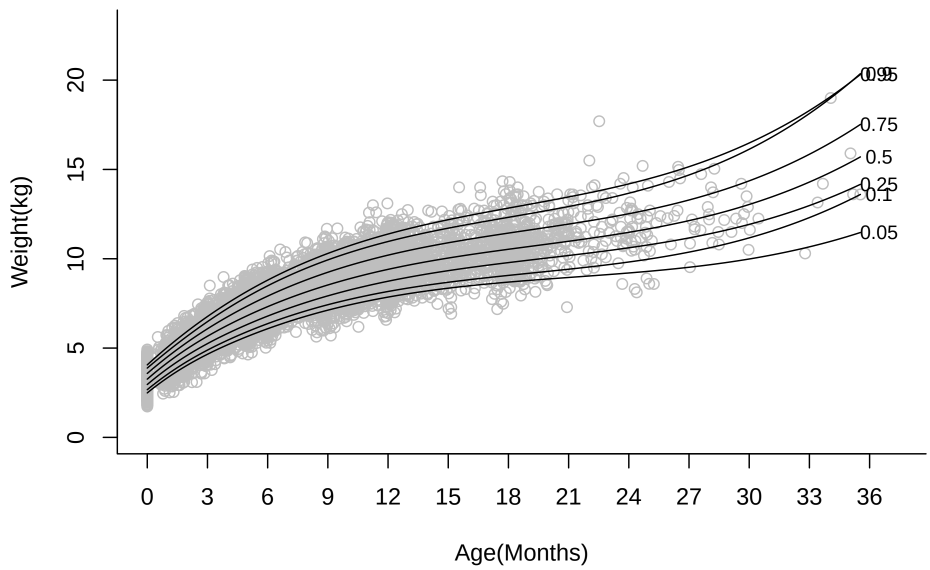
Fitting a quantile regression model with B-splines of quadratic and cubic degree with an interior quantile knot at 4.30 months, as shown in Figures 4.2a and 4.2b, to the full-term female data, produces sensible reference centile curves over the first 21 months of growth for both degrees of freedom. However after this point the curves struggle to capture the trend with the cubic spline model exhibiting signs of crossing quantiles, indicating that one knot is not sufficient.

Adding an additional interior quantile knot, so that interior knots are positioned at 2.59 and 6.43 months, improves the reference centile curves produced when quadratic B-splines are implemented, as illustrated in Figure 4.3a. However Figure 4.3a shows that the additional flexibility does not have a favourable effect when cubic B-splines are used.

Inclusion of a third interior quantile knot, so that the interior knots are now positioned at 1.77, 4.30 and 9.02 months, as displayed in Figure 4.4, when a quadratic degree of smoothing is applied, leads to quantile crossing. This implies that fitting a quadratic B-spline model with two interior quantile knots, is preferable. However there are still visible indications that this struggles to model the pattern in the later months, due to the sizeable amount of variability between the limited weights documented in the records at older ages. Such regions of sparse data, could lead to quantile crossing which is a problem with the model fit. This indicates that B-splines with unequally spaced knots where the emphasis is on positioning more knots at

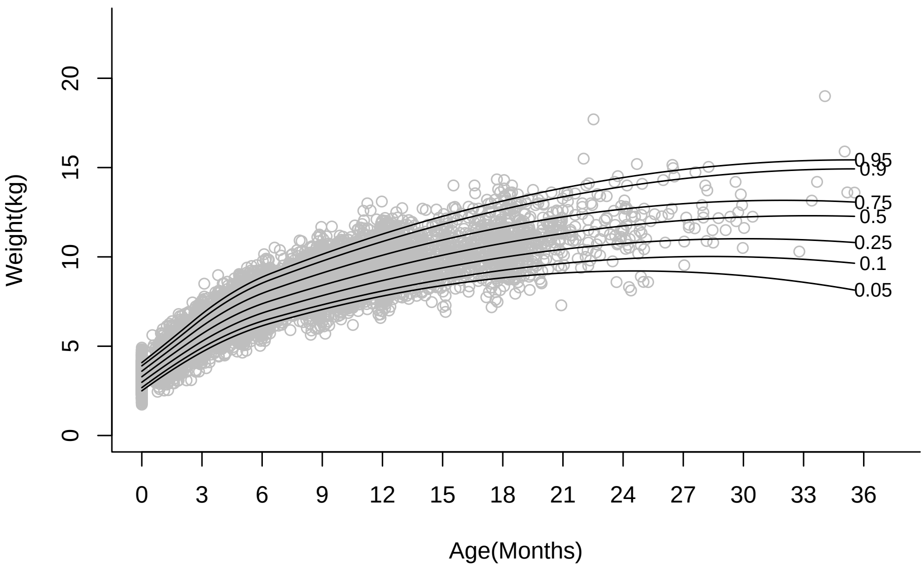


(a) Quadratic B-splines

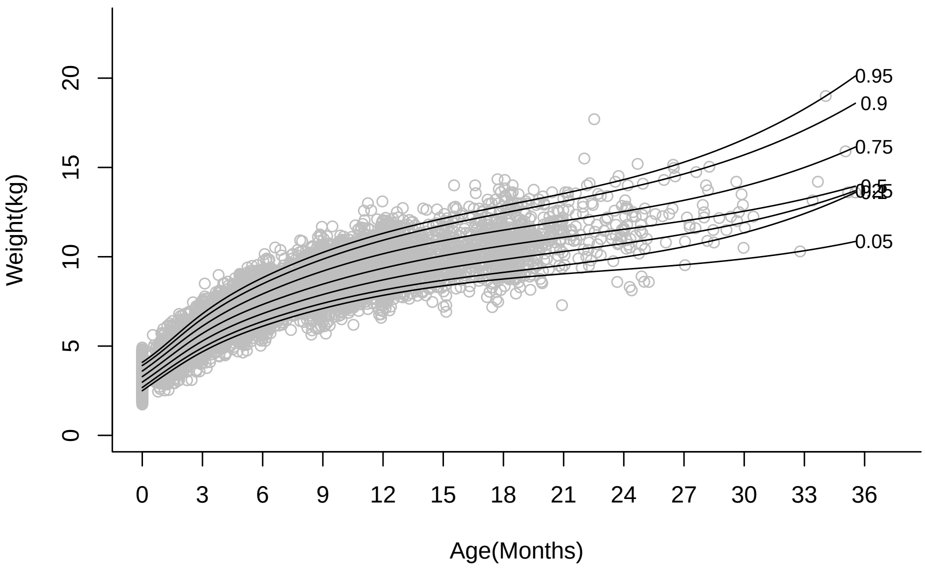


(b) Cubic B-splines

Figure 4.2: Weight growth charts for full-term female infants constructed by quantile regression models with quadratic and cubic B-splines, with one interior quantile knot at age 4.30 months.



(a) Quadratic B-splines



(b) Cubic B-splines

Figure 4.3: Weight growth charts for full-term female infants constructed by quantile regression models with quadratic and cubic B-splines, with two interior quantile knots at ages 2.59 and 6.43 months.

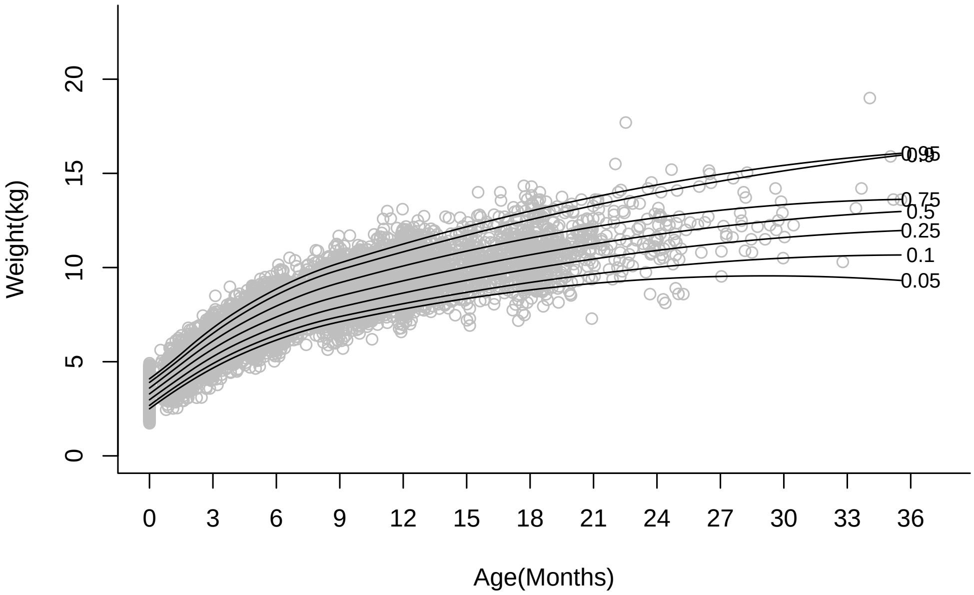


Figure 4.4: Weight growth chart for full-term female infants constructed by a quantile regression model with quadratic B-splines, with three interior quantile knots at ages 1.77, 4.30 and 9.02 months.

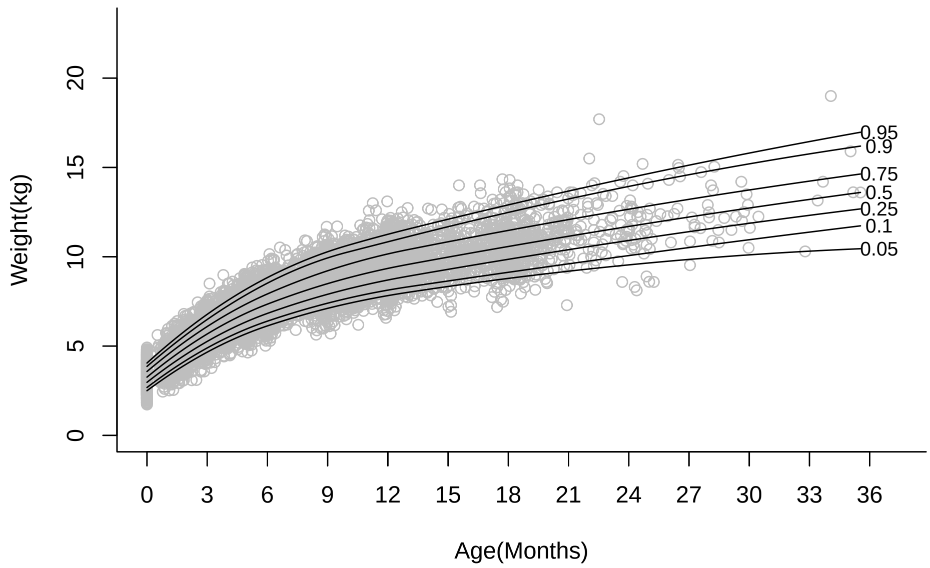
later months, may produce more favourable reference centile curves, handling this problem of variability.

The above process of model selection was repeated with alternatively differing positionings of the knots chosen. Figure 4.5a displays the reference centile curves produced using a quadratic B-spline with three interior knots at ages 6, 10 and 13 months. Although positioning knots at later ages leads to an improvement in the reference centile curves constructed they still struggle to capture the trend in the upper tail of the distribution. For the full-term females infant weight data a restricted quantile regression model from birth to 24 months of age, fitted using quadratic B-splines with interior knots at ages 2.5 and 10 months, which will be referred to as QR Model 1, gave the weight growth chart shown in Figure 4.5b. These curves are more robust than the ones for the full age range shown in Figure 4.5a, because they don't rely on a very small number of observations at the upper-end of the age scale.

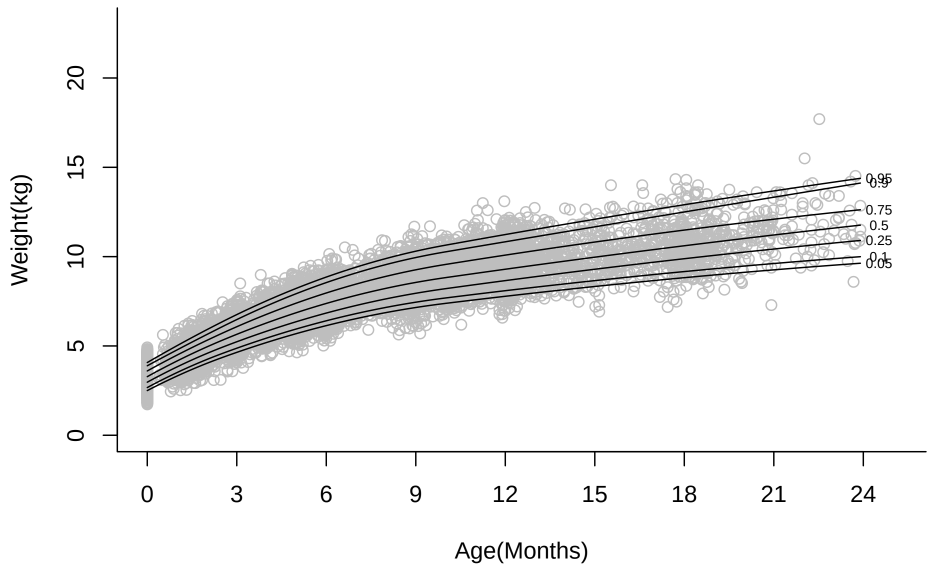
To assess the goodness of fit of such growth charts, estimated reference centile values are simulated from the quantile regression model as described in Wei et al. (2006). This is done by partitioning by age into equally spaced intervals between the maximum and minimum age points, with the full-term female and male weight data being partitioned at 20 and 24 ages, respectively. Then between each of these partitioned ages, at unique ages, weight measurements are predicted from the quantile regression model for each centile. These estimated reference centiles are compared directly to the true centile value at this unique age and thus for each interval in age an overall level of agreement is deduced.

Figures 4.6a and 4.6b illustrate the goodness of fit of the quantile regression models shown in Figures 4.5a and 4.5b to the data. It is clear that the fit of the models is not as good for the later months. This is illustrated by the fluctuating reference centile curves estimated under the quantile regression approach at the later months, with the difference between the distributions appearing to be particularly obvious in the 25th, 50th and 75th centiles. Figure 4.6a stresses how the agreement between the estimated and true reference centile curves at the very upper end of the age scale, where a smaller number of observations are available, depletes very dramatically.

By repeating the same procedure of model selection for the full-term male infant weight data, a sensible choice for a weight growth chart for full-term male infants was constructed form birth to 24 months by a quantile regression model with quadratic B-splines with two unequally spaced interior knots at ages 3 and 11.5 months. This model will be referred to as QR Model 2. Figures 4.7a and 4.7b illustrating this growth chart and its corresponding goodness of fit diagnostic plot, show that the reference centile curves fitted using QR Model 2 are a relatively good fit to that data. There is again some

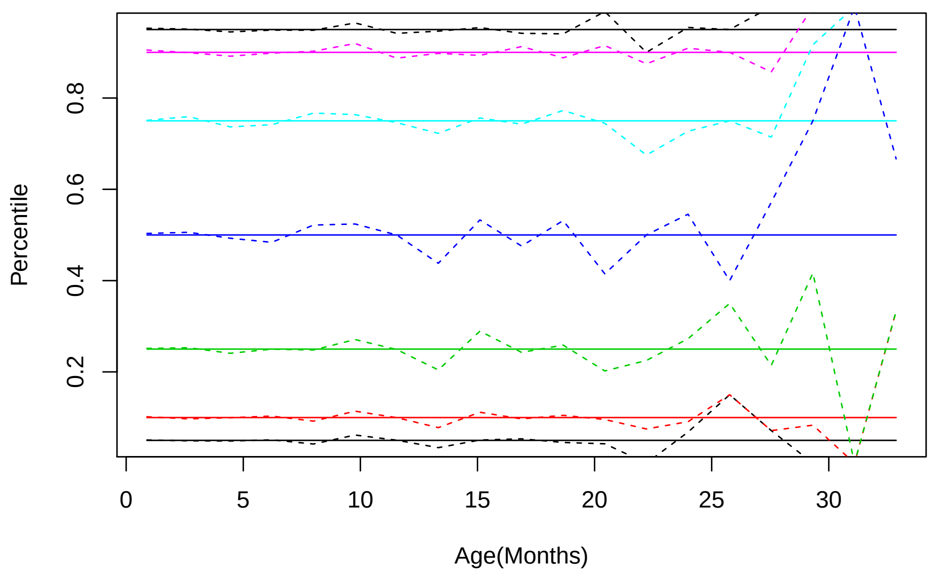


(a)

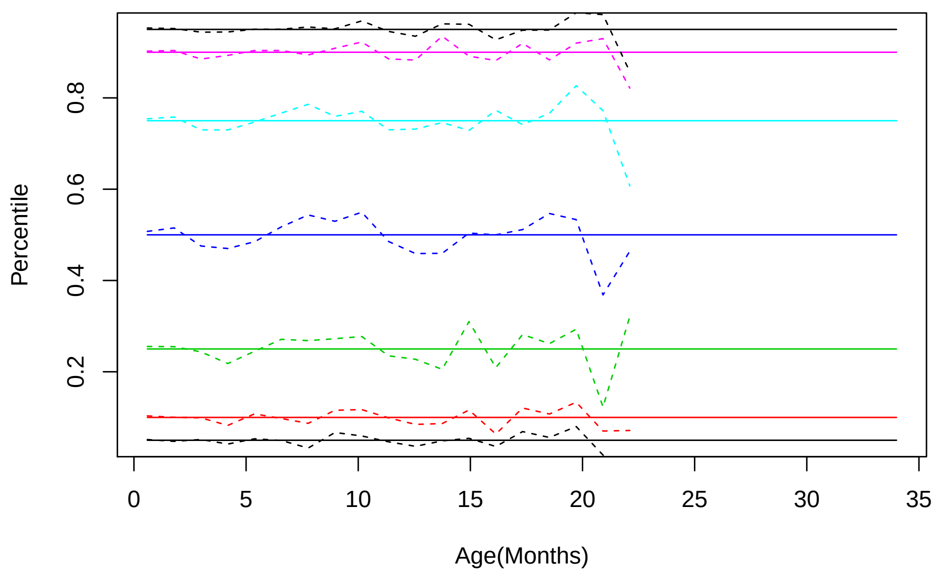


(b)

Figure 4.5: From top to bottom, weight growth charts for full-term female infants constructed by a quantile regression model with quadratic B-splines with three unequally spaced interior knots, at ages 6, 10 and 13 months modelling across the entire age range and a quadratic B-spline with two unequally spaced knots, at ages 2.5 and 10 months, applied to the restricted age range.



(a)



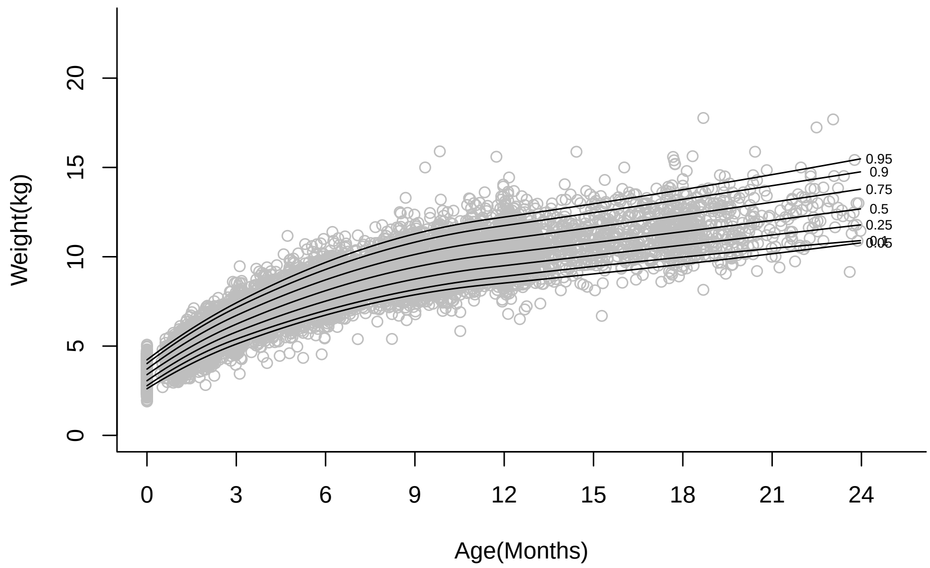
(b)

Figure 4.6: From top to bottom, diagnostic plots assessing the goodness of fit of the weight growth charts for full-term female infants shown in Figures 4.5a and 4.5b, respectively. The estimated reference centile curves under the quantile regression approach are represented as broken curves and the unbroken lines representing the true $\tau 100\%$ centile curves.

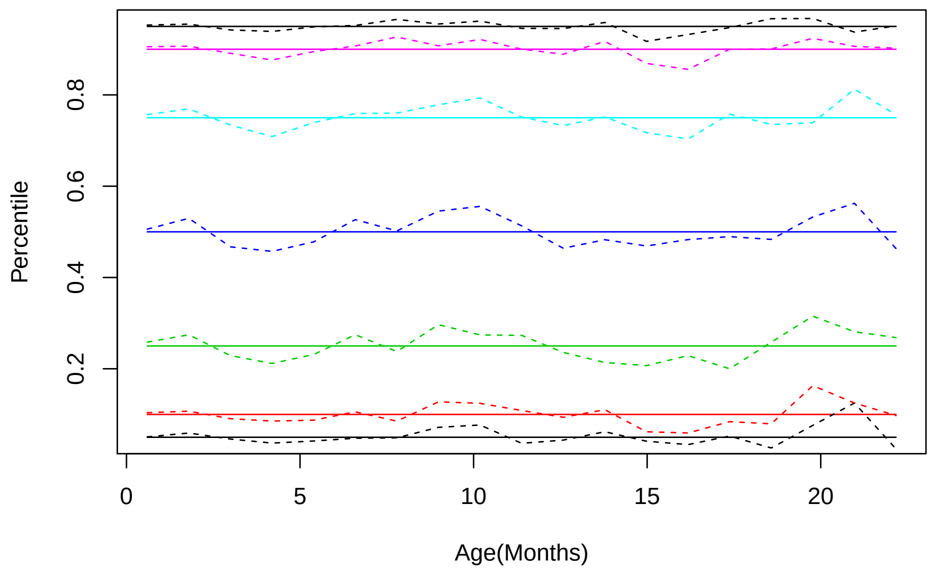
indication that at the later months, the fit is not as good, with in particular the difference between the distribution being more pronounced for the 25th, 50th and 75th centiles.

The gender-specific weight growth charts for full-term female and male infants, composed by QR Models 1 and 2, respectively, were compared directly to those produced by LMS Models 1 and 2. Figures 4.14a and 4.14b show that the two modelling approaches in fact produce relatively similar gender-specific growth charts between birth and two years of age, capturing how weight gradually increases with age and how the rate of increase steadily reduces. In particular only very minor differences are present between the gender-specific weight growth charts constructed by the two modelling approaches for the first 11 months. After this point a larger amount of variability is visible, with the quantile regression models tending to produce slightly lower reference centile curves. Overall the two approaches appear to produce satisfactory reference centile curves for modelling weights of full-term female and male infants. Although both approaches give similar results for ages with a large number of observations, the LMS approach appears to produce realistic reference centile curves in intervals of limited observations whereas the quantile regression approach struggles. This is a consequence of quantile regression being a non-parametric approach, which requires a sufficient number of observations for reliable estimation.

Screening based on the gender-specific weight growth charts for full-term infants can be performed using QR Models 1 and 2. For each of the four case infants described in section 1.3 centile estimates are used to evaluate the positioning of the weight measurements on the gender appropriate weight growth charts, observed at the screening age, as well as to directly compare between the growth charts constructed by LMS Models 1 and 2 and QR Models 1 and 2.

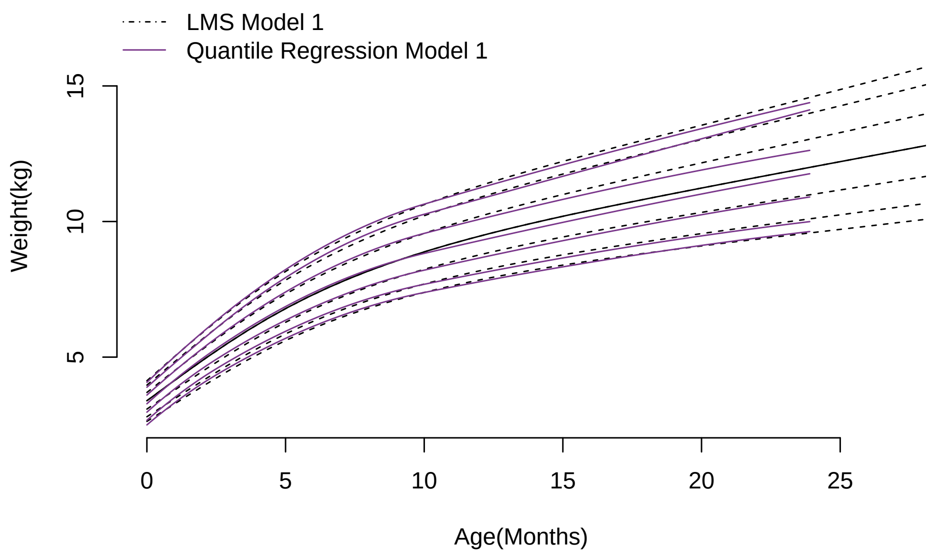


(a)

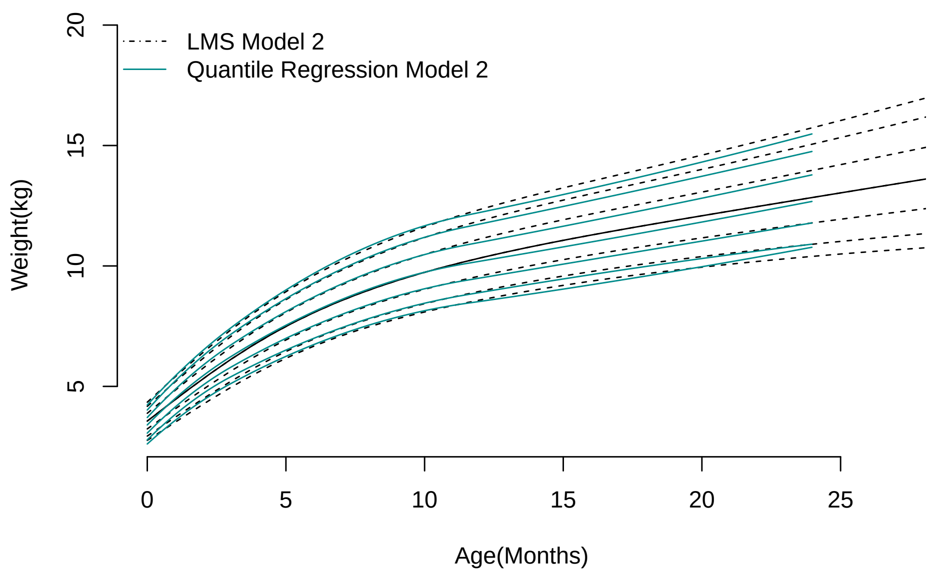


(b)

Figure 4.7: From top to bottom, weight growth chart for full-term male infants constructed by a quantile regression model with quadratic B-splines with two unequally spaced interior knots at 3 and 11.5 months, and a corresponding diagnostics plot assessing the curves' goodness of fit. The estimated reference centile curves under the quantile regression approach are represented as dashed curves and the solid lines representing the true $100\tau\%$ centile curves.



(a) Full-term Female Infants, LMS Model 1 and QR Model 1



(b) Full-term Male Infants, LMS Model 2 and QR Model 2

Figure 4.8: Comparison of the gender-specific weight growth charts for full-term infants constructed by LMS Models 1 and 2 and QR Models 1 and 2.

Figure 4.9a illustrates the growth curves for full-term female infants' weight measurements, estimated by QR Model 1 described previously. Subject 1500's growth follows an almost identical trend in relation to the reference centile curves as that seen in Figure 3.8a, falling consistently above the 5th reference centile curve. Table 4.1 detailing the centile estimate, suggests that at the age at which the screening decision is considered her centile estimate is 0.084, indicating that her weight value falls above the 5th reference centile curve estimate of full-term female infants of this age. This is slightly different to the LMS centile estimate of 0.014.

Subject 146's growth also follows an almost identical trend in relation to the reference centile curves as that seen in Figure 3.8a. However by her final measurement at roughly 12 months she falls well above the median reference centile curve whereas previously, based on the curves produced using the LMS Model 1, she fell just above. This is reflected in the centile estimate value, detailed in Table 4.1, with the quantile regression centile estimate of 0.615 indicating that her weight value is well above the median weight of full-term female infants of this age in comparison to the LMS centile estimate of 0.574 which suggests she is just above. This slight difference is due to the minor variation in the growth charts produced by the two methods.

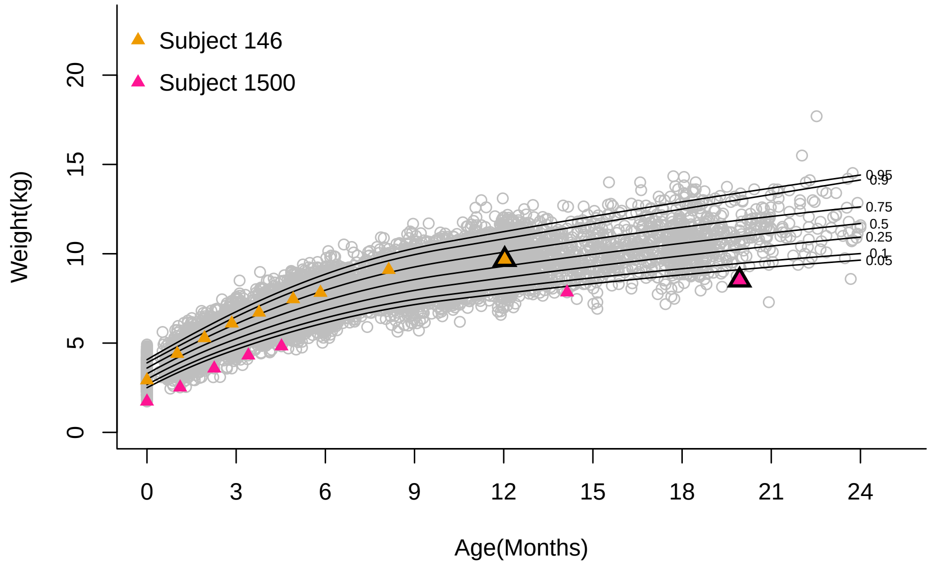
Figure 4.9b shows the growth curves for full-term male infants' weight measurements, estimated by QR Model 2 described previously. Subject 12 follows an almost identical trend in relation to the reference centile curves as that seen in Figure 3.8b, moving substantially between the reference centile curves with age. At the age at which screening is considered, roughly 6 months, he falls just above the 10th reference centile curve, which is very similar to his position in relation to the LMS reference centile curves at this screening age. This similarity is reflected in the centile estimates, detailed in Table 4.1, with the quantile regression and LMS centile estimates of 0.134 and 0.140 respectively, indicating only a minor difference in the growth chart at this age.

Subject	Gender	Screening Age(Months)	LMS centile estimate	QR centile estimate
1500	Female	19.93	0.014	0.084
146	Female	12.03	0.574	0.615
12	Male	5.87	0.140	0.134
1799	Male	11.34	0.997	0.979

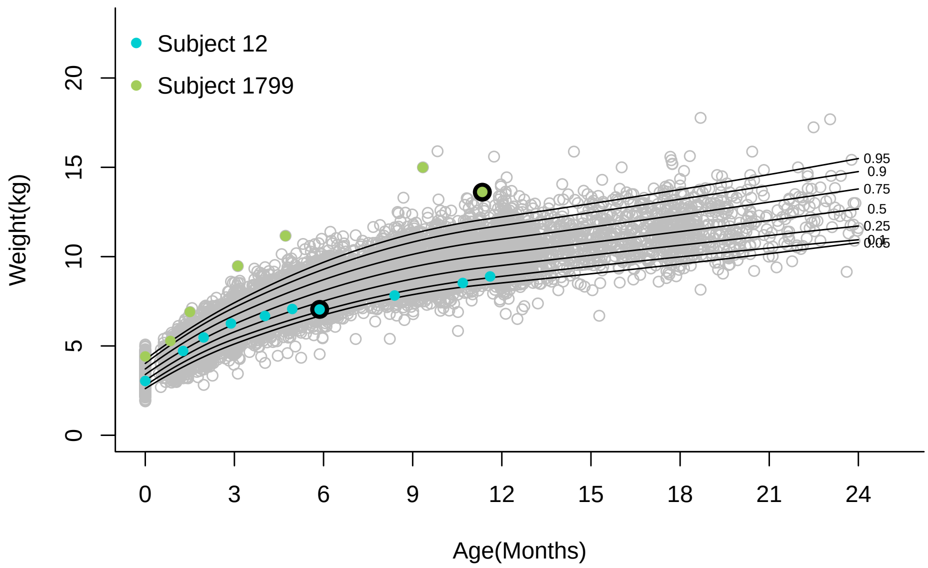
Table 4.1: Centile estimates for the four case infants, at the specified screening ages based on the gender-specific weight growth charts for full-term infants constructed by LMS Models 1 and 2 and QR Models 1 and 2.

Subject 1799 also follows an almost identical trend in relation to the reference centile curves as that seen in Figure 3.8b, falling just above the 95th centile curve at birth and thereafter well above. At the age at which screening is considered, 11.34 months, his centile estimate of 0.979 suggests that his weight value falls just above the 95th centile reference line. This is a slightly lower estimate than the LMS centile estimate of 0.997.

These differences are due to the minor variation between the growth charts produced by these two approaches at the later ages, with in particular the quantile regression approach generating centile curves that fall slightly above those produced by the LMS method. However this difference is only slight, so again these comparisons confirm the similarity in the growth charts produced by the LMS and quantile regression approach.



(a) Full-term Female Infants



(b) Full-term Male Infants

Figure 4.9: Weight growth charts for full-term female and male infants constructed by QR Models 1 and 2, respectively. Superimposed are the observed weight measurements of the four case infants, with the point bordered in black in each case denoting the observation at which the screening decision is considered.

4.5 Penalised Quantile Regression Model with Monotonicity and Non-Crossing Constraints Methodology

Theoretically reference centile curves should be non-crossing but while the LMS model method implicity leads to curves of this form via a scaling function, as seen in chapter 3.2, extra work is required when using the quantile regression model method to return non-crossing centile curves. A manual approach was adopted in section 4.3 to determine sensible choices of the quantile regression models, guaranteeing, that amongst other thing, this property of non-crossing centile curves held. The **R** package **quantregGrowth**, which is another approach for quantile regression, modifies the model so that any couple of adjacent quantiles are strictly non-crossing. This approach which also uses a B-spline term for the time covariate, can additionally apply a penalty to the B-spline and a monotonicity constraint, both of which are properties described in sections 2.3 and 2.4, respectively. When implementing a penalised quantile regression model, the $f_{\tau_k}(x_i)$ must be estimated using penalized spline methods, seen in Claeskens et al. (2009). Furthermore a full discussion of estimation of non-crossing and monotone quantile curves is given in Muggeo et al. (2012).

The model selection when implementing the **quantregGrowth** package was done by initially fitting a quantile regression model with B-splines with non-crossing and monotonicity constraints applied. Although the positioning of knots can not be specified, as by default equally spaced knots are implemented in this package, the number of interior knots and degrees of the spline polynomial can be. To determine the most effective choice of these properties, initially a model with limited number of interior knots for both B-splines of quadratic and cubic degree were specified. The reference centile curves produced were examined visually and compared using the Schwarz information criterion (SIC) (Schwarz, 1978), which was ascertained for each growth chart. The SIC is closely related to AIC, with SIC being optimal. If sensible reference centile curves were not obtained in this way then the number of interior knots was increased and the process is repeated again. If for a substantial number of interior knots a suitable growth chart was not found, then smoothing using P-splines with no monotonicity constraints was considered.

A similar model selection procedure to that discussed previously could again be used when implementing the penalised quantile regression model with non-crossing constraints. Additionally now the smoothing parameter λ and the difference order of the penalty had to be specified. For each combination of degree of spline polynomial and number of knots, a second and third order penalty was implemented for a spectrum of smoothing parameter values. This choice of smoothing parameter values was very subjective, however values 2, 6 and 10 were chosen as this hopefully allowed an adequate range of smoothing parameter values to be tried. If required, a monotonicity constraint could be applied to the model, which restricted the smoothing function to be strictly non decreasing. This required only a minor change in the model specification.

Clearly these selection criteria are rather subjective, which is one of the key limitations to this growth curve estimation technique.

4.6 Penalised Quantile Regression Model for the Growth and Development Study Data

The effectiveness of the **quantregGrowth** package for composing weight growth charts for full-term females infants by a quantile regression model using B-splines with monotonicity and non-crossing constraints, based on Growth and Development Study Data, was found to be very poor. The reference growth curves evidently struggled to capture the trend in the later

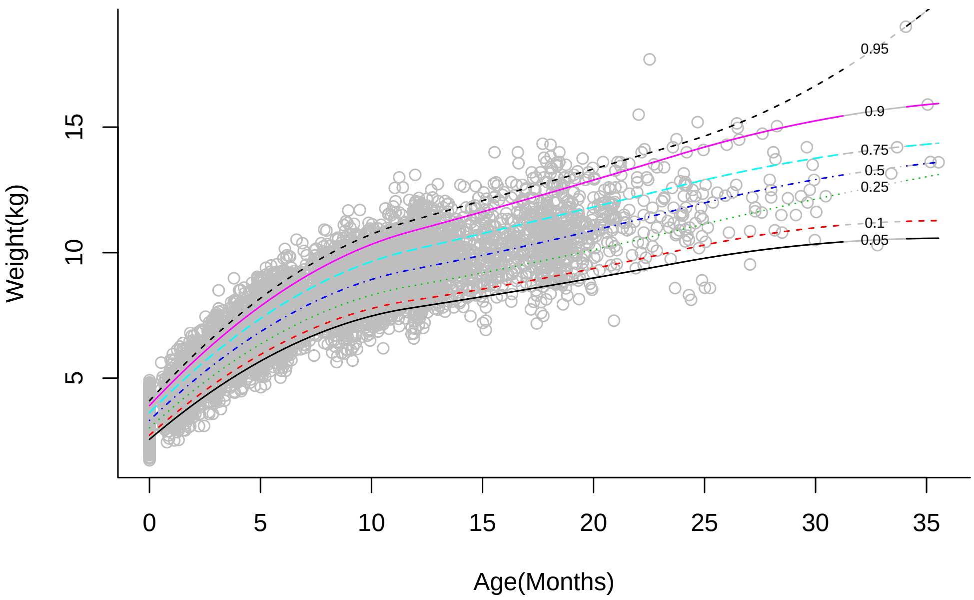


Figure 4.10: Weight growth charts for full-term female infants produced by a quantile regression model with quadratic B-splines with two interior knots.

months even when a wide variety of number of knots and degrees of spline polynomial were specified. For instance fitting a quadratic B-spline model with two equally spaced interior knots, as seen in Figure 4.10, illustrates a concerning pattern at the upper end of the age scale. This is a similar issue as that stressed when fitting using QR Models 1 and 2 for composing gender-specific growth curves over the entire age range.

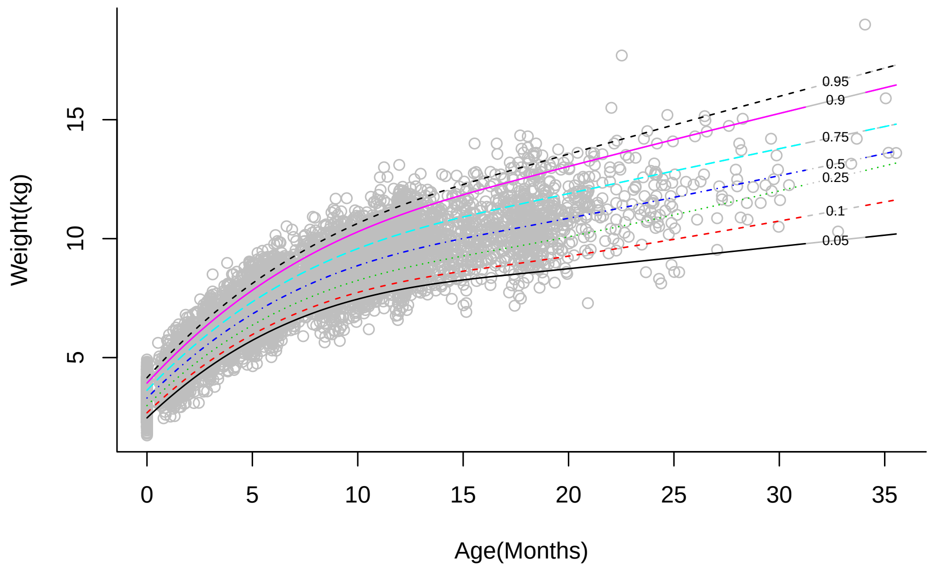
Fitting a penalised quantile regression model which has a non-crossing constraint using P-splines with one interior equally spaced knot, second order penalty, cubic degree of the P-splines and smoothing parameter value of 6, was visually found to produce a sensible weight growth chart for full-term female infants. This Model will be referred to as PQR Model 1, and is the 8th model down detailed in Table 4.2. This indicates that the issue of fluctuating curves at later months, which clearly does not reflect reality, can be resolved by smoothing using P-splines. The 9th model down detailed in

Number of Interior Knots	Spline Degree	λ	Difference Penalty	SIC
1	Quadratic	2	2	-7.394
1	Quadratic	6	2	-7.276
1	Quadratic	10	2	-7.170
1	Quadratic	2	3	-7.418
1	Quadratic	6	3	-7.371
1	Quadratic	10	3	-7.343
1	Cubic	2	2	-7.477
1	Cubic	6	2	-7.272
1	Cubic	10	2	-7.104
1	Cubic	2	3	-7.558
1	Cubic	6	3	-7.455
1	Cubic	10	3	-7.304

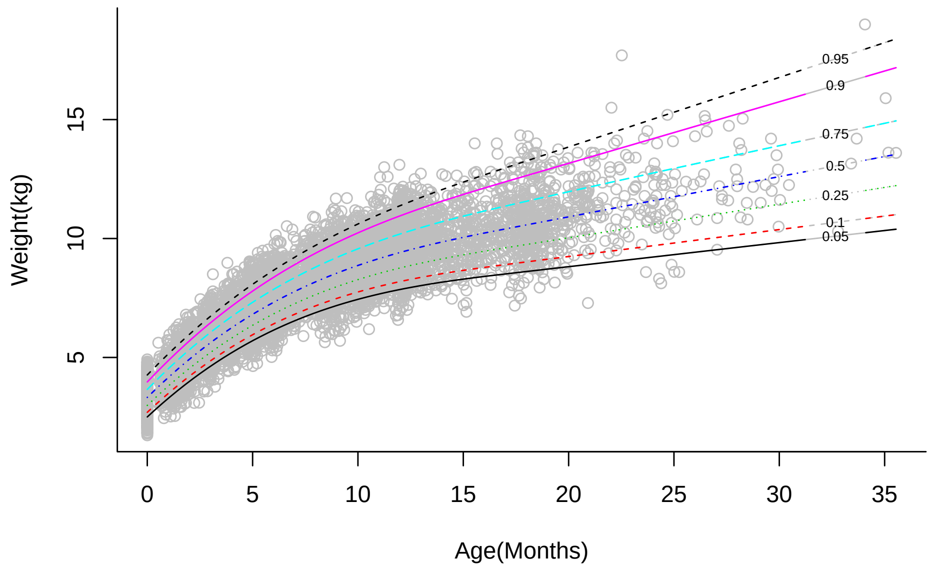
Table 4.2: SIC values for weight growth charts for full-term females infants constructed by penalised quantile regression models with a non-crossing constraint fitted with a series of P-spline curves with one interior equally spaced knot, quadratic and cubic degree of the P-splines, differing smoothing parameter λ values and second and third order difference penalty.

Table 4.2, which is an adaption of PQR model 1 with a reduced λ value of 2 has an SIC value of -7.104. This is a minor improvement on PQR model 1 SIC which is -7.272, however visually choosing $\lambda = 6$ seems more sensible, as seen when comparing Figure 4.11a and 4.11b. The spacing between centiles curves appears to be more even when PQR Model 1 is fitted, with it being comparatively more evident at the later months. The resulting reference centile curves are strictly non decreasing so there is no need to add a monotonicity constraint.

A similar selection process was used for the full-term male infant data. It



(a) Smoothing parameter $\lambda = 2$



(b) Smoothing parameter $\lambda = 6$

Figure 4.11: From top to bottom, weight growth charts for full-term female infants produced by penalised quantile regression models with a non-crossing constraint, cubic P-splines, one interior equally spaced knot, second order difference penalty and λ values 2 and 6, respectively.

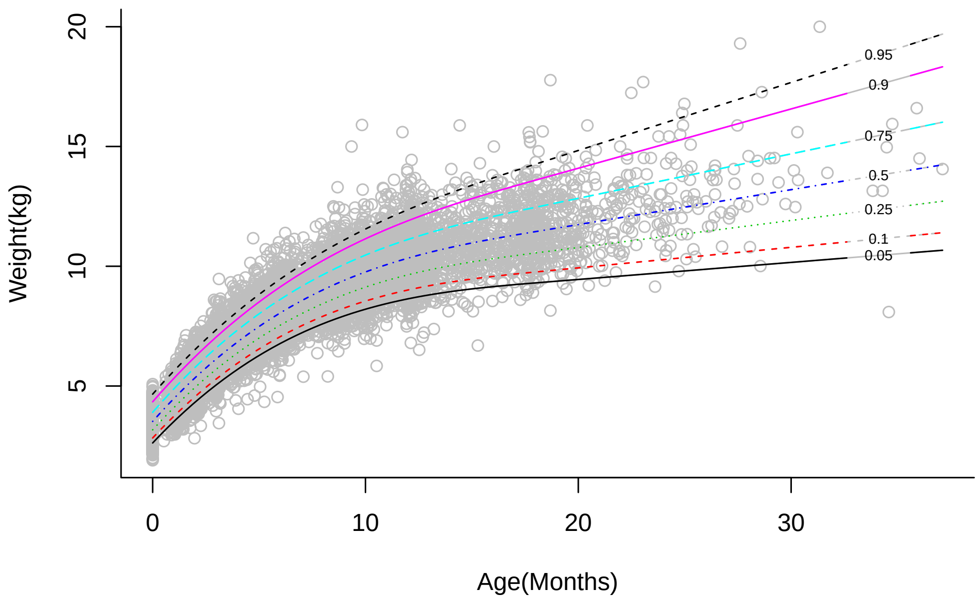


Figure 4.12: Weight growth charts for full-term male infants produced by penalised quantile regression models with a non-crossing constraint, cubic P-splines, one interior equally spaced knot, second order difference penalty and λ value 10.

was identified that the quantile regression models with B-splines with non-crossing and monotonicity constraints again performed badly, but cubic P-splines with a second order difference penalty, one interior equally spaced knot and $\lambda = 10$ was a sensible choice, visually producing a sensible weight growth chart for full-term male infants as seen in Figure 4.12. This chosen penalised quantile regression model (PQR Model 2) which has a non-crossing constraint has a very small SIC of -6.497 indicating that the model is a very good fit to the data. Again the resulting reference centile curves are strictly non decreasing so there is no need to add a monotonicity constraint.

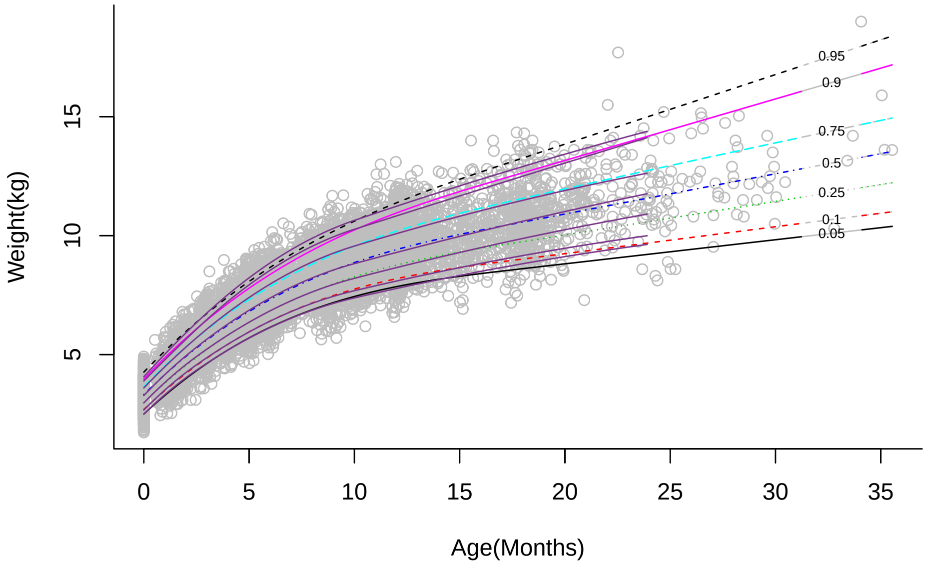
Comparing the gender-specific weight growth charts composed by the QR Models 1 and 2 to those produced by the PQR Models 1 and 2, as seen in Figures 4.13a and 4.13b, illustrates that the two approaches of modelling produce relatively different growth charts. Although a strong likeness is visible between the 25th, 50th and 75th reference centile curves composed by the two model approaches, the other curves evidently diverge from each other, with the difference greatest at later months. The gender-specific growth charts produced by the penalised quantile regression models with the non-crossing constraint appear to be more satisfactory, performing well across the entire age range. This stresses the advantage of modelling using quantile regression with penalised B-splines and a non-crossing constraint instead of an unconstrained quantile regression model.

Due to the evident similarities in the gender-specific weight growth charts for full-term infants produced by QR Models 1 and 2 and LMS Models 1 and 2, it was unsurprising that when directly comparing the growth charts composed by the PQR Models 1 and 2 to the LMS Models 1 and 2, a contrast in the curves is visible. Evidently the difference, as shown in Figures 4.14a and 4.14b, between the centile curves fitted becomes more pronounced at the later months.

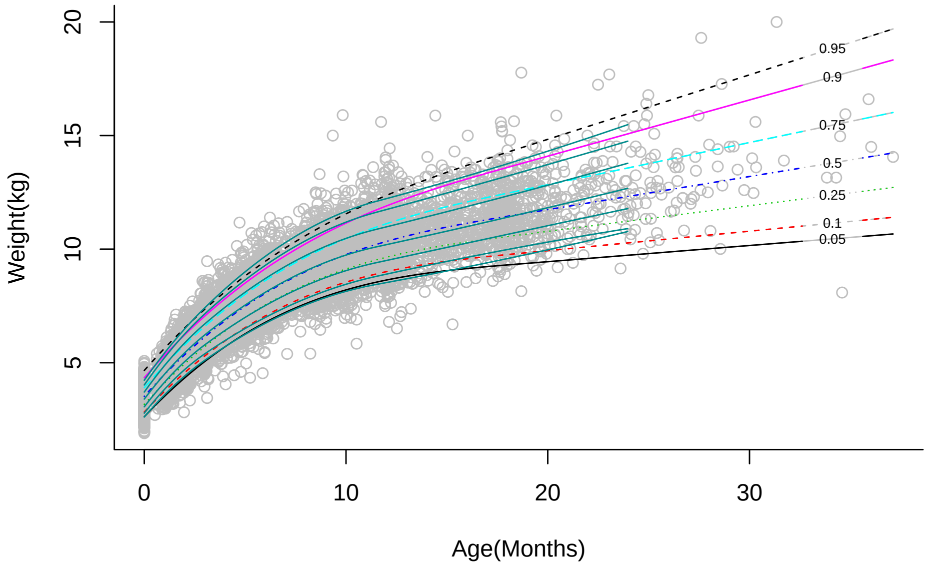
4.7 Summary

Although the quantile regression model appears to perform well at composing growth charts, particulary when using penalised quantile regression models, allowing estimation without a distribution specification, it relies on subjective choice of smoothing parameters, whereas the LMS method has the advantage of only requiring the specification of the e.d.f for each model.

Furthermore both the constrained and unconstrained quantile regression models of weight on age are models for cross-sectional data and would need to be generalised to accommodate longitudinal data.

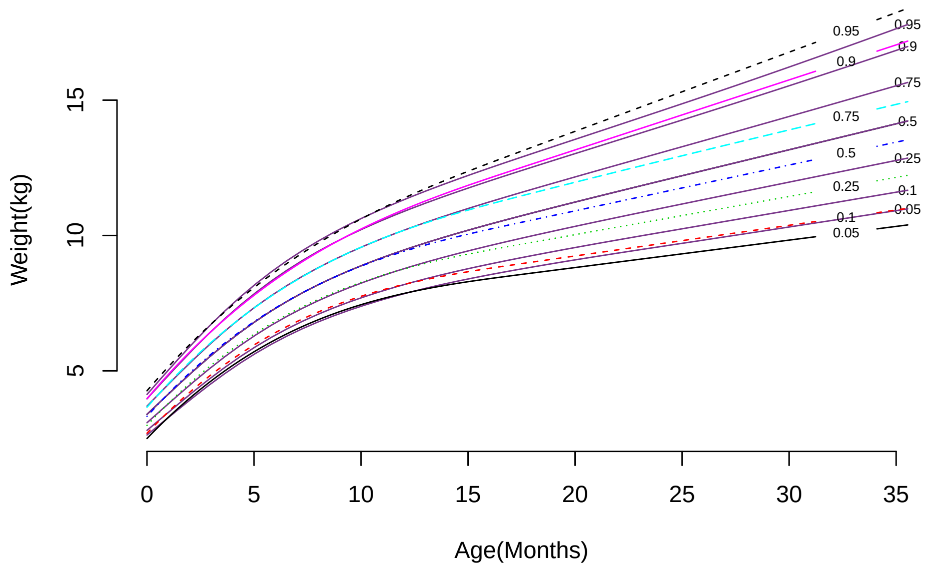


(a) Full-term female infants

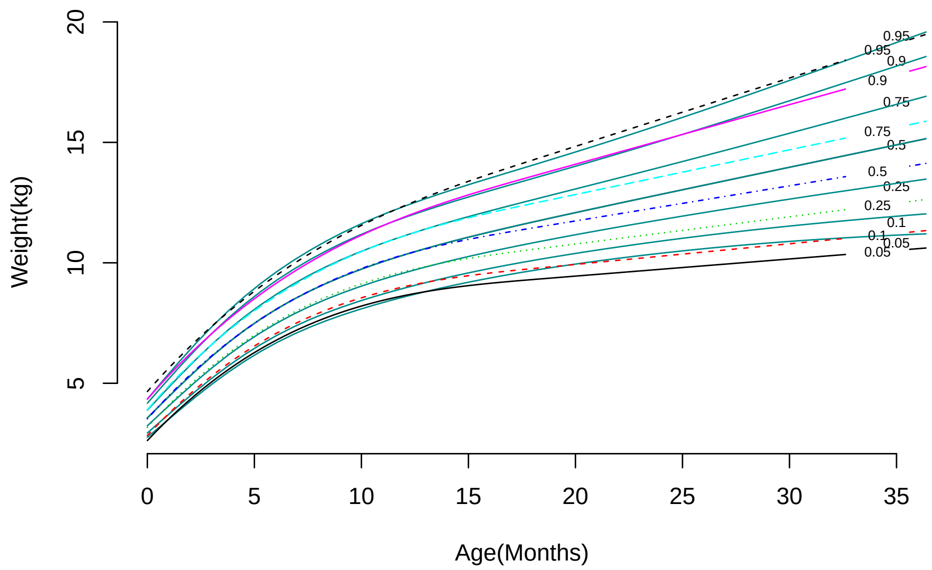


(b) Full-term male infants

Figure 4.13: From top to bottom, comparison of gender-specific growth charts of weights for full-term infants, fitted using the QR Models 1 and 2 to the PQR Models 1 and 2. The solid purple and turquoise lines represent the reference centile curves fitted by the QR models.



(a) Full-term Female Infants



(b) Full-term Male Infants

Figure 4.14: From top to bottom, comparison of the gender-specific weight growth charts for full-term infants constructed by LMS Models 1 and 2 to the PQR Models 1 and 2. The solid purple and turquoise lines represent the reference centile curves fitted by the LMS models.

While a variety of methods for longitudinal data have been proposed (e.g.Cole (1994)), software implementations were not readily available up to the time of writing this thesis. An exception was the method by Wei et al. (2006), for which R code was made available by the lead author. This method and its application to the growth data are presented in the remainder of this chapter.

Chapter 5

Quantile Regression Models for Longitudinal Data

The quantile regression approach to constructing growth curves can be extended to incorporate a subject's prior growth history as well as other informative covariates. This facilitates assessing unusual growth patterns, handling longitudinal growth data in such a way that a subject's current growth measurements can be compared directly to their previous measurements. It is also possible for crucial additional information to be incorporated into the model, enabling the subject's measurements to be put into perspective. Recent work on conditional quantile models for growth includes Wei et al. (2006) and Wei (2008), Geraci and Bottai (2006), Muggeo et al. (2012) and Geraci and Bottai (2013).

The major advantage of a model conditioning on age as well as prior growth history and additional information, in comparison to the methods conditioning on age alone, is that it gives a better understanding of a subject's current growth status, allowing comparisons to be made to subjects with similar growth paths. It also provides a complete view of how unmeasured variables are contributing towards the distribution pattern and allow subject-level variables such as parent's heights to be incorporated. Allowance of crucial additional information like this into the model, enables the subject's growth measurements to be put into perspective. In comparison, conditional reference growth curves, which condition on age alone, just provide a valuable snapshot of the dispersion of growth measurements at various values of age.

5.1 Longitudinal Model Methodology

A challenging aspect of most longitudinal growth data is that measurements are observed at unequally spaced time points. One way to address this difficulty is to be adopt a simple first order autoregression model where the AR(1) parameter is specified as a linear function of the time gap between successive measurements as proposed by Wei et al. (2006). Consider the case where we observe measurements, $\{Y_i(t_{i,j}) : j = 1, ..., T_i, i = 1, ..., m\}$ on m individuals. The quantile regression model proposed by Wei et al. (2006) is

$$\begin{gathered} Q_{Y_i(t_{i,j})}(\tau | t_{i,j}, Y_i(t_{i,j-1}), x_i) = g_\tau(t_{i,j}) + \\ [\alpha(\tau) + \beta(\tau)(t_{i,j} - t_{i,j-1})]Y_i(t_{i,j-1}) + x_i^T \gamma(\tau) \end{gathered} \tag{5.1}$$

where the τth conditional quantile function is additively decomposed into a nonparametric trend component, g_τ, an $AR(1)$ component and a partially linear component in the covariate vector x_i. The nonparametric trend component is expressed as a linear expansion in B-splines and is estimated by a suitable quantile regression model. Estimated standard errors of the parametric estimates for the additional covariates can be obtained by the bootstrap method.

5.2 Longitudinal Model for the Growth and Development Study Data

A model of the form (5.1) conditioning on age as well as one prior weight measurement was considered, taking the form

$$\begin{aligned} Q_{Y_i(t_{i,j})}(\tau|t_{i,j}, Y_i(t_{i,j-1}), x_i) = g_\tau(t_{i,j}) + \\ [\alpha_1(\tau) + \beta_1(\tau)(t_{i,j} - t_{i,j-1})]Y_i(t_{i,j-1}) \end{aligned} \tag{5.2}$$

where g is the nonparametric intercept function of the current measurement age estimated for the ith infant, for the seven quantile value by the gender appropriate quantile regression model (QR Model 1 or 2 for females and males respectively), determined in chapter 4. The models given by (5.2) will be referred to as Longitudinal Models 1 and 2, respectively. $Y_i(t_{i,j-1})$ denotes the ith infant's prior weight measurement, and $t_{i,j} - t_{i,j-1} = D_{i,j,1}$ is the age difference between their current and prior weight measurement. The coefficients of the autoregressive term of $Y_i(t_{i,j-1})$ are assumed to be linear functions of this age difference ($D_{i,j,1}$), where α_1 and β_1 are the estimated autoregressive parameters given for the seven quantile values.

The estimated autoregression effect for full-term infants reported in Table 5.1 declines quite dramatically as we move up through the conditional distribution of weight. In the lower tail, dependence on prior weight is quite strong indicating that infants in the lower tail of the weight distribution have a steeper growth profile, while infants in the upper tail have a much flatter profile. A possible explanation (but not necessarily the only explanation) for this is that infants with initial low weights will then put on weight more quickly, which is consistent with the weight catch-up growth hypothesis. Catch-up growth is the acceleration in growth of a majority of "*small for gestational age*" (SGA) infants soon after birth, where SGA refers to infants with a low birth weight with respect to a gestational age read on a reference chart (Han et al., 2010). Generally this catch up occurs during the first 6 months of life.

Table 5.1: Parametric components of Longitudinal Models 1(females) and 2(males), which condition on age as well as one prior weight measurement.

	Females		Males	
	$\hat{\alpha}_1(\tau)$	$\hat{\beta}_1(\tau)$	$\hat{\alpha}_1(\tau)$	$\hat{\beta}_1(\tau)$
0.05	0.865	0.037	0.860	0.035
	(0.260)	(<0.0005)	(<0.0005)	(0.179)
0.1	0.876	0.041	0.867	0.039
	(<0.0005)	(0.6864)	(<0.0005)	(0.002)
0.25	0.872	0.047	0.853	0.046
	(<0.0005)	(<0.0005)	(<0.0005)	(0.135)
0.5	0.853	0.056	0.831	0.052
	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)
0.75	0.804	0.064	0.767	0.060
	(<0.00005)	(<0.00005)	(<0.00005)	(<0.00005)
0.9	0.714	0.075	0.678	0.067
	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)
0.95	0.628	0.078	0.586	0.074
	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)

Estimates of the autoregressive parameters α_1 and β_1 are given for the seven indicated quantiles. P-values are given in brackets.

Longitudinal Models 1 and 2 were used to illustrate screening for the four cases individuals, detailed in section 1.3, who reported unusual growth experiences.

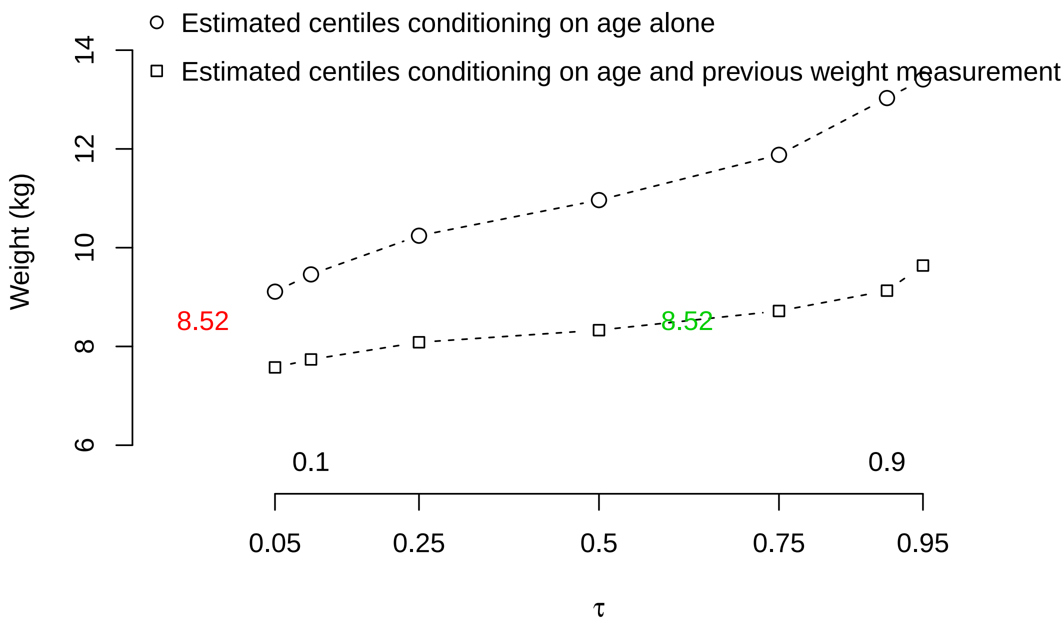
Figure 5.1 illustrates the predictive distributions of the full-term female weights conditional on age alone, produced by the QR Model 1 and those obtained by Longitudinal Model 1. The prediction distributions are the estimated quantiles from these models, deduced for the case infant at their screening age. It can be seen that additionally conditioning on the prior weight measurement reduces the dispersion of the predictive distributions in both cases.

For subject 1500, the observed weight of 8.52 kg at the age of 19.93 months is extremely unusual by the standard of QR Model 1, as discussed in detail in section 4.3, falling below the centile reference curves. However, additionally conditioning on her prior weight measurement at 14.13 months of age, her observed weight of 8.52kg seems reasonable, falling between the 50th and 75th centile reference curve, obtaining an centile estimate of 0.622. It seems reasonable to conclude that although she is slightly underweight for her age there is nothing unusual about the measurement made at the age of 19.93 months when compared to her own growth pattern.

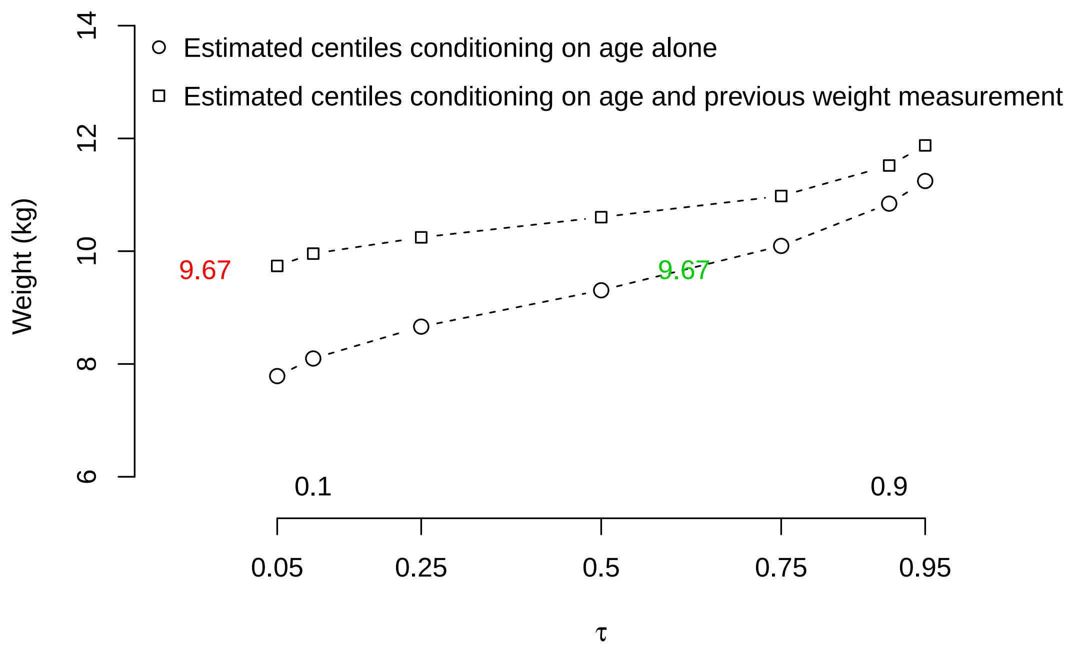
For subject 146 the observed weight of 9.67 kg at the age of 12.03 months is not at all unusual with respect to QR Model 1, falling above the median centile curve. Relative to Longitudinal Model 1 which additionally conditions on her prior weight of 9.1 kg at the age of 8.13 months, however, her weight measurement of 9.67 kg almost 4 months later is extremely unusual, falling below the estimated reference centile curves and attained an centile estimate of 0.098. Given this subject's prior growth history it seems reasonable to conclude that the deceleration in growth experienced by subject 146 is unusual and would therefore be a potential cause of concern and call for closer follow up.

Figure 5.4 illustrates the predictive distributions of QR Model 2 and Longitudinal Model 2 for the two male case infants. Conditioning on the prior weight measurement again reduces the dispersion of the predictive distributions in both cases.

For subject 12 the observed weight of 7.05kg at the age of 5.87 months seems reasonable by the standard of QR Model 2, as discussed in detail in



(a) Subject 146



(b) Subject 1500

Figure 5.1: Screening subjects 1500 and 146 based on QR Model 1 conditioning on age alone and Longitudinal Model 1 which additionally conditions on a prior weight measurement.

Subject	Gender	Screening Age(Months)	Screening Weight(kg)	Prior Age(Months)	Prior Weight(kg)	Centile Estimate
1500	Female	19.93	8.52	14.13	7.84	0.622
146	Female	12.03	9.67	8.13	9.1	0.098
12	Male	5.87	7.05	4.95	7.08	0.087
1799	Male	11.34	13.61	9.34	15	0.076

Table 5.2: Centile estimates for the case infants, obtained at the specified screening ages by the gender appropriate longitudinal model(Longitudinal Model 1 or 2), which conditions on age and a prior weight measurement.

section 4.3, falling just above the 10th centile reference curve. However, additionally conditioning on his prior weight of 7.08kg at the age of 4.95 months, his observed weight of 7.05kg seems extremely unusual, falling below the centile reference curves and obtaining a centile estimate 0.087. This slight reduction in weight experienced by subject 12 highlighted by Longitudinal Model 2 is highly unusual and would therefore be a potential cause of concern and call for closer follow-up.

For subject 1799 the observed weight of 13.61 at the age of 11.34 months is unusually high with respect to QR Model 2, falling above the estimated quantiles for full-term male infants of this age. However relative to Longitudinal Model 2 that additionally conditions on a prior weight of 15 kg at the age of 9.34 months, the 13.61 kg measurement appears to be unusually low when compared to his prior measurements, attaining a centile estimate of 0.076. Given this subjects' prior growth history it seems reasonable to conclude that the sudden drop in weight after the age of 9.34 months may call for further investigation however the slowing down may not be a cause for concern as the subject was at the top of the weight distribution to begin with.

This emphasises the differences in results obtained from using QR Models 1 and 2 and Longitudinal Models 1 and 2, indicating that incorporating subject's prior measurement gives a more comprehensive understanding of the current weight observed.

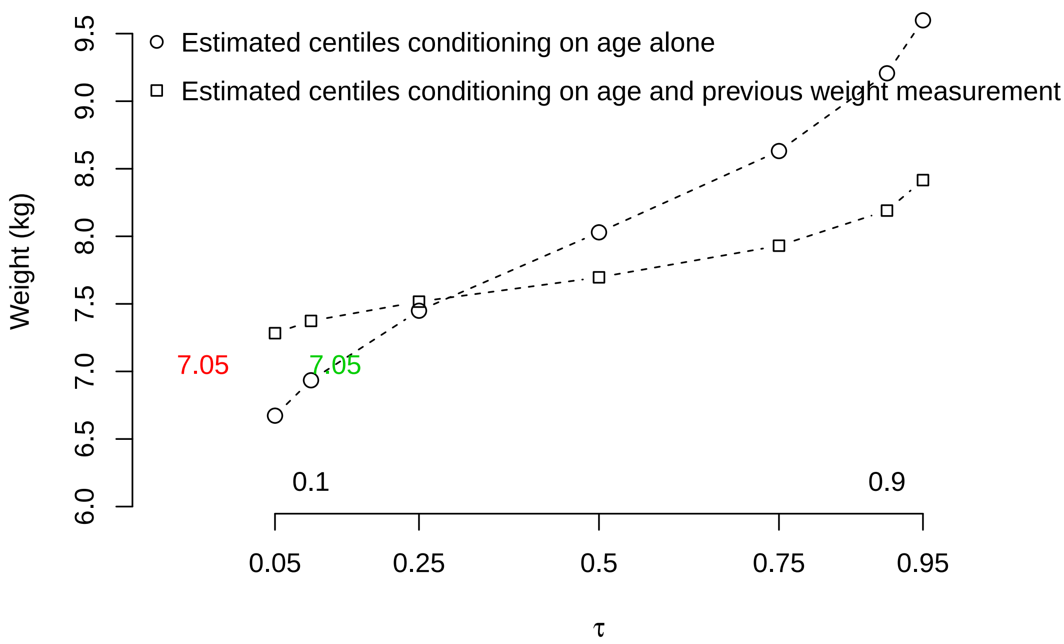
The AR(1) model may be too simple for describing the longitudinal nature of the data. To explore this, a model which conditions on age as well as two previous weight measurement was considered, taking the form

$$Q_{Y_i(t_{i,j})}(\tau|t_{i,j}, Y_i(t_{i,j-1}, t_{i,j-2})) = g_\tau(t_{i,j}) + [\alpha_1(\tau) + \beta_1(\tau)(t_{i,j} - t_{i,j-1})]Y_i(t_{i,j-1}) \\ +[\alpha_2(\tau) + \beta_2(\tau)(t_{i,j} - t_{i,j-2})]Y_i(t_{i,j-2}).$$

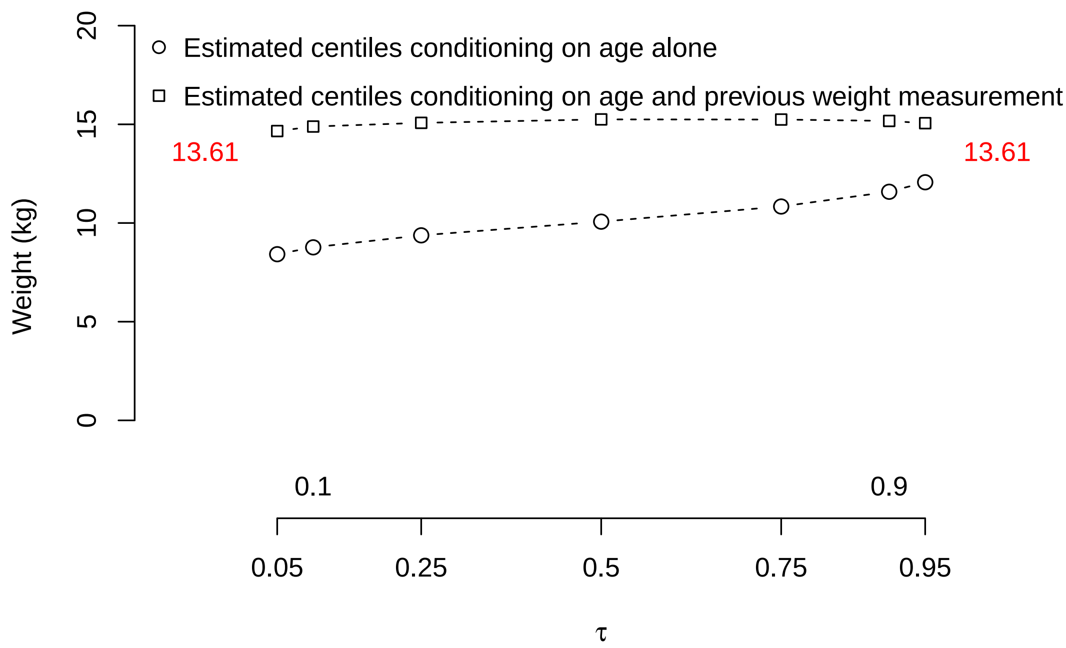
where the terms specified earlier in Longitudinal Model 1 and 2 have the same interpretation. $Y_i(t_{i,j-2})$ denotes the ith infant's second prior weight measurement, and $t_{i,j} - t_{i,j-2} = D_{i,j,2}$ is the age difference between their current and second prior weight measurement. The coefficient of the autoregressive function of $Y_i(t_{i,j-2})$ are linear functions of this age difference ($D_{i,j,2}$), where α_2 and β_2 are the estimated autoregressive parameters given for the seven quantile values. The full-term female and male weight growth longitudinal models of the form (5.3) will be referred to as Longitudinal Models 3 and 4, respectively.

The estimates of the autoregression parameters $\alpha_1(\tau)$ given for the seven indicated quantiles as reported in Table 5.3 are highest at the median quantiles and reducing towards the 5th and 95th centiles, where as the estimates of the autoregression parameters $\alpha_2(\tau)$ and $\beta_2(\tau)$ illustrate that as we move up through the conditional distribution for both genders of full-term infants the estimated autoregression effect declines quite dramatically, as before.

Directly comparing Table 5.2 and Table 5.4 illustrates that the same conclusions would be drawn when screening each case infant based on Longitudinal Models 1 and 3 or 2 and 4, with almost identical centile estimates being observed. This was found to hold in general when screening infants besides the



(a) Subject 12



(b) Subject 1799

Figure 5.2: Screening subjects 12 and 1799 based on QR Model 2 which conditions on age alone and Longitudinal Model 2 which additionally conditions on a prior weight measurement.

Table 5.3: Parametric components of Longitudinal Models 3 and 4, which condition on age as well as two prior weight measurements.

	Females				**Males**			
	$\hat{\alpha}_1(\tau)$	$\hat{\beta}_1(\tau)$	$\hat{\alpha}_2(\tau)$	$\hat{\beta}_2(\tau)$	$\hat{\alpha}_1(\tau)$	$\hat{\beta}_1(\tau)$	$\hat{\alpha}_2(\tau)$	$\hat{\beta}_2(\tau)$
0.05	0.867	0.035	0.024	0.001	0.893	0.036	-0.013	-0.004
	(0.377)	(0.051)	(0.104)	(<0.0005)	(<0.0005)	(0.131)	(<0.0005)	(0.218)
0.1	0.908	0.042	-0.012	-0.004	0.895	0.039	-0.013	-0.001
	(0.008)	(0.135)	(0.003)	(<0.0005)	(<0.0005)	(<0.0005)	(0.073)	(0.162)
0.25	0.918	0.465	-0.023	-0.001	0.914	0.046	-0.035	-0.001
	(<0.0005)	(<0.0005)	(<0.0005)	(<0.00005)	(<0.00005)	(<0.0005)	(0.002)	(<0.00005)
0.5	0.927	0.055	-0.050	-<0.0005	0.909	0.051	-0.044	<0.0005
	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)
0.75	0.926	0.062	-0.087	0.001	0.889	0.058	-0.060	-0.001
	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)
0.9	0.892	0.071	-0.105	0.002	0.831	0.065	-0.069	0.002
	(<0.0005)	(<0.0005)	(<0.0005)	(0.002)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)
0.95	0.827	0.077	-0.108	0.002	0.788	0.068	-0.081	0.003
	(<0.0005)	(<0.0005)	(<0.0005)	(0.002)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)

Estimates of the autoregressive parameters α and β are given for the seven indicated quantiles. P-values are given in brackets.

Subject	Gender	Screening Age(Months)	Screening Weight(kg)	Prior Age(Months)	Prior Weight(kg)	2nd Prior Age(Months)	2nd Prior Weight(kg)	Centile Estimate
1500	Female	19.93	8.52	14.13	7.84	4.52	4.82	0.561
146	Female	12.03	9.67	8.13	9.1	5.84	7.82	0.098
12	Male	5.87	7.05	4.95	7.08	4.03	6.68	0.076
1799	Male	11.34	13.61	9.34	15	4.72	11.17	0.072

Table 5.4: Centile curve estimates for the case infants, deduced at the specified screening ages by the gender appropriate longitudinal model(Longitudinal Model 3 or 4), which conditions on age and two prior weight measurements.

case infants. It can therefore be concluded that the AR(2) model is an option too, but the AR(1) model is chosen for simplicity, requiring less information on the infant's growth history.

In addition to an infant's prior measurements, other factors might play a role. To illustrate this a model which conditions on age, one prior weight measurement and incorporates a linear term for average parental height, $\gamma(\tau)x_i$, was considered. The parental height value, available only for infants included in the follow up study performed five years after the initial Growth and Development study, was calculated as

$$(\text{Father's Height} + \text{Mother's Height})/2,$$

and the model fitted therefore takes the form

$$Q_{Y_i(t_{i,j})}(\tau|t_{i,j}, Y_i(t_{i,j-1}), x_i) = g_\tau(t_{i,j}) +$$
$$[\alpha(\tau) + \beta(\tau)(t_{i,j} - t_{i,j-1})]Y_i(t_{i,j-1}) + \gamma(\tau)x_i,$$

where x_i is the average parental height for the ith infant and again the terms specified earlier in Longitudinal Model 1 and 2 have the same interpretation. The model of the form (5.3) will be referred to as Longitudinal Models 5 and 6 for full-term female and male weight growth longitudinal models, respectively.

The estimated autoregression effect for full-term infants reported in Table 5.5 again declines quite dramatically as we move up through the conditional distribution of weight. In the lower tail dependence on prior weight is quite strong indicating that infants in the lower tail of the weight distribution have a steeper growth profile, while infants in the upper tail have a much flatter profile. The effect of average parental height is weaker in the lower tail and more strongly significant for both female and male infants in the upper tails.

Screening based on Longitudinal Models 5 and 6 which condition on age, a prior weight measurement and average parental height, was performed on

Table 5.5: Parametric components of Longitudinal Models 5 and 6, which condition on age as well as a prior weight measurement and average parental height.

	Females			Males		
	$\hat{\alpha}_1(\tau)$	$\hat{\beta}_1(\tau)$	$\hat{\gamma}_1(\tau)$	$\hat{\alpha}_1(\tau)$	$\hat{\beta}_1(\tau)$	$\hat{\gamma}_1(\tau)$
0.05	0.856	0.036	0.003	0.880	0.039	0.005
	(<0.0005)	(<0.0005)	(0.130)	(0.026)	(0.547)	(0.255)
0.1	0.870	0.039	0.003	0.879	0.039	0.005
	(0.085)	(0.002)	(0.042)	(0.217)	(0.645)	(0.622)
0.25	0.893	0.046	0.004	0.889	0.046	0.005
	(<0.0005)	(0.066)	(<0.0005)	(<0.0005)	(0.147)	(0.006)
0.5	0.887	0.054	0.007	0.872	0.052	0.006
	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(0.001)
0.75	0.850	0.062	0.010	0.821	0.059	0.008
	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)
0.9	0.802	0.070	0.011	0.755	0.067	0.013
	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(0.003)
0.95	0.726	0.071	0.018	0.705	0.068	0.009
	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(<0.0005)	(0.071)

Estimates of the autoregressive parameters α and β, and the average parental height effect, γ are given for the seven indicated quantiles. P-values are given in brackets.

		Measurement							
	Subject	1	2	3	4	5	6	7	8
Age(Months)	28	Birth	1.11	1.80	3.02	4.10	5.25	9.15	21.61
Weight(Kg)	28	3.29	4.96	5.58	6.46	7.37	8.32	10.1	12.35

Table 5.6: Weight Measurements of Subject 28, a full-term female infant.

the individuals with unusual growth experiences whose parental heights had been recorded. Of the four case infants initial described in section 1.3 only subject 12, a male full-term infant, had parental height information recorded. Therefore an additional full-term female case infant, Subject 28, who observed an unusual growth pattern as well as having her parental information recorded was selected for screening from the Growth and Development Study data.

Figure 5.3 illustrates that subject 28 was an average weight for a full-term female infants at birth, showing an initial growth spurt in her first month. She then grew steadily thereafter up to the age of roughly 9 months. However by her next weight measurement, the age at which screening is considered, her weight does not appear to have increased substantially. Table 5.6 giving more precise details on these measurements indicates that at the age of 21.61 months, almost 13 months since her last measurement, subject 28 is reported to have gained only 2.25 kg.

Figure 5.4a illustrates the predictive distributions of Longitudinal Model 1 which conditions on age and a prior weight measurement and Longitudinal Model 5 which additionally contains a covariate for average parental height. Subject 28's weight measurement at 21.61 months is relatively unusual by the standard of both growth charts, falling between the 5th and 10th centile reference curves. Table 5.7 detailing her centile estimates at the screening age based on Longitudinal Models 1 and 3 demonstrates the similarity in

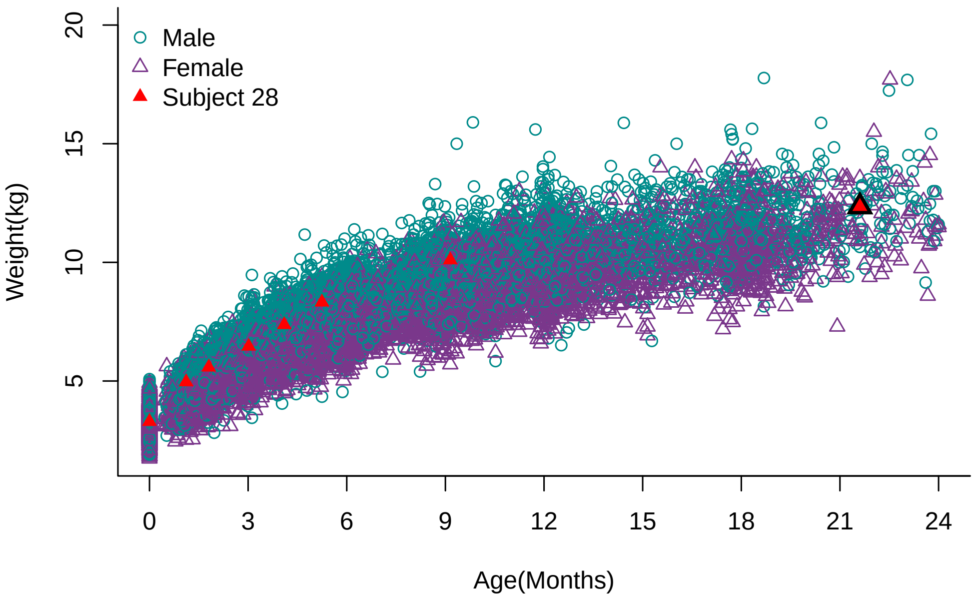


Figure 5.3: Plot of weight measurements of full-term Infants in the Growth and Development Study by gender. Highlighted are the weight measurements observed for Subject 28, with the point bordered in black denoting the observation at which the screening decision is considered.

Subject	Gender	Average Parental Height(PH)	Not Containing PH	Containing PH
28	Female	163.83	0.084	0.066
12	Male	165.1	0.076	0.083

Table 5.7: Centile estimates for the case infants, deduced at the specified screening ages, by Longitudinal Models 1 and 5 or 2 and 6).

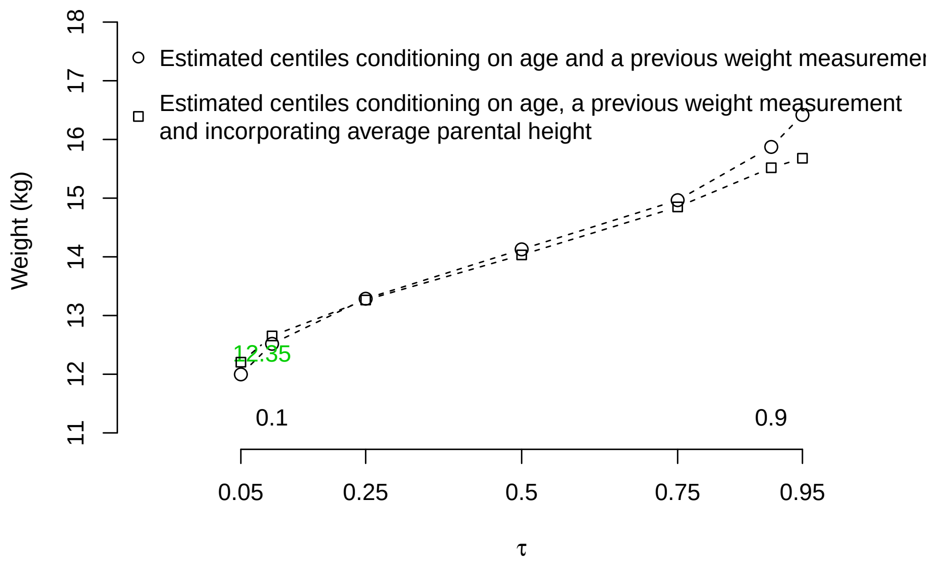
results obtained.

Figure 5.4b illustrates the predictive distributions of Longitudinal Model 2 which conditions on age and a prior weight measurement and Longitudinal Model 6 which additionally contains a covariate for average parental height. Subject 12's weight measurement at 5.87 months is unusual by the standard of both growth charts, falling below the 5th reference centile curve. Again Table 5.7 detailing his centile estimates at the screening age, based on both these models, demonstrates the similarity in results obtained.
This suggests that adding a linear trend component for average parental height does not in this case have any influence on the conclusions drawn. This is likely to be because parental height is not a very strong predictor of weight and would perhaps be a more relevant additional covariate to add when modelling height growth charts.

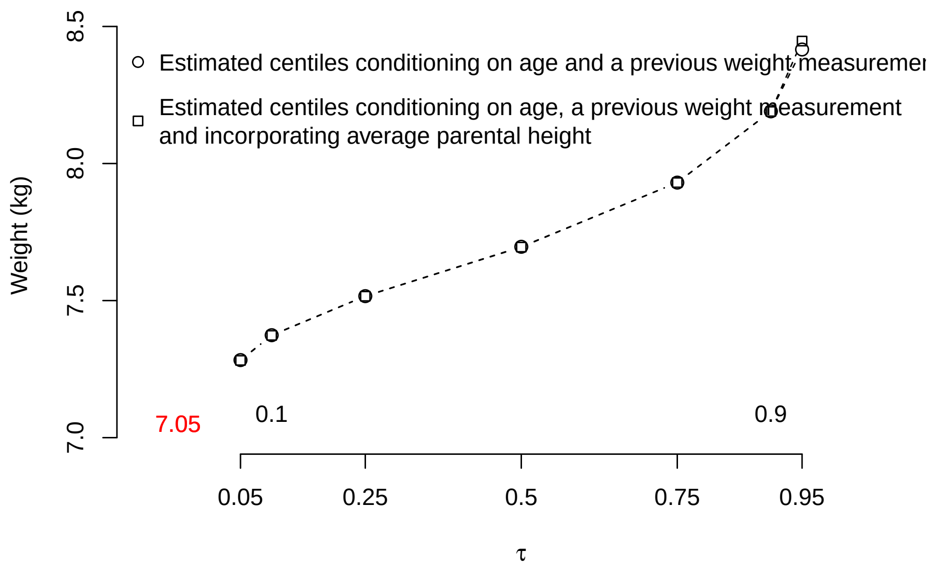
5.3 Summary

To conclude, a model conditioning on age and a previous weight measurement appears to allow satisfactory monitoring of a full-term infant's weight at a particular age, whilst considering their prior growth history.

This longitudinal model which is an extension of the quantile regression model which performs smoothing using a B-spline basis function, has similar



(a) Subject 28



(b) Subject 12

Figure 5.4: Screening subjects 28 and 12 based on Longitudinal Model 1 or 2, which conditions on age and a prior weight measurement, and Longitudinal Model 5 or 6 which additionally incorporates a covariate for average parental height.

properties and limitations as those discussed in chapter 4 for QR Models 1 and 2. Some of these limitations may potentially be resolved by instead modelling the nonparametric trend component in the longitudinal model by a P-spline quantile regression model which has additional options of non-crossing and monotonicity constraints. This is because, as seen in chapter 4, PQR Models 1 and 2 can construct relatively robust reference centile curves in covariate intervals when there are limited observations.

Chapter 6

Web Application

In order to demonstrate how some of the methods described in the previous chapters can be used in practice, a user friendly interactive web application was designed using the **R** package **shiny**. The application allows monitoring of new infant's weights based on reference growth charts modelled on the Growth and Development Study data, composed via several of the modelling approaches discussed in this thesis. The modelling approaches included in the growth monitoring web application are the LMS Models 1 and 2, QR Models 1 and 2 and Longitudinal Models 1 and 2. However, currently accessibility to this web application is restricted to University of Glasgow computers only, because of confidentiality of the data. Such an application can, in principle, run from any web browser without the need for any specialised statistical software to be installed on the user's computer.

When first accessing the web application, the start up page, as displayed in Figure 6.1, requires the infant's sex, date of birth, date of weight measurement and subsequent observed weight (kg) at this date to be imputed. This information is used to deduce the infant's screening age, the age at which the weight measurement is recorded. The web application defaults to a female case infant and date 1st January 2000. For the purpose of illustration, suppose that the application is used to monitor a female infant, with a date of birth of 20th February 2005 and observed screening date of 3rd

December 2005, at which she weighs 8 kg. Once the infant's information has been entered, the user has the option to choose between reference growth charts fitted using the LMS or the quantile regression modelling approach or both. The LMS and quantile regression reference growth curves are presented by black and turquoise curves, respectively. As seen in Figures 6.2, 6.3 and 6.4 both of these modelling approaches produce reference growth charts with the infant's weight measurement at the screening age symbolised by a red triangle. This allows monitoring of the infant's growth visually, however additionally a centile estimate is shown for the weight measurement at the specified screening age when the LMS modelling approach is implemented. The quantile regression approach returns centile estimates for each of the seven reference centiles(the 5th, 10th, 25th, 50th, 75th, 90th and 95th) at the specified screening age, which can be used to determine the positioning of the infant's weight in comparison to the reference growth chart.

If additional information on the infant's prior weight measurements is available this can be included also, with the date of the previous weight measurement and observed weight needing to be specified. This consequently allows modelling using Longitudinal Model 1 or 2, which composes a growth chart which conditions on age and a prior weight measurement. This approach outputs centile estimates for each of the seven reference centile curves at the specified screening age. If modelling using the quantile regression approach is also specified, a direct visual comparison between these centile estimates and those from QR Model 1 or 2 can be made. These are presented by turquoise and purple dashed lines, for the centile estimates produced by the quantile regression models and longitudinal models, respectively. The positioning of the observed weight at the screening age in comparison to both of these lines is also shown, illustrating were the infant's weight measurement falls within the centile estimates. If this is within the reference range, in other words between the 5th and 95th centiles, this is shown by the observed weight value in green, otherwise it is coloured red. Figure 6.5 illustrates such an output, for

Figure 6.1: Growth and Development Study data web application start up page.

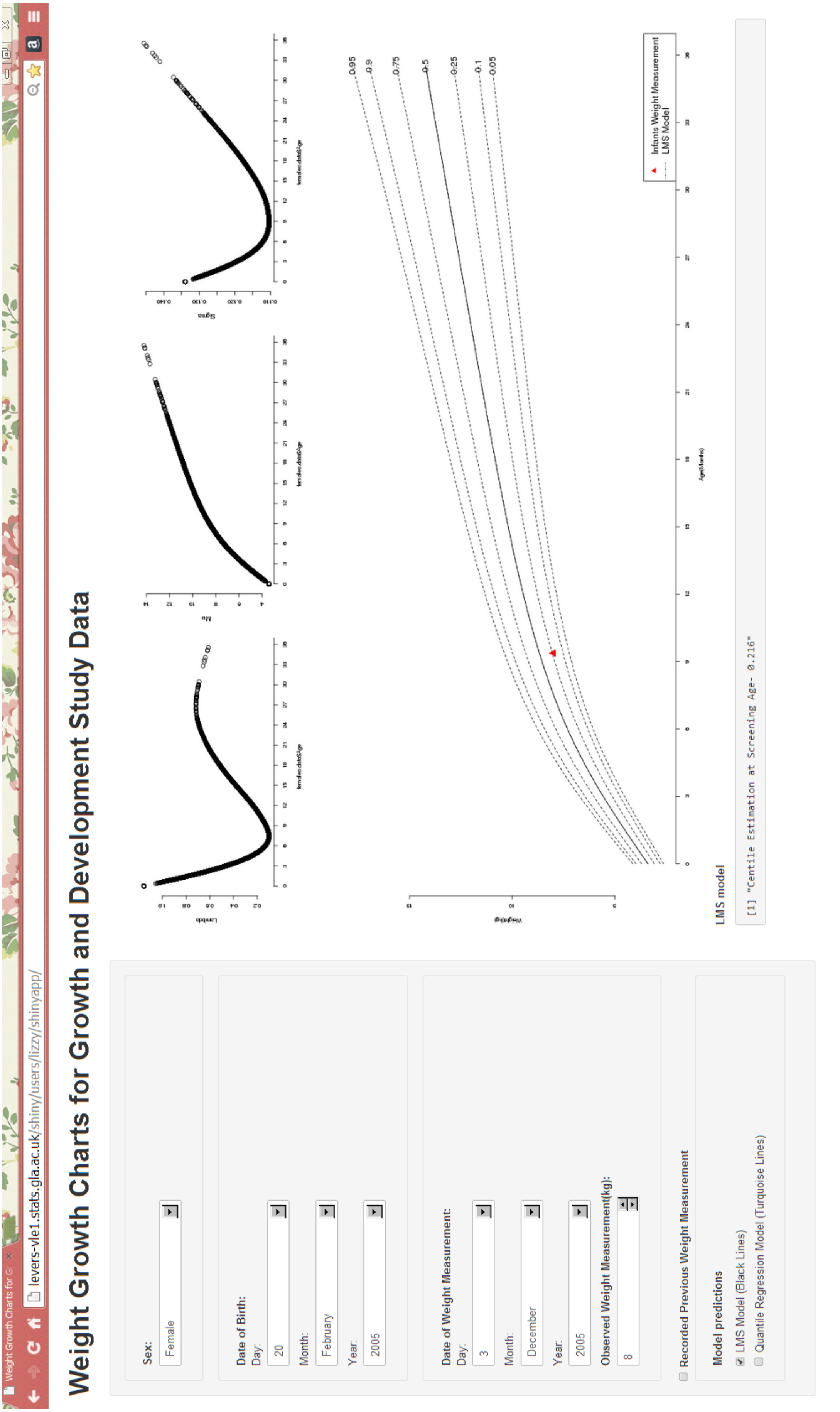


Figure 6.2: Growth chart web application with LMS modelling approach specification.

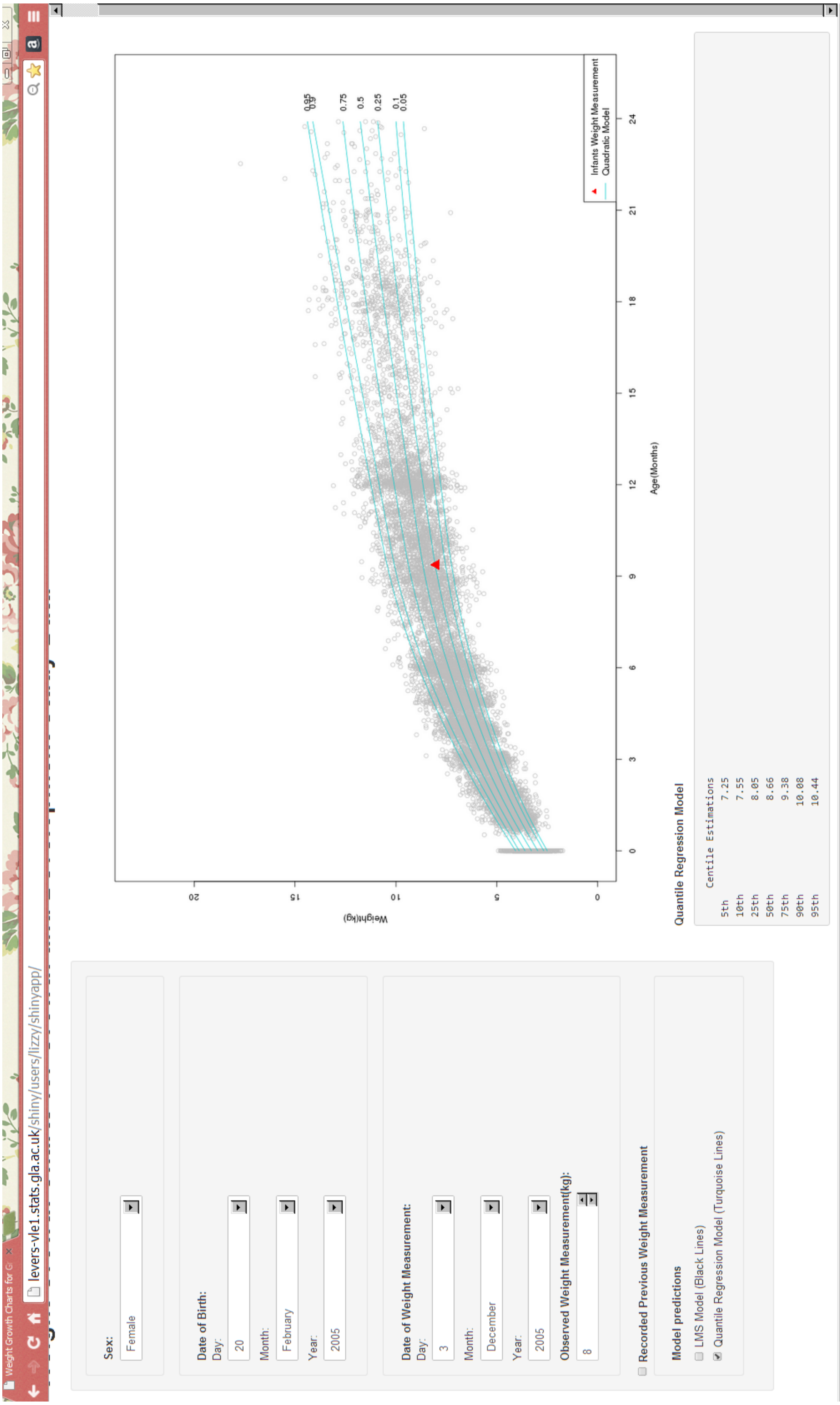


Figure 6.3: Growth chart web application with quantile regression modelling approach specification.

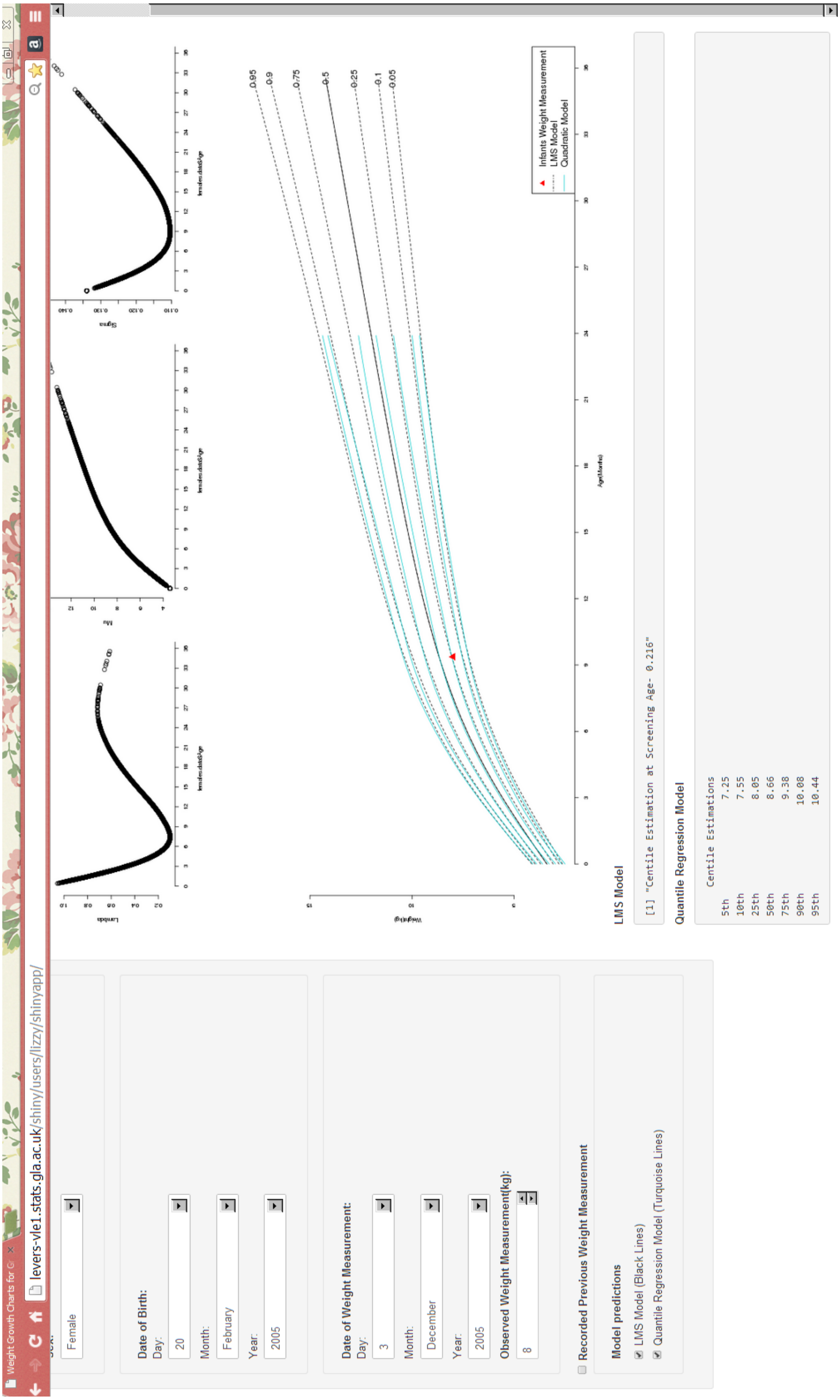


Figure 6.4: Growth chart web application with LMS and quantile regression modelling approach specification.

the example case used throughout this chapter, if her prior weight is defined as 6kg on 18th April 2005.

Specification of the LMS and/or quantile regression modelling approaches (LMS Model 1 or 2 and or QR Model 1 or 2), when a prior weight measurement is available, has no influence on the growth charts composed as neither of these approaches can incorporate this additional information. The charts do however now have both the weight measurement at the screening age as well as that observed at the prior weight measurement shown on the curves, presented by a red triangle and pink circle, respectively. Figure 6.6 gives a representation of this for the illustrative example case of a female infant.

The web application designed in this research is purely for illustration purposes of the different modelling methods of composing growth charts for the Growth and Development study data. Further improvements could be made to allow the incorporation of further modelling approaches, enhancements to it visually and minimisation of the time taken for the web application to run once specification of the modelling approach has been made.

Such an application could be used by health professionals, advising them on an infant's growth and helping them detect unusual growth patterns that could potentially merit further attention and monitoring. Clearly the charts composed via the Growth and Development Study Data are not suitable for monitoring all infants internationally as the data is very specific to a particular population and is an old data set. However, if appropriate longitudinal datasets were available, international or country-specific reference growth charts could be composed using the outlined approaches, and a more concise, user friendly web application could be designed which could be used by health professionals for monitoring infants' growth.

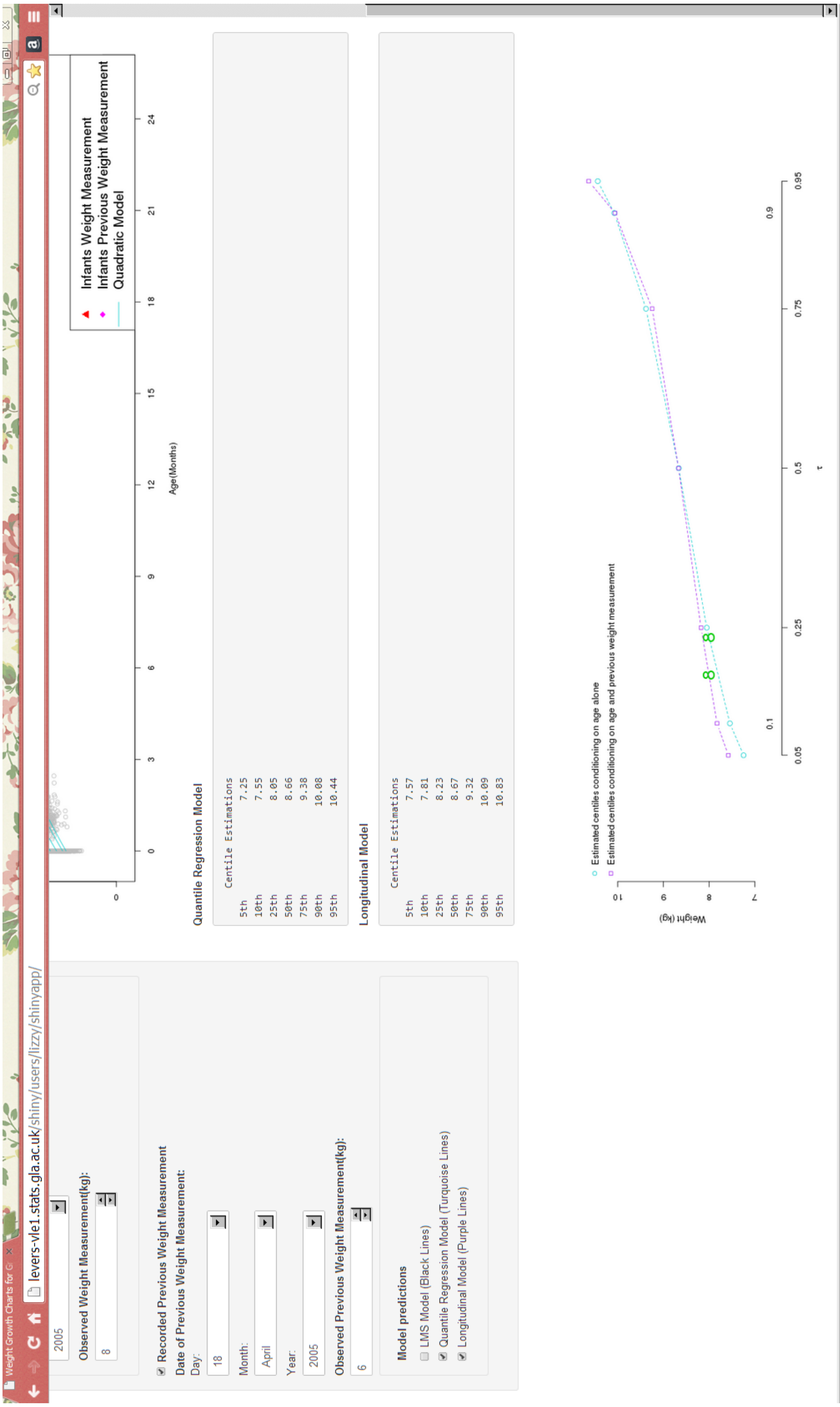


Figure 6.5: Growth chart web application with quantile regression and longitudinal model specification, when a prior weight measurement information is available.

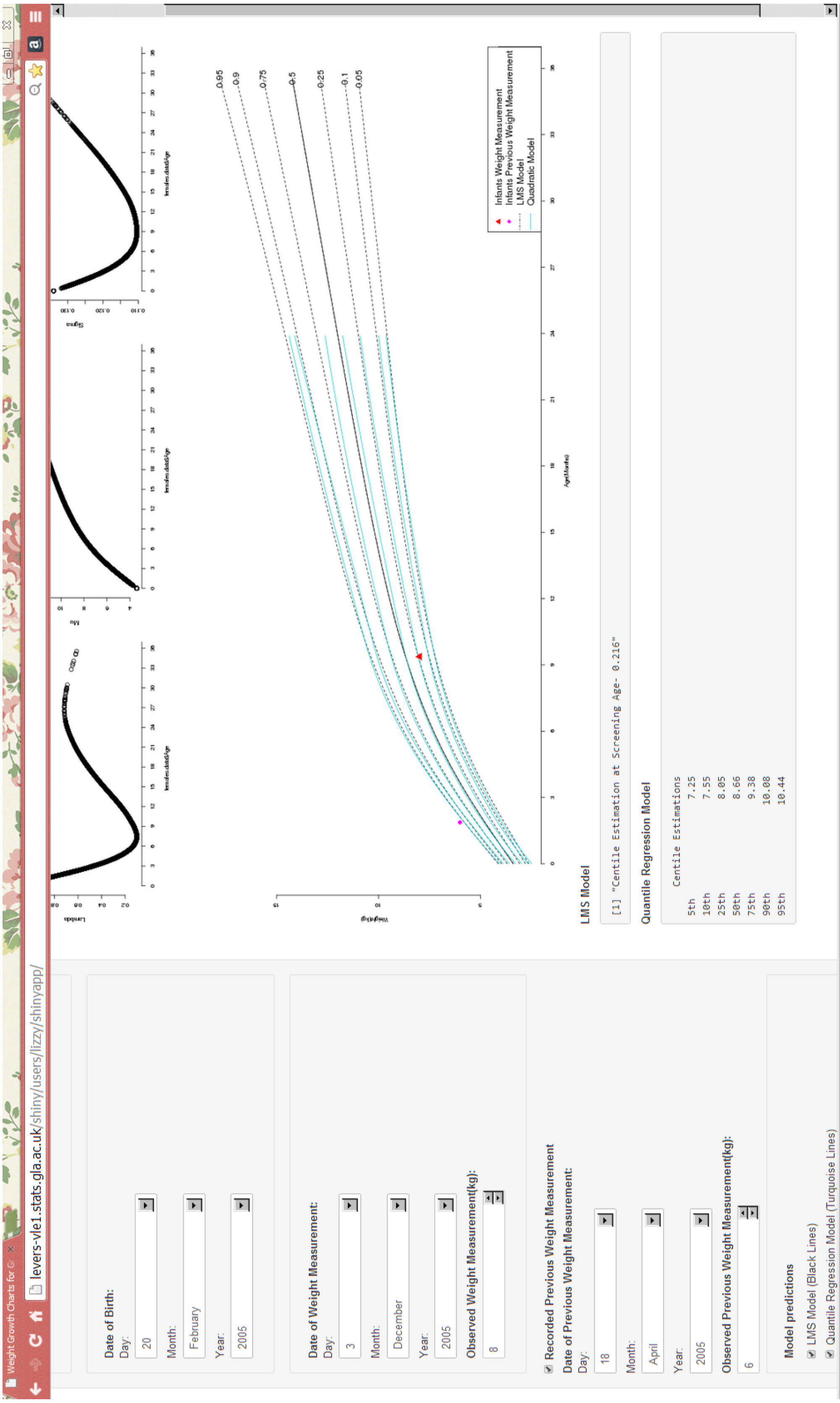


Figure 6.6: Growth chart web application with LMS and Quantile Regression modelling approaches specification, when a prior weight measurement information has been inputed.

Chapter 7

Concluding Remarks

This chapter briefly explores possible areas for further work, gives a summary of the main findings of the research conducted, discusses limitations and future research directions.

7.1 Conditional Gain SD score

One more approach to constructing growth gain references that was explored briefly, is that of conditional gain SD score measurements. These allow conditionality on a time covariate, often age, as well as a prior growth measurement, by looking at the change in SD scores, see page 38. This approach calculates growth gain in term of the change in the subject's current growth SD score (SDS_2) at time t_2 and the value computed from their growth SD score on their previous measurement (SDS_1) at time t_1 (Cole, 1995). The conditional gain SD score is therefore defined as

$$SDS\text{gain} = SDS_2 - SDS_1. \tag{7.1}$$

The World Health Organisation's Child Growth Velocity Standards for weight uses this approach for construction of these standards. These contain gender-specific velocity curves for one, two, three, four and six month weight increments conditional on age, again based one the WHO Multicentre Growth Reference Study, a study discussed in detail in chapter 1. These curves demonstrate a change in the weight value expressed in units per time period (g/time). The tables contain centile estimates for ages 0-1, 1-2, ..., 11-12 months (de Onis et al., 2009), as shown in Figure 7.1a for example for the female one-month increment curves, whereas for instance for the female two-month increment curves, centile estimates are given for ages 0-2, 1-3, ..., 22-24 months as illustrated in Figure 7.1b. Estimates are also provided in other accessible tables in SD form. Clinicians can use these tables to monitor infants, looking at the increment table which most closely approximates the interval over which the infant's weight is being monitored.

The WHO Child Growth Velocity standards for weight could lead to earlier identification of weight growth problems than the simpler WHO weight-for-age child growth standards but are not yet used as widely as a generic approach for monitoring infants. This is partly due to the scarcity of appropriate longitudinal datasets available for composing the standards, which results in far fewer velocity references (de Onis et al., 2009).

The SD score gain methodology can be extended to adjust not only for the age covariate but also for regression to the mean. The theoretical basis for this approach is the expectation that over time subjects' growth measurements, such as weight, drift towards the median from the tails of the distribution. The conditional gain SD score, adjusted for regression to the mean is defined as

$$\text{SDSgainAdjusted} = \frac{SDS_2 - (r.SDS_1)}{\sqrt{1-r^2}} \tag{7.2}$$

where r is the correlation coefficient between SDS_1 and SDS_2 which determines the expected slope of change in the subject's growth between the two times of measurement (Cole, 1995). When measurements are observed at equally spaced time intervals, r is simply the correlation amongst SD scores (SDS_1 and SDS_2) for the subset of subjects whose growth measurements were observed at both time t_1 and t_2.

1-month weight increments GIRLS

Birth to 12 months (percentiles)

					Percentiles (weight increment in g)										
Interval	L	M[a]	S	δ	1st	3rd	5th	15th	25th	50th	75th	85th	95th	97th	99th
0 – 4 wks	0.7781	1279.4834	0.21479	400	280	388	446	602	697	879	1068	1171	1348	1418	1551
4 wks – 2 mo	0.7781	1411.1075	0.19384	400	410	519	578	734	829	1011	1198	1301	1476	1545	1677
2 – 3 mo	0.7781	1118.0098	0.19766	400	233	321	369	494	571	718	869	952	1094	1150	1256
3 – 4 mo	0.7781	984.8825	0.20995	400	133	214	259	376	448	585	726	804	937	990	1090
4 – 5 mo	0.7781	888.9803	0.22671	400	51	130	172	286	355	489	627	703	833	885	983
5 – 6 mo	0.7781	801.3910	0.24596	400	-24	52	93	203	271	401	537	611	739	790	886
6 – 7 mo	0.7781	744.3023	0.26515	400	-79	-4	37	146	214	344	480	555	684	734	832
7 – 8 mo	0.7781	710.6923	0.28409	400	-119	-44	-2	109	178	311	450	526	659	711	811
8 – 9 mo	0.7781	672.6072	0.30106	400	-155	-81	-40	70	139	273	412	489	623	675	776
9 – 10 mo	0.7781	644.6032	0.31676	400	-184	-110	-70	41	110	245	385	464	598	652	754
10 – 11 mo	0.7781	633.2166	0.33208	400	-206	-131	-89	24	95	233	378	459	598	653	759
11 – 12 mo	0.7781	631.7383	0.34627	400	-222	-145	-102	15	88	232	383	467	612	670	781

WHO Growth Velocity Standards

[a] The L, M, S values provided are estimated based on the modelling of the shifted observations (i.e. by the addition of delta to the actual increment) which explains the difference (equals to delta) in value between the "M" and the 50[th] centile values.

(a) One-month increments

2-month weight increments GIRLS

Birth to 24 months (percentiles)

					Percentiles (weight increment in g)										
Interval	L	M[a]	S	δ	1st	3rd	5th	15th	25th	50th	75th	85th	95th	97th	99th
0 – 2 mo	0.4599	2497.0406	0.18000	600	968	1128	1216	1455	1604	1897	2210	2386	2696	2820	3062
1 – 3 mo	0.3294	2314.2285	0.17612	600	890	1030	1107	1317	1450	1714	2000	2163	2452	2569	2799
2 – 4 mo	0.3128	1907.0116	0.17761	600	625	740	804	978	1088	1307	1545	1681	1922	2020	2213
3 – 5 mo	0.3560	1673.5778	0.18421	600	451	556	615	773	874	1074	1290	1413	1632	1720	1894
4 – 6 mo	0.4264	1482.7466	0.19524	600	295	395	450	600	695	883	1085	1200	1403	1486	1646
5 – 7 mo	0.5002	1342.3734	0.20864	600	170	267	321	468	560	742	938	1048	1243	1321	1473
6 – 8 mo	0.5699	1251.4869	0.22315	600	76	175	229	377	469	651	846	955	1147	1223	1372
7 – 9 mo	0.6268	1181.4135	0.23586	600	3	103	157	306	399	581	775	883	1072	1147	1293
8 – 10 mo	0.6730	1116.8192	0.24680	600	-59	40	95	243	336	517	708	814	999	1073	1215
9 – 11 mo	0.7102	1078.3961	0.25656	600	-104	-3	53	203	297	478	670	776	960	1033	1174
10 – 12 mo	0.7382	1058.4112	0.26494	600	-135	-31	26	179	274	458	652	759	944	1018	1159
11 – 13 mo	0.7605	1040.8737	0.27292	600	-163	-57	1	157	254	441	637	745	932	1005	1147
12 – 14 mo	0.7762	1027.9459	0.28011	600	-185	-78	-19	140	238	428	626	736	924	999	1142
13 – 15 mo	0.7864	1019.6870	0.28705	600	-204	-95	-35	127	227	420	621	732	924	999	1144
14 – 16 mo	0.7913	1016.4898	0.29343	600	-219	-108	-47	118	220	416	622	735	930	1007	1154
15 – 17 mo	0.7922	1017.5335	0.29961	600	-231	-118	-55	112	216	418	627	743	943	1021	1172
16 – 18 mo	0.7902	1017.2241	0.30592	600	-243	-128	-64	106	212	417	631	750	954	1035	1189
17 – 19 mo	0.7866	1012.8511	0.31201	600	-255	-139	-75	97	205	413	631	751	959	1041	1199
18 – 20 mo	0.7827	1007.2711	0.31824	600	-267	-151	-86	88	196	407	628	751	962	1046	1206
19 – 21 mo	0.7795	1001.8324	0.32415	600	-279	-162	-97	79	188	402	626	750	965	1050	1213
20 – 22 mo	0.7771	993.3265	0.33014	600	-291	-174	-109	67	178	393	620	745	963	1049	1214
21 – 23 mo	0.7755	980.7096	0.33605	600	-305	-189	-124	53	164	381	608	735	954	1040	1207
22 – 24 mo	0.7743	967.2057	0.34166	600	-318	-202	-137	39	150	367	596	723	942	1029	1197

WHO Growth Velocity Standards

(b) Two-month increments

Figure 7.1: WHO Child Growth Velocity Standards for female infants.

An attempt was made to apply the SD score gain approach with adjustment for regression to the mean to the Growth and Development Study Data, but as the data has unequally spaced measurements this did not yield satisfactory correlation estimates in some cases. As a result the conditional gain SD scores, adjusted for regression to the mean are not always meaningful or reliable. The SD scores used in the calculations were those ascertained using the gender-specific weight growth chart composed by LMS Models 1 and 2. The correlation coefficients were computation by only incorporating pairs of SD scores when their corresponding weight measurements were observed at most 0.5 months from the desired ages t_1 and t_2, from weight measurements observed on infants of the same sex as the infant whose correlation coefficient was being calculated. Figure 7.2 illustrates an example where this computation method for calculating the correlation coefficient fails, where the red points denoting subject 1799's SDS_1 and SDS_2 are particularly unusual in comparison to the other SD scores which are incorporated into the correlation coefficient computation, leading to an unrealistic positive correlation coefficient estimation.

The centile estimations for the four case infants at their screening ages, obtained from the conditional gain SD score and the gender appropriate longitudinal model(Longitudinal Model 1 or 2), as illustrated in Table 7.1, are similar to each other. This implies that these estimates are more sensible and are an appropriate approach of monitoring.

To conclude, the standard approach of unadjusted conditional gain SD score, which is used by WHO for their Child Growth Velocity Standards for weight, seems to produce feasible centile and SD score estimates which broadly agree with the quantile estimates from Longitudinal Models 1 and 2, which incorporates prior growth history. Both approaches appear to give sensible results and can be useful for monitoring changes in growth patterns. The regression to the mean adjustment for SD score gain is an interesting extension which

Subject	Gender	Screening Age	SDS_1	SDS_2	r	SDSgain	SDSgainAdjusted
146	Female	12.03	0.913	0.187	-0.053	-0.726	0.235
						(0.234)	(0.593)
1500	Female	19.93	-2.064	-2.112	0.0596	-0.048	-1.992
						(0.481)	(0.023)
12	Male	5.87	-0.460	-1.100	-0.048	-0.640	-1.232
						(0.261)	(0.131)
1799	Male	11.34	4.408	2.747	0.0103	-1.661	2.702
						(0.048)	(0.997)

Table 7.1: Conditional gain SD scores, not adjusted and adjusted for regression to the mean, calculated for the four case infants at their screening ages. Centile estimates are given in brackets.

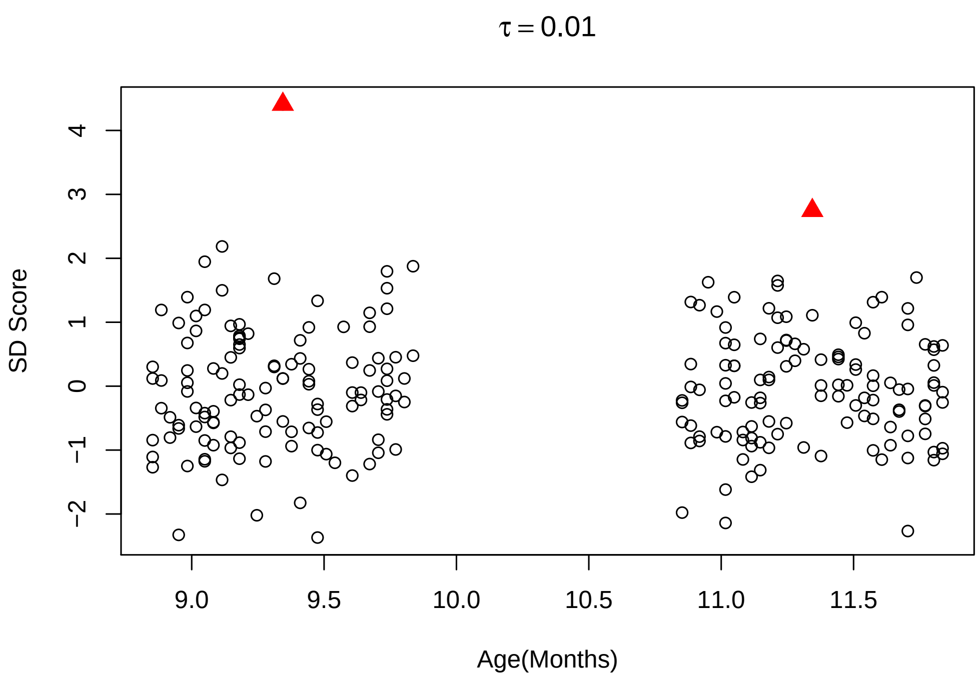


Figure 7.2: Correlation coefficient computation for male case infant subject 1799, for the first centile estimates.

could be explored further in future research, subject to availability of a suitable data set with equally spaced measurements.

The conditional gain SD score approach requires an LMS model to be initially fitted to determine the SD scores, so the limitations associated with this modelling technique, as discussed in chapter 3 still exist. However the advantage of this approach to a longitudinal model approach is that it is able to estimate across the whole age range of the study, determining the e.d.f values is relatively straighthforward, and appears to be a very stable approach in general.

7.2 Summary and Conclusions

This research, which studied current statistical methods for composing growth charts, found each of the proposed methods to be satisfactory when modelling the Growth and Development Study data.

The LMS model approach, discussed in chapter 3, which can be implemented in **R** by the **lmsqreg** package and the **lms** function in the **GAMLSS** package, appears to be a very reliable method for composing gender-specific growth charts for full-term infants' weights, modelled on the Growth and Development Study data. Although a robust approach exists for determining the most desirable e.d.f values for the L, M and S curves, this modelling approach also has some limitations. These include its inability to allow additional covariates to be included in the model as well as assuming that the response variable is normalised once a suitable power transformation has been performed on it. This assumption, which appeared reasonable for the Growth and Development Study Data, may not be a feasible assumption for all growth datasets.

The nonparametric quantile regression modelling approach, detailed in chapter 4, adjudicates this limitation, having the ability to estimate quantiles given a time covariate without the usual distribution assumptions. One limitation associated with this modelling approach is that no automatic procedure for selecting the smoothness parameters is currently implemented in the available software, so determining a suitable smoothing parameter choice can be a lengthy procedure. The quantile regression model approach of composing growth charts, which smooth using B-splines, as discussed in section 4.3, appears to perform reasonably well on the Growth and Development Study Data. When modelling across the entire age range, however, the gender-specific growth charts composed by this approach for the full-term infants' weight measurements, illustrate unrealistic fluctuations and often depleting reference centile curves at later months, where fewer observation were observed for which the data is sparse. Furthermore while the LMS model approach implicitly leads to non crossing curves via a scaling function, the quantile regression model approach needs some extra work to return non-crossing quantile curves. Discussion of this is given in Muggeo et al. (2012). The quantile regression model approach requires careful selection of the smoothing parameters to avoid this issue.

Some of these issues, can be addressed my modelling using the **R** package **quantregGrowth**, which allows growth chart composition using penalised B-splines in a quantile regression model with a non-crossing constraint. This model is discussed in section 4.5. In particular, allowance of smoothing using P-splines appears to have a very positive influence of the growth charts produced, eliminating the problem of fluctuating reference centile curves at the later months, which is present when smoothing using B-splines. The **quantregGrowth** package is still under development so a limitation associated with this package is that screening based on the growth charts composed is currently not implemented.

Longitudinal models, described in chapter 5, which are simply an extension of the quantile regression model approach, allow incorporation of individuals' prior growth history and other informative covariates. The limitation associated with the LMS and quantile regression model approaches, which only condition on age is hence resolved. Longitudinal models, therefore, crucially allow assessment of an individual's growth patterns, in such a way that their current growth measurement is compared directly to their own previous measurement and comparison to individuals in the reference population with similar growth patterns are made. In contrast, approaches conditioning only on age allow monitoring of an individual's measurement at a specific screening age, comparing it directly to the reference population's growth measurements at that value of the covariate. Applying the longitudinal model approach to the Growth and Development Study data appears to allow satisfactory monitoring of a full-term infants' weight at a particular age, whilst considering their prior growth history, giving sensible conclusions.

Looking at the application of the methods in studying the weights of the four case infants, from the Growth and Development Study, at their screening ages, the following conclusions were reached. Although the growth charts constructed using the LMS and nonparametric quantile regression modelling approaches lead to almost identical conclusions for each case infant, the longitudinal models were able to put these weight measurements into more context by drawing on prior weight. On balance the longitudinal approach was therefore found to be the more informative for following individual infants' growth

The web application designed in this research, which is purely for illustration purposes of the different modelling methods of composing gender-specific growth charts for monitoring full-term infants' weights on the Growth and Development study data, shows the efficiency of the shiny package at producing user-friendly web applications and the potential of such a web application to be used as an alternative to paper-based growth charts.

The conditional gain SD score approach is an alternative approach to constructing growth gain references, allowing conditionality on age as well as a prior growth measurement, by looking at the change in SD scores. The conclusions drawn from the standard conditional gain SD scores seemed to be feasible although SD scores may not be as easy to interpret as quantiles.

7.3 Further Work

Further work which would be considered, includes the extension of the longitudinal models to allow specification of the non-parametric trend component by penalised B-splines subject to a monotonicity and non-crossing constraint, as implemented in the package **quantregGrowth**. This would hopefully lead to an improvement in the reliability of conclusions drawn, as this research has undoubtedly shown the advantages of modelling using this package. To compose growth charts by such a model, software development in **R** would be required, combining and extending the **R** code currently available on the longitudinal models to include monotonically constrained non-crossing P-splines. A **predict** function could also be designed in **R**, to allow screening of an individual's growth measurement based on the growth charts composed by this model.

Further investigation of the conditional gain SD score approach with application to longitudinal data studies where measurements were observed at equally spaced intervals in order to determine the effectiveness of this approach when adjusting for regression to the mean, would also be interesting additional work.

Application to current and relevant longitudinal data sets of the model approaches outlined in this research to construct growth charts would also be considered if appropriate datasets were available. More concise, user friendly web applications could be designed, containing such growth charts, which could be used by health professional for monitoring individuals' growth.

Chapter 8

Appendix

Model	Model Description	Gender	Chapter Introduced
LMS Model 1	LMS model with L, M and S curves smoothed by natural cubic spline curves each with 4 e.d.f.	Females	3
LMS Model 2	LMS model with L, M and S curves smoothed by natural cubic spline curves with 3, 4 and 5 e.d.f, respectively.	Males	3
LMS Model 3	LMS model with L, M and S curves smoothed by cubic P-spline curves with 1, 5 and 5 e.d.f, respectively.	Females	3
LMS Model 3	LMS model with L, M and S curves smoothed by cubic P-spline curves with 1, 3 and 2 e.d.f, respectively.	Males	3
QR Model 1	Restricted quantile regression model from birth to 24 months of age, fitted using quadratic B-splines with two unequally spaced interior knots at ages 2.5 and 10 months.	Females	4
QR Model 2	Restricted quantile regression from birth to 24 months, fitted using quadratic B-splines with two unequally spaced interior knots at ages 3 and 11.5 months.	Males	4
PQR Model 1	Quantile regression model with cubic P-splines with a second order difference penalty, one interior equally spaced knot and $\lambda = 6$.	Females	4
PQR Model 2	Quantile regression model with cubic P-splines with a second order difference penalty, one interior equally spaced knot and $\lambda = 10$.	Males	4
Longitudinal Model 1	Extension of QR Model 1 with an additional $AR(1)$ component, allowing conditionality on age and a prior weight measurement.	Females	5
Longitudinal Model 2	Extension of QR Model 2 with an additional $AR(1)$ component, allowing conditionality on age and a prior weight measurement.	Males	5
Longitudinal Model 3	Extension of QR Model 1 with an additional $AR(2)$ component, allowing conditionality on age and two prior weight measurements.	Females	5
Longitudinal Model 4	Extension of QR Model 2 with an additional $AR(2)$ component, allowing conditionality on age and two prior weight measurements.	Males	5
Longitudinal Model 5	Extension of QR Model 1 with an additional $AR(1)$ and a partially linear component in the covariate vector x_i. This allows conditionality on age, a prior weight measurement and incorporation of a linear term for average parental height.	Females	5
Longitudinal Model 6	Extension of QR Model 2 with an additional $AR(1)$ and a partially linear component in the covariate vector x_i. This allows conditionality on age, a prior weight measurement and incorporation of a linear term for average parental height.	Males	5

Table 8.1: Complete list of models used in thesis to compose gender-specific weight growth charts from the Growth and Development Study data for full-term infants, including detailed description of each model and which gender it is modelled on.

Bibliography

H. Akaike. A New Look at the Statistical Model Identification. *Transactions on Automatic Control*, 19:716–723, 1974.

K. Bollaerts, P. H. C. Eilers, and I. V. Mechelen. Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology*, 59:451–469, 2006.

G. E. P. Box and D. R. Cox. An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B*, 26:211–252, 1964.

V. J. Carey. *An R package for Cole-Green Reference Centile Curves*. R Development Core Team(2004), 2002. URL `http://www.biostat.harvard.edu/~carey`.

G. Claeskens, T. Krivobokova, and J. D. Opsomer. Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544, 2009.

T. J. Cole. Fitting Smoothed Centile Curves to Reference Data. *Journal of the Royal Statistical Society, Series A*, 151:385–418, 1988.

T. J. Cole. The LMS method for constructing normalised growth standards. *European Journal of Clinical Nutrition*, 44:45–60, 1990.

T. J. Cole. Growth charts for both cross-sectional and longitudinal data. *Statistics in Medicine*, 13:2477–2492, 1994.

T. J. Cole. Conditional reference charts to assess weight gain in British infants. *Archives of Disease in Childhood*, 73:8–16, 1995.

T. J. Cole and P. J. Green. Smoothing reference centile curves: The LMS Method and penalized likelihood. *Statistics in Medicine*, 11:1305–1319, 1992.

M. de Onis, A. Onyango, E. Borghi, A. Siyam, and A. Pinol. *WHO Child Growth Standard,Length/height-for-age, weight-for-age, weight-for-length,weight-for-height and body mass index-for-age: methods and development.* Department for Nutrition and Development, World Health Organization (1211 Geneva 27, Switzerland), 2006.

M. de Onis, A. Onyango, E. Borghi, A. Siyam, and A. Pinol. *WHO child growth standards : growth velocity based on weight, length and head circumference : methods and development.* Department for Nutrition and Development, World Health Organization (1211 Geneva 27, Switzerland), 2009.

P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121, 1996.

J. Fox. *An R and S-PLUS Companion to Applied Regression*. Sage Publications, Thousand Oaks, California, 2002.

A. M. Fredriks, S. Van Buuren, R. J. Burgmeijer, J. F. Meulmeester, R. J. Beuker, E. Brugman, M. J. Roede, S. P. Verloove-Vanhorick, and J. M. Wit. Continuing positive secular change in the Netherlands, 1955-1997. *Pediatric Research*, 47:316–323, 2000a.

A. M. Fredriks, S. Van Buuren, J. M. Wit, and S. P. Verloove-Vanhorick. Body index measurements in 1996-7 compared with 1980. *Archives of Childhood Diseases*, 82:107–112, 2000b.

M. Geraci and M. Bottai. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8:140–154, 2006.

M. Geraci and M. Bottai. Linear quantile mixed models. *Statistics and Computing*, 2013.

P. J. Green. Penalised Likelihood for General Semi-parametric Regression Models. *International Statistical Review*, 55:245–259, 1987.

P. J. Green. In the Discussion of Fitting Smoothed Centile Curve to Reference Data. *Journal of the Royal Statistical Society, Series A*, 151:410–411, 1988.

T. Han, X. Wang, Y. Cui, H. Ye, X. Tong, and M. Piao. No Weight Catch-Up Growth of SGA Infants Is Associated with Impaired Insulin Sensitivity during the Early Postnatal Period. *International Journal of Pediatrics*, 2010. doi: 10.1155/2010/704642.

J. Hardin and J. Hilbe. *Generalized Linear Models and Extensions*. Stata Press, College Station, second edition, 2007. ISBN 1-59718-014-9.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, first edition, 1990. ISBN 978-0-412-34390-2.

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

R. Koenker and G. Bassett. Regression Quantiles. *Econometrica*, 46:33–50, 1978.

J. H. Mathews and K. K. Fink. *Numerical Methods Using Matlab, 4th Edition*. Prentice-Hall Inc., Upper Saddle River, New Jersey, USA, 2004.

P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, Boca Raton, second edition, 1989. ISBN 0-412-31760-5.

M. C. Meyer. Constrained penalized splines. *The Canadian Journal of Statistics*, (1):190–206, 2012.

V. M. R. Muggeo, M. Sciandra, A. Tomasello, and S. Calvo. Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. *Environmental Ecological Statistics*, 2012. doi: 10.1007/s10651-012-0232-1.

Z. Nie and J. S. Racine. The crs Package: Nonparametric Regression Splines for Continuous and Categorical Predictors. *The R Journal*, 4, 2012.

H. Pan and H. Goldstein. Multilevel models for longitudinal growth norms. *Statistics in Medicine*, 16:2665–2668, 1997.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location scale and shape. *Journal of Statitsical Society, Series C*, 54:507–554, 2005.

R.A Rigby and D.M Stasinopoulos. Smooth centile curves for skew and kurtotic data modelled using the box-cox power exponential distribution. *Statistics in Medicine*, 23:3053–3076, 2004.

Royal College of Paediatrics and Child Health. UK-WHO 0-4 years growth chart resources, Accessed August 2013. URL `http://www.rcpch.ac.uk/child-health/research-projects/uk-who-growth-charts/uk-who-growth-chart-resources-0-4-years/uk-who-0`.

G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6 (2):461–464, 1978.

B. W. Silverman. Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting. *Journal of the Royal Statistical Society, Series B*, 47:1–52, 1985.

S. Stanojevic, T. J. Cole, J. Stocks, A. L. Coates, J. L. Hankinson, and A. M. Wade. Age- and size-related reference ranges: A case study of spirometry through childhood and adulthood. *Statistics in Medicine*, 28: 880–898, 2009.

D. M. Stasinopoulos and R. A. Rigby. Generalized Additive Models for Location, Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, 23, 2007.

Wales National Health Service. Child Health System, Accessed August 2013. URL `http://www.wales.nhs.uk/nwis/page/52630`.

Y. Wei. An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. *Journal of the American Statistical Association*, 103:397–409, 2008.

Y. Wei, A. Pere, R. Koenker, and X. He. Quantile regression methods for reference growth charts. *Statistics in Medicine*, 25:1369–1382, 2006.

S. N. Wood. *Generalized Additive Models: An introduction with R*. Chapman and Hall/CRC, 2006. Texts in Statistical Science Series.

C. M. Wright and T. D. Cheetham. The strengths and limitations of parental heights as a predictor of attained height. *Archives of Disease in Childhood*, 81:257–260, 1999.

C.M. Wright, J.N. Matthews, A. Waterson, and A. Aynsley-Green. What is a normal rate of weight gain in infancy? *Acta Paediatrica*, 83:351–356, 1994.