Channgam, Taweesak (2024) *Flexible quantile regression and Bayesian quantile modelling for longitudinal child growth data.* PhD thesis.

https://theses.gla.ac.uk/84689/

Flexible quantile regression and Bayesian quantile modelling for longitudinal child growth data

Taweesak Channgam,
B.Sc., M.Sc.

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Mathematics and Statistics
College of Science and Engineering
University of Glasgow

March 2024

Abstract

Numerous studies have been conducted to understand the comprehensive mechanisms of child growth and development. Among these, the longitudinal study is a key approach that can provide a complete characterisation of these aspects. However, the greater benefits of this approach often come with increased complexities inherent in the data from such studies. These complexities include non-linear trajectories, within-subject correlation, heterogeneity of individual baseline and dynamic growth characteristics, and auto-correlation within individuals. Moreover, given the diverse range of growth patterns in children, it is particularly relevant to evaluate risk factors that exhibit significant associations with growth measurements across different locations in the distribution, rather than focusing solely on the central location, such as the mean or median. This thesis aims to address these complexities by examining longitudinal child growth data (LCGD) through appropriate statistical models. In order to characterise LCGD and describe the entire distribution of growth, the additive quantile mixed model (AQMM) has been reviewed. This approach has been incorporated into both the mixed-effects model framework and the additive model, enabling the analysis of longitudinal data. Simulation studies conducted within this thesis assess the performance of AQMM under various experimental designs in the context of LCGD. Overall, AQMM performed well in predicting simulated data. Subsequently, the LCGD in Scotland was used to construct reference growth curves and identify risk factors associated with physical growth measurements. However, AQMM lacks a method for determining appropriate random effects to capture individual variability.

To address this limitation, a Bayesian variable selection method within quantile mixed models (QMMs) is proposed. This novel methodology combines several key components: the Bayesian sparse group LASSO method with spike and slab priors, a likelihood function based on the scale mixture representation of the asymmetric Laplace (AL) distribution, and the utilisation of mixed models based on a decomposition for the covariance matrix of random effects. By incorporating these elements, the methodology establishes a comprehensive framework to tackle challenges associated with the selection and estimation of fixed and random effects in the context of QMMs. To evaluate the performance of the novel method, simulation experiments are conducted. Furthermore, the proposed approach is applied to the analysis of LCGD in Scotland, thus providing practical insights into its real-world applicability. Overall, the novel model demonstrates strong performance in variable selection with simulated data. When applied to LCGD in Scotland, the selected risk factors were consistent across both lower and upper quantiles of physical growth measurements in school-age children and young people in primary or secondary education. The selection of both random intercepts and random slopes suggests variability in individual linear trends.

Contents

List of Tables

List of Figures

Acknowledgements

Undertaking a PhD without any help, support, or encouragement is indeed an impossible task. I have received a great deal of these from many people. Firstly, I would like to extend my heartfelt thanks to my supervisors, Dr Craig Anderson and Dr Tereza Neocleous, for providing opportunities, invaluable guidance, and support throughout my PhD journey. Without them, this work would not have been possible. I am also grateful to Dr Amira El-Ayouti from Cairo University for her invaluable feedback and comments, especially in the initial stage of my thesis. Furthermore, my sincere appreciation goes to the Scottish Centre for Social Research for generously providing the Growing up in Scotland data, which was crucial for my analysis. Their support has been a cornerstone in the completion of this study. Moreover, I would like to thank Professor Eric Ohuma and Dr Nema Dean for their valuable comments, which have greatly strengthened this thesis.

Secondly, I am also deeply grateful to Maejo University and my colleagues for enabling me to obtain a scholarship to pursue my PhD. Additionally, I gratefully acknowledge the funding received for my PhD from the Ministry of Higher Education, Science, Research and Innovation, Royal Thai Government. Moreover, I would like to thank the OEAUK for supporting my life in Scotland.

To my best friends: Aim, Kwang, and Aung, who have always supported me in both good times and bad. I am profoundly grateful for your willingness to pick up my calls in the middle of the night, Thailand time, to listen to my troubles. Thank you so much. To my flatmate, Naruemon Tiyasangthong, thank you so much for being a wonderful flatmate and for your help in many things. To my friend, Kamol Sanittham, thank you for sharing valuable PhD experiences and guiding me in various aspects of my PhD journey. To my brotherly friend, Puchong Paophan (X), thank you for kindly sharing academic knowledge and for the enjoyable moments we shared.

Most importantly, to my family, Mom and Dad, your love has encouraged me in making my dream come true and in becoming the man I am today. And to myself, I feel a profound sense of pride.

Declaration

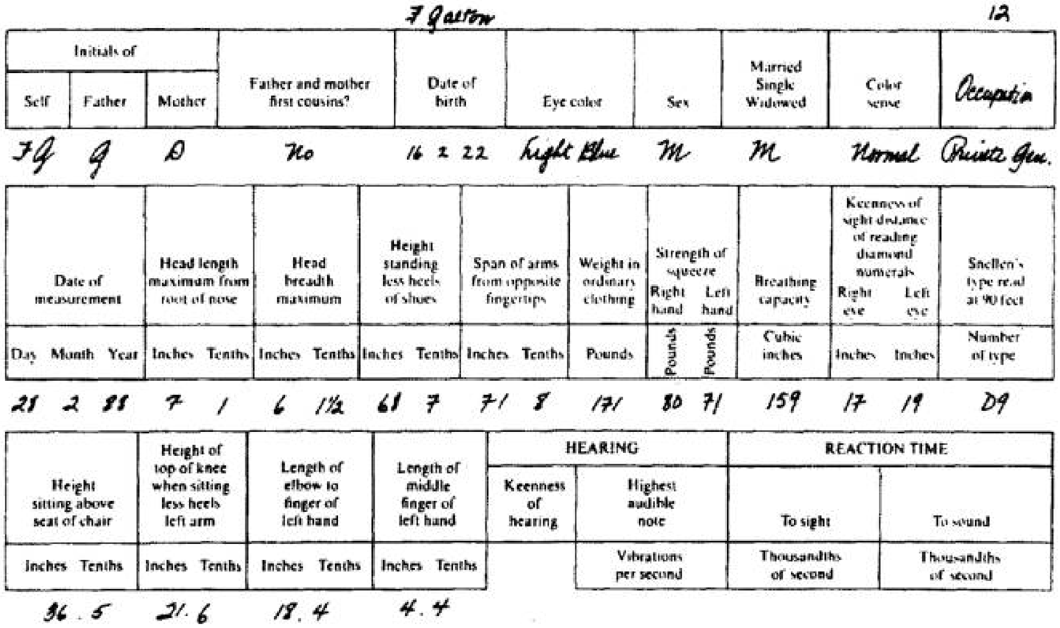
I, Taweesak Channgam, hereby declare that this thesis, titled "Flexible quantile regression and Bayesian quantile modelling for longitudinal child growth data", represents my own work, completed after registration for the degree of PhD at the University of Glasgow. This work has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree. I have acknowledged all sources used and have cited these in the bibliography section.

Further, I delivered a poster presentation on this work described in Chapter 4 at the RSS International Conference in Aberdeen, UK, in 2022. Additionally, I gave a talk on the work from Chapter 5 at the Research Students' Conference in Probability and Statistics in Sheffield, UK, in 2023.

Chapter 1

Introduction

From ancient civilisations to modern times, the study of human growth and development has remained a central focus for societies. Evidence of writings about human growth has been found in ancient societies such as the Babylonians and Egyptians (Tanner, 1981). In modern times, several movements related to the study of human growth explicitly began in the 18th century, with the primary focus of research being on measuring physical dimensions and describing patterns of growth. Assessments of child growth and development started appearing in scientific documents. In 1777, Buffon, a French naturalist, mathematician, cosmologist, and encyclopedic author, published a chart in *Histoire Naturelle* demonstrating the height of De Montbeillard's son from 1759 to 1777 across different ages.



F Galton 12

Initials of			Father and mother first cousins?	Date of birth	Eye color	Sex	Married Single Widowed	Color sense	Occupation
Self	Father	Mother							
FG	G	D	No	16 2 22	Light Blue	M	M	Normal	Private Gen.

Date of measurement			Head length maximum from root of nose		Head breadth maximum		Height standing less heels of shoes		Span of arms from opposite fingertips		Weight in ordinary clothing	Strength of squeeze Right hand	Strength of squeeze Left hand	Breathing capacity	Keenness of sight distance of reading diamond numerals Right eye	Left eye	Snellen's type read at 90 feet
Day	Month	Year	Inches	Tenths	Inches	Tenths	Inches	Tenths	Inches	Tenths	Pounds	Pounds	Pounds	Cubic inches	Inches	Inches	Number of type
21	2	88	7	1	6	1½	61	7	71	8	171	80	71	159	17	19	D9

Height sitting above seat of chair		Height of top of knee when sitting less heels left arm		Length of elbow to finger of left hand		Length of middle finger of left hand		HEARING		REACTION TIME	
								Keenness of hearing	Highest audible note	To sight	To sound
Inches	Tenths	Inches	Tenths	Inches	Tenths	Inches	Tenths		Vibrations per second	Thousandths of second	Thousandths of second
36	5	21	6	18	4	4	4				

Figure 1.1: Anthropometric laboratory card of Francis Galton, sourced from R. C. Johnson et al. (1985)

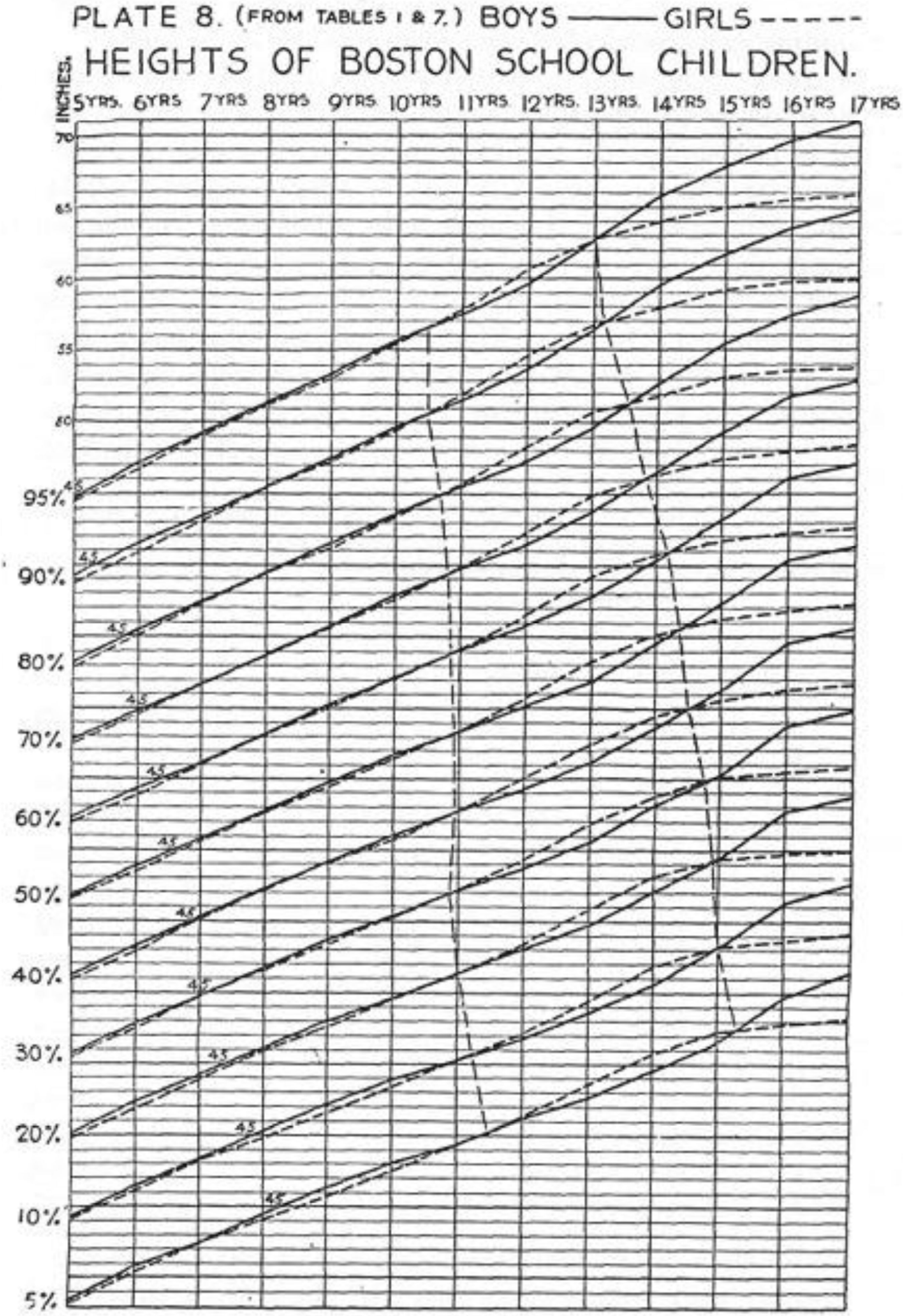


Figure 1.2: Height versus age for each percentile, sourced from Bowditch (1891)

Research in human growth became more scientific and rapidly increased in the 19th century, particularly in the study of child growth. In Europe, the assessment of child growth through physical dimensions remained prominent, encompassing more than just height measurement. Other measurements, such as weight, arm span, and head circumference, were used to describe patterns of human growth (refer to Figure 1.1) (Galton, 1883). The Body Mass Index (BMI), known as the Quetelet index, was also introduced in this context (Quetelet, 1832; Sarton, 1935). The concept of percentile curves was initially used to describe growth patterns. One notable application was by Henry Bowditch (1840–1911), who introduced height charts based on percentiles for each age group (see Figure 1.2) (Bowditch, 1891). It has since become the standard format for growth charts used today. It is important to note that these growth charts were constructed using single-time-point measurements per child, a collection method known as "cross-sectional data". Another important finding this century is that a country's socio-economic status can impact an individual's adult height (Villermé, 1835). Moreover, it was during this time that the adolescent growth spurt was first described (Kotelmann, 1879).

In the 20th century, the individual growth chart known as the "height-for-age" curve or "height distance" (see Figure 1.3) was first constructed and published in 1927 by Richard E. Scammon (1883 - 1952). Scammon utilised serial height measurements collected over time from De Monteilard's son (Scammon, 1927). This growth chart basically describes the progress of a child's growth at any particular age. Furthermore, during this period, many "longitudinal studies" were conducted, collecting anthropometric data from individuals over long-term periods. The Fels Growth Study, which began in 1929, was one of the earliest studies in this regard (Roche, 1992). This study collected anthropometric data from American children and expanded to include other measurements, such as ultrasound, skeletal maturation, and dual-energy X-ray absorptiometry (DXA), and continues to do so to this day. In the UK, one of the first large-scale birth cohort studies was conducted in 1946 by the National Study of Health and Development (Wadsworth et al., 2006). This cohort was formed to investigate the social and biological factors affecting health and development from birth into adulthood. The data from longitudinal studies were crucial for offering a comprehensive and detailed view of how human growth over time. Due to this benefit, the introduction and use of velocity growth charts (see Figure 1.4) represented a revolutionary innovation in human growth assessment, becoming essential tools for the clinical evaluation of individuals (Tanner, 1962). In the late 20th century, advanced statistical methods, such as the Lambda-Mu-Sigma (LMS) method, were used to construct growth reference charts/tables to serve this purpose (Cole, 1997, 2012; Cole & Green, 1992). It revolutionised the construction of growth charts by allowing the adjustment of skewness in the growth data, making the data more normally distributed and improving the fit of the growth charts.

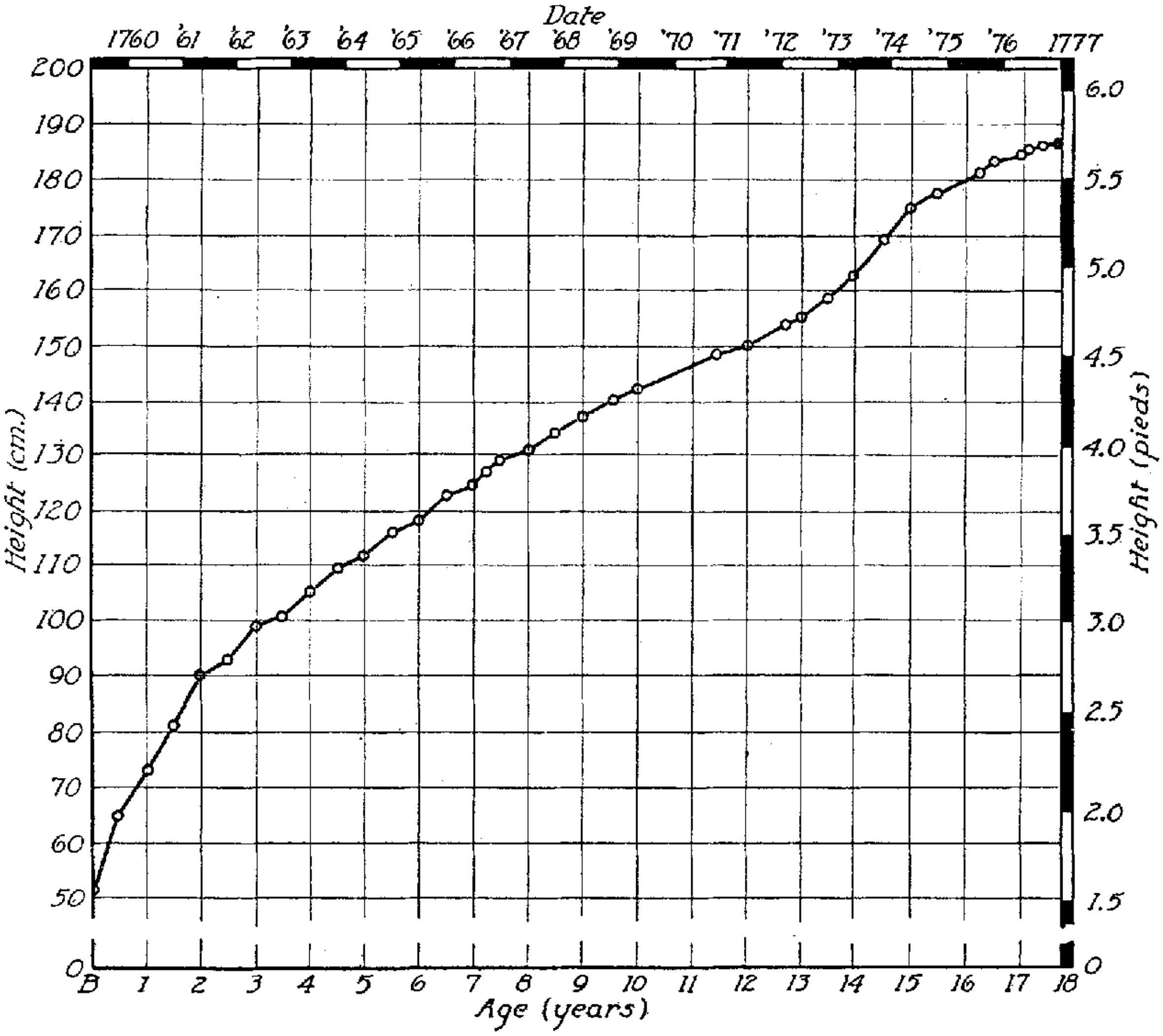


Figure 1.3: The height distance growth of De Montbeillard's son as constructed and published by Scammon, sourced from Scammon (1927)

In the 21st century, studies on human growth and development have expanded, offering deeper insights into growth patterns and influencing factors (Fogel et al., 2008). Longitudinal studies have extended their duration, increased sample sizes, expanded coverage across different countries, and enhanced data collection methods. Notably, global growth standards, such as those established by the World Health Organization (WHO) (WHO Multicentre Growth Reference Study Group & de Onis, 2006), have been developed and implemented during this period. Many countries have developed their own growth assessment tools, including growth tables and charts, to monitor and screen the health of children, such as in Belgium (Roelants et al., 2009), Germany (Rosario et al., 2014), Norway (Júlíusson et al., 2009), etc. Various statistical methods, such as Generalised Additive Models for Location, Scale and Shape (GAMLSS) (Rigby & Stasinopoulos, 2005), Box-

Cox-Power-Exponential (BCPE) method with splines (Rigby & Stasinopoulos, 2004), have been proposed for constructing these growth charts and analysing child growth data.

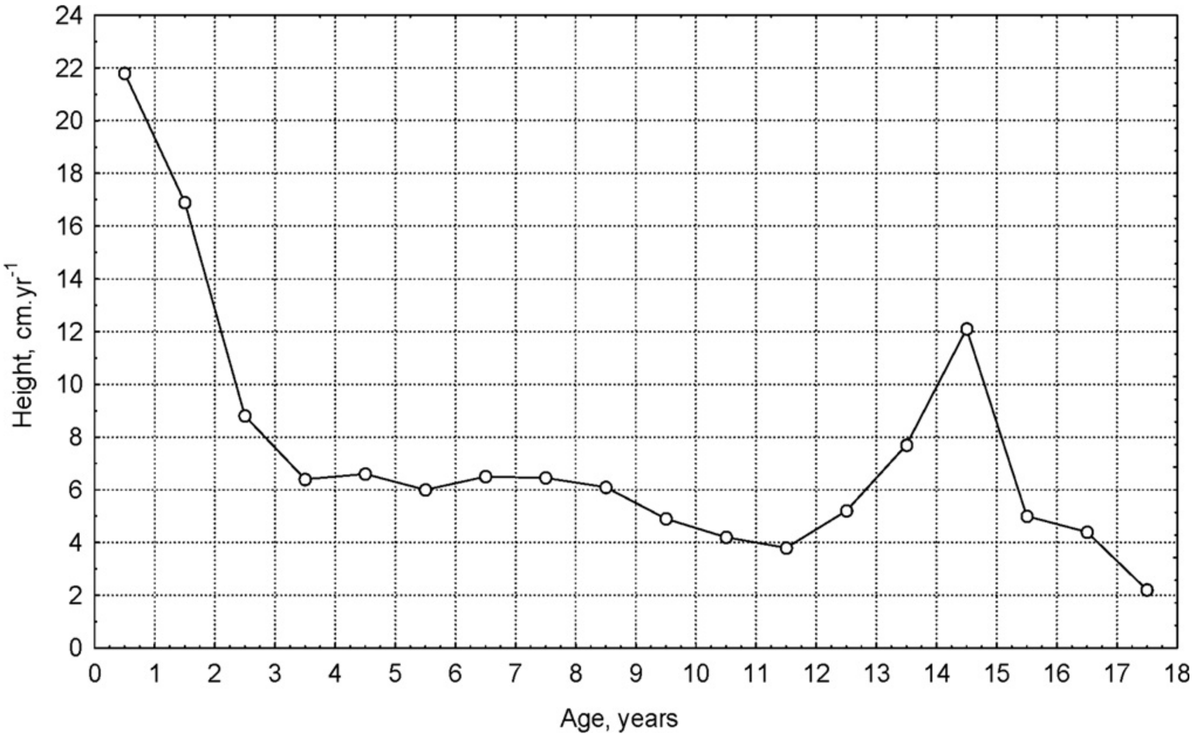
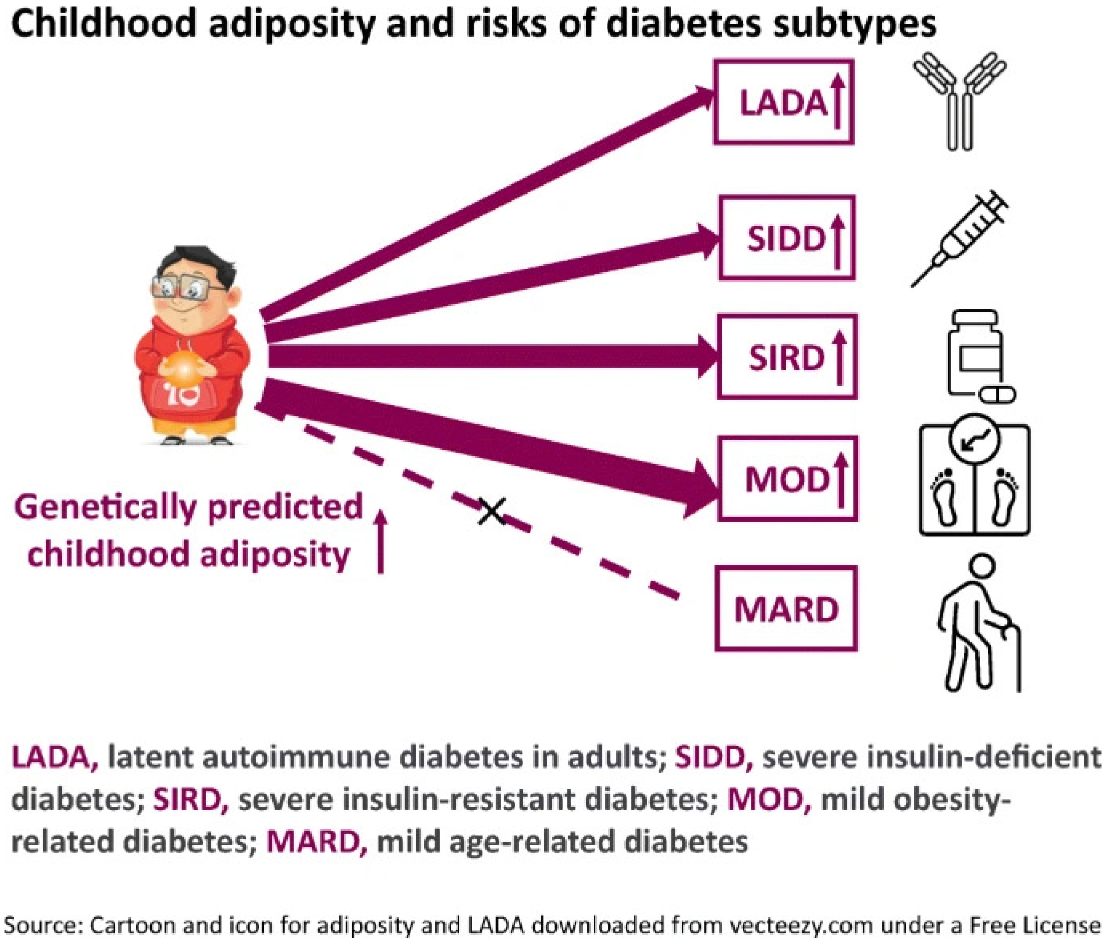


Figure 1.4: The velocity growth of De Montbeillard's son, sourced from Tanner (1962)

Therefore, from the past to the present day, studies on human growth and development have consistently aimed to understand growth and improve health outcomes across the human lifespan, including childhood, adolescence, and adulthood. Each of these stages have its own unique processes, requiring specific knowledge and tools for understanding and improvement. Childhood, in particular, is a crucial period of physical growth and development in human beings, setting the trajectory for an individual's future health, cognitive abilities, and overall well-being (Berk & Meyers, 2016). Childhood typically consists of various sub-period stages (Berk & Meyers, 2016), including the prenatal period, infancy and toddlerhood, early childhood and middle childhood. Each stage has its own distinct mechanism of physical growth; for instance, infancy is characterised by rapid weight gain and considerable increases in length, while early childhood generally exhibits a steadier, slower rate of growth in height and weight (Berk & Meyers, 2016; Bukatko & Daehler, 1995). During each stage, children require age-appropriate nutrition and an environment conducive to their growth. Abnormal growth during these stages can impact subsequent developmental phases (Barker, 1990). For example, there is a strong connection between being overweight or obese and adiposity in children (Guo & Chumlea, 1999; Krebs et al., 2007; Y. Wang, 2017). Consequently, children with higher adiposity may be at a greater risk of developing diabetes subtypes in adulthood compared to those of normal weight (see Figure 1.5). Additionally, stunted children may face long-term consequences such as impaired development and learning capacity, insulin resistance, and a heightened risk of developing diabetes (De Sanctis et al., 2021). In this regard, having appropriate tools (e.g. growth charts) and information (e.g. risk factors) can help prevent inappropriate growth and support children in experiencing normal growth.


Childhood adiposity and risks of diabetes subtypes

LADA, latent autoimmune diabetes in adults; SIDD, severe insulin-deficient diabetes; SIRD, severe insulin-resistant diabetes; MOD, mild obesity-related diabetes; MARD, mild age-related diabetes

Source: Cartoon and icon for adiposity and LADA downloaded from vecteezy.com under a Free License

Figure 1.5: Graphical abstract from *Childhood adiposity and novel subtypes of adult-onset diabetes: a Mendelian randomisation and genome-wide genetic correlation study* by Wei et al. (2023)

Beyond childhood, the adolescent period is also a major stage of child development, marking the transition between childhood and adulthood. During this stage, children undergo several biological changes. Growth spurts are initially evident, with rapid increase in height and weight (Berk & Meyers, 2016; Cameron & Bogin, 2012; Marshall, 1978). Subsequently, secondary sexual characteristics typically develop during this period, known as *puberty* (Berk & Meyers, 2016; Cameron & Bogin, 2012; Marshall, 1978). Girls experience breast development, the onset of menstruation (menarche), and widening of hips, while boys experience the growth of facial and body hair, deepening of the voice, and enlargement of the testes and penis. In addition to these changes, adolescents also undergo hormonal changes, bone growth, muscle development, skin changes, and brain development. As mentioned previously, rapid developmental changes and physical growth occur during this period, indicating that monitoring growth at this stage is essential. For instance, early onset of puberty can lead to shorter adult height due to the premature closure of growth plates (Carel & Léger, 2008). Children with growth hormone deficiency (GHD) may experience shorter stature compared to their peers and delayed puberty (Cohen et al., 2008).

In the family unit, parents play a crucial role, bearing the responsibility of properly monitoring their child's growth, using appropriate tools and seeking guidance from health professionals. To support these efforts, several countries have established their own organisations, such as the Royal College of Paediatrics and Child Health in the UK (www.rcpch.ac.uk). On a global scale, the World Health Organization (WHO, www.who.int) assumes the responsibility of promoting and safeguarding child growth and development in various means. These include setting global standards, conducting research, developing policies, implementing health programmes, enhancing education, supporting healthcare systems, and responding to emergencies. This underscores the significance of child growth and development, reinforcing the importance of family, local, and global perspectives.

For decades, child growth charts have been essential tools used to monitor children's physical growth across different ages (Cole, 2012). These charts comprise several centile or quantile curves that depict the comprehensive profile of physical growth measurements. The monitoring process involves plotting a child's physical measurements on this chart to demonstrate their growth pattern at specific ages, usually separated by sex. Essentially, if a child's growth follows a relatively consistent curve or pattern within the growth chart, it can indicate normal growth. However, it is important to note that child growth charts are not diagnostic tools in themselves; rather, they can signal when further evaluation might be necessary. Another consideration is that these charts were constructed from cross-sectional data; therefore, they cannot measure growth velocity, as this would require data collection from a representative subset of the target population at multiple ages.

Unlike cross-sectional data, longitudinal data provide another means to capture all important aspects of outcomes, especially in characterising trends and changes in each individual's outcome over time (Fitzmaurice & Ravichandran, 2008). This type of data involves the repeated observation of the same individuals over a period, which is a key difference compared to cross-sectional data. In the context of child growth studies, longitudinal data have become increasingly popular for exploring temporal variation in child growth measurements. However, these advantages come at a cost of increased complexity in analysis, for instance, within-subject correlation, heterogeneity of individual baseline and dynamic growth characteristics, and autocorrelation within individuals.

Linear mixed-effects models (LMMs) are a statistical method that has been widely used to analyse longitudinal data. The LMMs proposed by Laird and Ware (1982) are well-known and popular in this field. A key contribution of these models is the incorporation of random effects, capturing individual-specific variations, while simultaneously accounting for population-level effects or fixed effects that are thought to have a consistent impact on growth across all children. The inclusion of random effects for each subject accommodates the assumption that measurements from the same subject are more likely to be similar or correlated, making these models particularly suitable for longitudinal child growth data (LCGD). However, these models are limited by the assumption that growth patterns are linear. To allow for more flexibility, they have been extended to model non-linear trajectories using a nonparametric approach, such as splines, leading to what are known as semiparametric mixed-effects models (Aniley et al., 2019; Durbán et al., 2005; Ruppert et al., 2003).

Models based on mixed-effects models, as mentioned previously, are conditional mean models. This means they focus only on the central location of the growth measurements. Consequently, they cannot provide a comprehensive analysis of the complete distribution of child growth measurements. As a result, these models do not adequately address questions regarding children at the lower and upper ends of the growth spectrum, who may have different growth mechanisms and risk factors associated with growth. Notably, over 340 million children and adolescents were living with overweight or obesity, which usually places them at the upper end of the growth spectrum (WHO, 2021). Also, in Scotland, the Royal College of Physicians and Surgeons of Glasgow (2023) reported that in 2021, 18% of Scottish children were at risk of obesity, and there are concerns that this percentage may rise in the future, as shown in Figure 1.6. These figures highlight the urgency of addressing child growth and development through effective tools, identifying risk factors, establishing policies, and implementing prevention programmes for this specific child group.

Similarly, another important group of children includes those affected by thinness, stunting and underweight, who are typically found at the lower end of the growth spectrum. According to WHO (2024), an estimated 190 million children were living with thinness, defined by BMI-for-age more than two standard deviations below the reference median, and approximately 149 million children under the age of 5 years suffered from stunting in 2022. Although only 1% of Scottish children were reported as underweight in 2019 (Vosnaki et al., 2019), this remains a concern similar to that of overweight children.

Quantile regression (QR) is a suitable approach (Koenker, 2005) to address the afore-

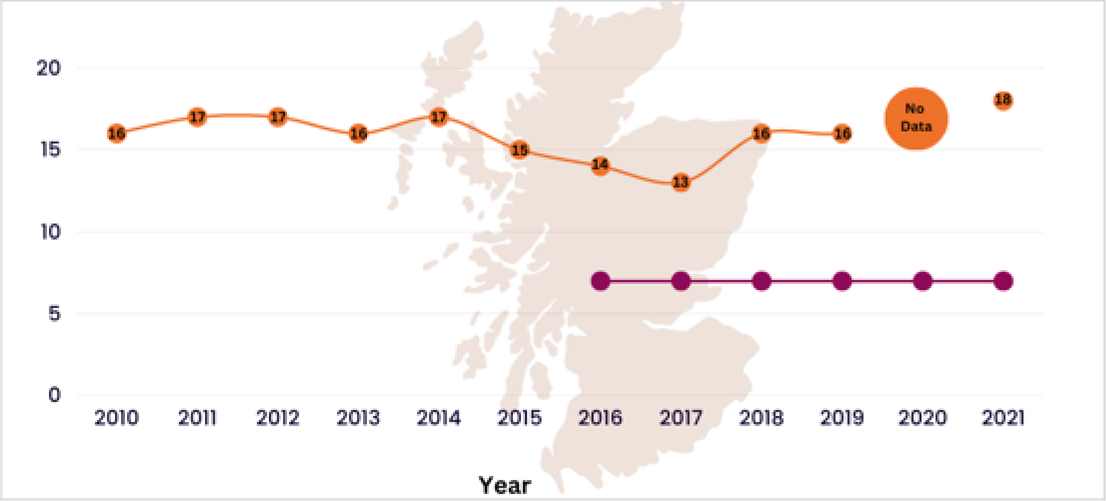


Figure 1.6: Proportion of children at risk of obesity in Scotland 2010 - 2021 (based on Scottish Health Survey data): the purple line represents the goal achievement of child obesity prevalence, and the orange line represents the actual rate of children at risk of obesity (Royal College of Physicians and Surgeons of Glasgow, 2023)

mentioned limitations. It offers a valuable way to describe the entire distribution of a target response conditional on the independent variables. Growth curve estimation has been successfully addressed with this approach in cross-sectional data, as demonstrated by Carey et al. (2004) and Muggeo et al. (2013). However, applying QR to longitudinal data, particularly LCGD, remains challenging due to the inherent complexity of such data. Traditional QR relies on a distribution-free framework, which complicates the incorporation of other frameworks necessary for characterising longitudinal data. The literature includes a few proposed QR approaches to handling LCGD. One such approach is the quantile specific autoregressive model (QSAM), a traditional non-parametric QR method proposed by Wei et al. (2006). QSAM uses a non-parametric function, such as B-splines, to model growth measurements, allowing for the assessment of unusual growth patterns. The main feature of this method is its combination of the first-order autoregressive (AR(1)) model to account for the dependence between growth measurements of the same child. While this model performs well, it is limited to the AR(1) framework and provides only the population-quantile effects. As a result, it cannot predict or estimate individual growth trajectories.

Recently, Geraci (2019) proposed a new QR approach for longitudinal data, utilising a likelihood based on an asymmetric Laplace (AL) distribution to estimate quantile functions. This approach also incorporates additive models (smooth terms) and mixed-effects models to capture non-linear dependencies and account for individual-specific variations, respectively. Additionally, the inclusion of random effects in the model explicitly addresses within-subject correlation. Consequently, this approach is named the additive quantile mixed model (AQMM) (Geraci, 2019). AQMM appears to be an attractive approach for analysing or modelling LCGD, as it is a flexible method capable of capturing non-linear growth patterns and accounting for child-specific variations and handeling correlation in repeated growth observations for the same child. However, AQMM allows for specifying smooth terms in various ways. This raises questions about the suitability of smooth terms for modelling non-linear growth patterns. To address this, P-splines (Eilers & Marx, 1996) are chosen in this thesis for their robustness, flexibility, and ease of use. P-splines combine B-splines with a difference penalty to ensure smoothness and can handle complex patterns without overfitting. In the context of AQMM, P-splines are advantageous as they allow for the construction of basis functions with equidistant knots, simplifying parameter specification and eliminating the need for optimal knot placement. Furthermore, AQMM is limited to specifying linear random effects (e.g. random intercepts and random slopes), which raised questions about whether this limitation adequately captures the variability present in LCGD. Therefore, its effectiveness should be confirmed through simulation experiments in the context of LCGD, focusing on the two main aspects mentioned.

Moreover, AQMM still lacks the capability to ascertain which random effects are appropriate for capturing individual variations. Regarding the selection of fixed effects, an inferential method based on the resampling technique, such as bootstrapping, is used for hypothesis testing. Yet, this approach has been debated in various aspects. In parametric bootstrap, particularly in models with penalties, the coefficients of bases may result in a smoothing bias in parametric model. Meanwhile, under-smoothing can occur due to the sampling-with-replacement process (e.g. some data points may be sampled twice) in the nonparametric bootstrap (Wood, 2006). These factors can influence hypothesis testing.

To circumvent this issue, the Bayesian LASSO method (Park & Casella, 2008) is recommended. This Bayesian approach assumes that the regression parameters follow independent Laplace priors and performs variable selection using Markov chain Monte Carlo (MCMC)-based computational techniques, such Gibbs sampling. However, it does not support grouped variables, such as a group of bases forming a smoother, because it treats each variable independently and applies shrinkage individually to each predictor. Moreover, the estimates obtained from this method may not precisely converge to zero (Alhamzawi & Ali, 2018; Park & Casella, 2008; Xu & Ghosh, 2015), which can potentially impact the accuracy of variable selection. To accommodate grouped variables and ensure exact zero convergence, Xu and Ghosh (2015) proposed two variants of the Bayesian LASSO method with spike and slab priors. These two variants consist of Bayesian group LASSO and Bayesian sparse group LASSO. The first handles the selection process for entire groups of variables, while the second goes a step further by also encouraging selection within the groups. Meanwhile, the spike and slab priors are employed to construct the point mass at zero for regression parameters. However, these methods were only applied to independent data and the mean model. Therefore, the novel methodology developed in this thesis extends these two Bayesian models in the context of quantile regression: Bayesian group LASSO qunatile regression with spike and slab (BGLSSQR) and Bayesian spare group LASSO quantile regression with spike and slab prior (BSGSSQR). Subsequently, the BSGSSQR is incorporated with a reparameterisation of the random effects component within the linear mixed model proposed by Kinney and Dunson (2007), enabling the simultaneous selection of both fixed effects and random effects. This novel method is named Bayesian spare group LASSO-mixed quantile regression with spike and slab prior (BSGSSMQR). This method also addresses relevant questions in the analysis of longitudinal child growth data in Scotland, particularly in identifying risk factors affecting to child physical growth.

The primary aim of this thesis is to examine longitudinal child growth data and describe child growth patterns using appropriate models, as outlined in the following steps. Firstly, specific locations within the distribution of physical growth measurements are represented by several quantiles, ranging from 0.004 to 0.996. In particular, three quantiles, the 0.10th, 0.50th and 0.90th, correspond to the lower, middle, and upper tails of the distribution, representing children in the lowest 10% (indicative of underweight or short stature), middle 50% (indicative of average weight or stature), and highest 90% (indicative of overweight or taller stature). These specific quantiles are utilised throughout the thesis. Each quantile is estimated using conditional quantile models with two existing methods (QSAM and AQMM) and the novel method (BSGSSMQR). Specially, the growth patterns are presumed to follow non-linear trajectories to fit these models. The QSAM model is based on classical quantile regression with splines, while the AQMM and BSGSSMQR models comprise two components: fixed effects and random effects. The former represents population-level effects with both non-linear fixed effect and linear fixed effects, while the latter demonstrates individual-level effects through individual linear trends, including intercepts and temporal slopes. Furthermore, another main focus of this thesis is the development of a statistical model to identify risk factors associated with specific child physical growth measurements, such as raw weight and Weight-for-Age Z-score (WAZ), for children in the lowest 10% and highest 90% of these physical growth measurement distributions, represented by the 0.10th and 0.90th quantiles. These quantiles correspond to children with underweight and overweight, respectively. Additionally, these measurements are chosen due to the growing concern over child obesity in Scotland connected to overweight children, as shown in Figure 1.6. This also includes determining suitable random variation in the intercept and slope coefficients that vary among subjects. More details of the methods are discussed in Chapters 4 to 5, with applications presented in Chapter 6.

The remainder of this thesis is organised as follows: Chapter 2 provides an overview of child growth and its relevance. This includes a brief discussion of child growth and development, child growth studies, child physical growth measurements, child growth charts, and risk factors affecting child growth and development. Furthermore, it introduces the longitudinal study, "the Growing up in Scotland study" (GUS), which furnishes LCGD used in this thesis. Data exploration is presented in thesis chapter, along with a summary.

Chapter 3 provides detailed outlines of the statistical methodologies involved in the analysis of LCGD, forming the fundamental basis for Chapter 4 and Chapter 5 of this thesis. It begins by covering methodologies, particularly those in conditional mean models, used to account for within-subject variations within LCGD. This includes both linear mixed-effects models and flexible mixed-effects models. In the area of conditional quantile models, quantile regression for independent data is initially discussed, including two estimation methods for quantile functions: the distribution-free method (Koenker, 2005) and the likelihood-based method (Geraci & Bottai, 2007; Koenker & Machado, 1999b; Yu & Moyeed, 2001). Furthermore, the Bayesian methods in the context of quantile regression are provided, including two specific working likelihoods based on the asymmetric Laplace (AL) distribution (Yu & Moyeed, 2001) and a location-scale mixture of normals representation of AL (Alhamzawi & Yu, 2013; Kozumi & Kobayashi, 2011; Tsionas, 2003). The earliest works on quantile regression models for longitudinal data are also reviewed in this chapter. Additionally, splines, which are techniques that facilitate the representation of non-linear dependence in the context of regression methods, are outlined. The regression methods involving these splines are presented, including regression splines, penalised regression splines with P-splines by Eilers and Marx (1996), and the representation of P-splines as mixed-effects models (Currie & Durban, 2002; Eilers, 1999).

Chapter 4 specifically focuses on flexible models in the context of quantile regression for modelling longitudinal child growth data. Two flexible models are discussed, including the quantile specific autoregressive model (QSAM) (Wei & He, 2006) and the additive quantile mixed model (AQMM) (Geraci, 2019). These two models rely on different methods to model non-linear trajectories: QSAM uses regression splines, while AQMM employs penalised regression splines. Moreover, for estimating quantile functions, QSAM is based on distribution-free method, whereas AQMM relies on the likelihood-based method. QSAM has been applied to LCGD to estimate reference growth curves. However, to the best of our knowledge, AQMM has not yet been applied in this context. Therefore, this chapter gives several experiments through simulation studies to explore the behaviour and assess the performance of AQMM when implemented with LCGD. It also provides a summary of the advantages and limitations of this method.

After exploring the AQMM approach in Chapter 4, it becomes evident that AQMM lacks the capability to identify random effects capturing individual variations in LCGD. Although AQMM has an inferential method to identify the fixed effects (risk factors) via the bootstrap method, questions have been raised about the accuracy of this method in model selection. Since this thesis focuses on models for longitudinal child growth data, such models must possess the ability to identify potential features/effects in a manner that appropriately reflects the mechanism of LCGD. Chapter 5 will present the methodologies used for the simultaneous selection of fixed and random effects in the context of quantile mixed models, which are specifically designed to analyse longitudinal child growth data. Initially, two methodologies based on Bayesian LASSO-type methods with spike and slab priors (Bayesian group LASSO (BGLSSQR) and Bayesian sparse group LASSO (BSGSSQR)) are applied for selecting fixed effects in the quantile models. Subsequently, the BSGSSQR method is extended by introducing a decomposition for the covariance matrix of random effects based on the reparameterisation of these effects within the linear mixed model, enabling the simultaneous selection of both fixed and random effects. This novel method, abbreviated as BSGSSMQR, stands for Bayesian spare group LASSO-mixed quantile regression with spike and slab prior. Unlike the AQMM method, this approach relies on Bayesian computation, and both fixed and random effects are selected via MCMC-based computation techniques. Several simulation studies are conducted to investigate the performance of the models used in this chapter.

Chapter 6 demonstrates the implementation of the AQMM approach reviewed in Chapter 4, along with the method developed in Chapter 5 for modelling LCGD from the Growing up in Scotland study. The main goals of this implementation are to identify potential risk factors (fixed effects) associated with child weight measurements and to select the appropriate individual variations (random effects) among Scottish children, particularly for school-age children and young people aged 4 to 14 years in primary or secondary care. Finally, Chapter 7 provides a summary of this thesis and discusses potential future developments arising from the work undertaken in this research.

Chapter 2

Child growth and its relevance

This chapter offers an overview of child growth and introduces the Growing up in Scotland study (GUS), a longitudinal investigation into child growth that furnishes the data analysed in this thesis. This chapter is structured into seven sections. The opening section (Section 2.1) explores the fundamentals of growth and development in children. Section 2.2 provides an overview of two distinct child growth studies. Section 2.3 elucidates on the physical child growth measurements, comparing two different scales. Section 2.4 introduces the concept of child growth charts that have been used to monitor child physical growth. Section 2.5 offers a summary of risk factors associated with child growth and development. In Section 2.6, details pertaining to the GUS study are presented, and its data are explored. Finally, Section 2.7 concludes with a comprehensive summary of the chapter.

2.1 Growth and development in children

Human growth and development have been studied since antiquity (Tanner, 1981). At that time, these studies relied on observation and philosophical speculation rather than through systematic scientific inquiry, as is the case today. Although these early studies were based on observation, they highlight the importance of understanding human growth and development. In the modern world, these studies are conducted with a more scientific approach, such as understanding normal development, and identifying and diagnosing disorders. A well-known example is Buffon's 1777 publication on annual height measurements of the son of Montbeillard (Tanner, 1981). In the 19th century, numerous human growth studies were conducted in Europe, particularly focusing on child growth. During this period, the main emphasis was on measuring physical dimensions, referred to as "anthropometry", such as height, weight, and head circumference of children. One of the earliest and most notable instances is the work of Adolphe Quetelet (1796–1874), who studied cross-sectional height and weight of newborns and children (Quetelet, 1832;

Sarton, 1935). His key contribution is best known for creating the Quetelet index, which later became known as the Body Mass Index (BMI). Additionally, the study of adolescent growth spurt was advanced during this period, such as the work of Kotelmann (1879).

In the 20th century, the main focus included conducting long-term studies to track the growth and development of children over extended periods, known as "longitudinal studies", and developing standardised growth charts. The first emphasis was on measuring mean growth velocity (the rate of change in size over time) to enable healthcare providers to detect deviations from normal growth patterns early. This capability can facilitate early intervention and treatment of growth disorders such as nutritional deficiencies. In 1929, the Fels Longitudinal Study became well-known in this regard (Roche, 1992), as it was the world's largest and longest-running study of human growth.

On the other hand, the second focus was on constructing charts to allow for the comparison of an individual child's growth to a reference population, facilitating the identification of growth disorders and malnutrition. An earlier set of growth charts in this regard were Stuart and Meredith's growth charts (Stuart & Meredith, 1946). These charts were constructed based on data from 750 northwest European children aged 5 to 18 years. Several other standards were developed in the late 20th century, including the Tanner growth charts (Tanner et al., 1966), the NCHS growth charts (for Health Statistics et al., 1977), and the 2000 CDC charts. (Kuczmarski, 2000). The Tanner growth charts were specifically developed using British children to assess and monitor the physical development of children and adolescents, particularly during puberty. In contrast, the NCHS growth charts were constructed using data from American children. Meanwhile, the 2000 CDC charts were an updated version of the NCHS growth charts, incorporating data from more recent surveys.

In the 21st century, human growth studies have become broader than in previous eras. Studies now encompass not only physical growth measurements in a specific population but also have a more comprehensive and global focus. Various aspects are explored, including language development, cognitive development, social and emotional development, lifespan development, child development, adolescent development, gerontology, neurodevelopment, educational psychology, among others (Berk, 2013; Berk & Meyers, 2016; Bornstein & Lamb, 2015; Cavanaugh & Blanchard-Fields, 2015; Craik & Salthouse, 2015; Jensen, 2016; Papalia et al., 2004; Smith et al., 2015; Steinberg, 2011; Steinberg & Lerner, 2009). Therefore, the study of human growth and development aims to understand not only individuals but also how groups progress through various aspects of life stages, from birth to adulthood.

Childhood is one of these critical stages, as abnormal growth in children may influence their subsequent developmental phases. An important hypothesis related to this is the Barker hypothesis (Barker, 1990), which suggests that children with adverse nutrition, including prenatal nutrition, are at a higher risk of developing metabolic syndrome, obesity, diabetes, insulin resistance, hypertension, and other related conditions (Edwards, 2017). For instance, children who are overweight or obese face a higher risk of developing type 2 diabetes (T2D) in adulthood (Fang et al., 2019). Within childhood, there are sub-periods, often determined by factors like age. For example, the periods of development based on age are typically defined as five sub-periods by Berk and Meyers (2016), as shown in Table 2.1. These stages displays variations in growth and development. During infancy, for instance, children usually experience rapid weight gain, especially in the first six months (Berk & Meyers, 2016; Bukatko & Daehler, 1995). Their length also increases notably. In contrast, early childhood shows a steady but slower rate of growth in height and weight compared to infancy. Many facets of growth and development during the childhood stage have been examined, including physical growth, cognitive and language development, personality, and social development (Berk, 2013; Berk & Meyers, 2016; Smith et al., 2015).

Table 2.1: Child growth development sub-period stages

Age	**Sub-period stages**
Conception to birth	Prenatal period
Birth to 2 years	Infancy and toddlerhood
2 to 6 years	Early childhood
6 to 11 years	Middle childhood
11 to 18 years	Adolescence

Physical growth is one of the most critical aspects of child growth development because it reflects a child's overall health, well-being, and readiness to learn. The term "auxology" is employed to encompass a study of human physical growth and development, whether in children or adults (Hermanussen & Bogin, 2014). This growth influences their cognitive, emotional, and social development, establishes the foundation for lifelong health, and facilitates the achievement of developmental milestones. Monitoring and promoting healthy physical growth are essential for a child's success and future well-being. Typically, the study of this growth begins in infancy and continues into toddlerhood and beyond. Height has been a widely used measurement to represent child physical growth since ancient times (Hermanussen & Bogin, 2014), while weight became another importance metric with the advent of measurement technology. Both measurements are typically assessed in conjunction with age. However, in contemporary times, additional metrics such as body mass index, head circumference, and arm circumference have also been utilised (WHO Multicentre Growth Reference Study Group & de Onis, 2006). The details of physical growth measurements are discussed further in Section 2.3.

Another crucial period of growth in children is adolescence, which marks the transition between childhood and adulthood (Cameron & Bogin, 2012). Reproductive capacity, referred to as puberty (Marshall, 1978), is generally the defining sign of this period. However, the development of secondary sexual characteristics (e.g. growth of breasts and nipples in girls, and growth of facial hair in boys) and a notable growth spurt (i.e. a rapid and considerable increase in height and weight) are additional signs that coincide with puberty. In general, the development of physical growth in adolescence is markedly different compared to childhood. During adolescence, children experience rapid and considerable increases in their physical dimensions, such as height and weight (Cameron & Bogin, 2012; Tanner, 1970), but the specific timing of these growth spurts can vary widely among individuals. Specifically, the pattern of height growth typically follows three stages: early adolescence, mid-adolescence, and late adolescence. In early adolescence, children experience their initial growth spurt, which tends to be more pronounced in females than in males. The most rapid growth, known as peak height velocity (PHV), typically occurs in mid-adolescence. Note that the "velocity" term refers to the rate of change or speed at which growth occurs over a specific period of time. In late adolescence, the rate of growth begins to slow down. On the other hand, weight typically reaches its peak velocity later than height growth, and its changes tend to be slower compared to height growth. Several growth disorders may arise during adolescence (Allen & Cuttler, 2013; Sybert & McCauley, 2004; Taylor-Miller & Simm, 2017), including short stature, growth hormone deficiency, Turner syndrome, and others. Therefore, screening, surveillance, monitoring, and promoting healthy growth require appropriate tools.

As mentioned earlier, monitoring a child's physical development during both childhood or adolescence is of utmost importance. Therefore, health professionals and parents require reliable tools to monitor this phenomenon. Over the past decades, essential tools have emerged to assist in monitoring a child's physical growth, one of which is the child growth chart. To construct this chart, two approaches have been used: the "Prescriptive" and "Descriptive" approaches (Bertino et al., 2007). Indeed, there are differences between them in many aspects, such as purpose, data collection, statistical methods, interpretation, and validation. This section focuses on detailing the main differences between these approaches.

The prescriptive approach has been used to construct growth chart that represent an ideal standard of growth for a defined population, serving as a reference for what is considered typical or optimal growth in healthy children. Data used to construct charts based on this approach must come from a large and representative sample of healthy children. On the contrary, the descriptive approach has been used to describe the actual growth patterns observed in a population, regardless of health status, providing a snapshot of how children are growing in a specific population without prescribing what growth should be. When constructing growth charts using this approach, data should include a broad rage of children, including those with various health conditions or environmental factors.

One of the well-known child growth charts is constructed and published by the World Health Organization (WHO Multicentre Growth Reference Study Group & de Onis, 2006). Its primary aim to provide a universal standard as a reference for child growth. However, due to variations in growth patterns among different ethnic groups, many countries have developed their own growth charts to better reflect their unique population characteristics. For example, in the UK, the current growth chart is the UK-WHO growth charts constructed and published in 2010 (Royal College of Paediatrics and Child Health, 2013). More details about child growth charts are discussed in Section 2.4.

2.2 Child growth studies

Child growth studies play a crucial role in comprehending the essential process of physical development in children. These studies provide valuable insights into how children grow, the factors influencing their growth, and the establishment of growth standards and reference charts for healthcare and public health purposes. Two fundamental approaches employed in child growth studies are cross-sectional and longitudinal studies, each offering distinct advantages and insights into the multifaceted nature of child development.

2.2.1 Cross-sectional studies

Cross-sectional studies are a commonly used method in child growth research, involving data collection from a representative subset of the target population at a specific age to describe child growth and development. In essence, each growth measurement in the sample is collected from a different child at a single point in time, resulting in what is referred to as “cross-sectional data.” Due to the nature of data collection, data often lack the ability to capture dynamic or periodic features (Cole, 1994), which is a significant limitation of this approach. However, it offers convenience as it allows researchers to study child growth and development within a relatively short time-frame (Bukatko & Daehler, 1995).

2.2.2 Longitudinal studies

In contrast, longitudinal studies represent another method in which growth observations and relevant information are collected repeatedly over time from the same group of children. The data collected from this type of study is typically referred to as "longitudinal child growth data" (LCGD). This kind of data has the unique advantage of reflecting trends and changes in each individual child's growth over time. Due to this advantage, longitudinal studies are particularly valuable for addressing the fundamental scientific questions in child growth research more effectively than other study designs (Bukatko & Daehler, 1995; Cole, 1994; Grammer et al., 2013; Weiss & Ware, 1996). However, these benefits can be influenced by various characteristics inherent in the data, including different growth patterns, heterogeneity, and autocorrelation within individuals. As a result, appropriate statistical approaches are required to account for these characteristics.

2.3 Child growth measurements

In contemporary studies of child physical growth, various growth measurements are employed to monitor child growth and development. Each of these measurements serves a specific purpose in addressing the scientific inquiries of child growth studies. Broadly, these measurements can be categorised into two main types: raw and z-scale growth measurements.

2.3.1 Raw growth measurements

Raw growth measurements are fundamental and traditionally used metrics in the field of child development and healthcare. They serves as cornerstones for understanding the physical growth and changes in children over time (Bukatko & Daehler, 1995). Such measurements are directly observed physical attributes such as height, weight, and others, at specific ages without any adjustments or transformations. Essentially, they provide a straightforward and unaltered reflection of a child's physical growth status at a particular point in time. Consequently, in this section, a brief overview of common raw growth measurements is provided.

Length or height-for-age

"Length or height-for-age" is commonly used as one of the primary raw growth measurements. This metric directly reflects the increase in a child's physical size concerning their age and sex. In practice, "length" is the term used when measuring children up to about two years old because young infants and toddlers are often unable to stand unassisted or may not be able to maintain an upright position as needed. Beyond that age, the term "height" is generally adopted for measurements of both children and adults (WHO Multicentre Growth Reference Study Group & de Onis, 2006).

Weight-for-age

“Weight-for-age” is another fundamental raw measurement used in assessing child growth and pediatric healthcare. It involves measuring a child’s weight in relation to their age. This measurement assists healthcare professionals in determining whether a child is underweight, overweight, or within a healthy weight range for their age.

Weight-for-length or height

In cases where the chronological age of the child is unknown, “weight-for-length or height” is an alternative raw measurement that can be used to identify issues such as undernutrition (weight below the expected range for length or height and sex), overnutrition (weight above the expected range), and normal weight (Mwangome & Berkley, 2014).

Body mass index-for-age (BMI-for-age)

“BMI-for-age” is a growth measurement that combines two raw measurements: length (or height)-for-age and weight-for-age. It can still be classified as raw measurement because it directly utilises these primary data points. This type of growth measurement provides a comprehensive assessment of a child’s nutritional status and growth since it considers both their weight or height. There are two common versions of BMI calculation based on different units of measurement:

Using kilograms and metres (or centimetres):

$$\text{BMI}_1 = \frac{\text{Weight in kilogram}}{(\text{Height in meters})^2},$$

Using pounds and inches:

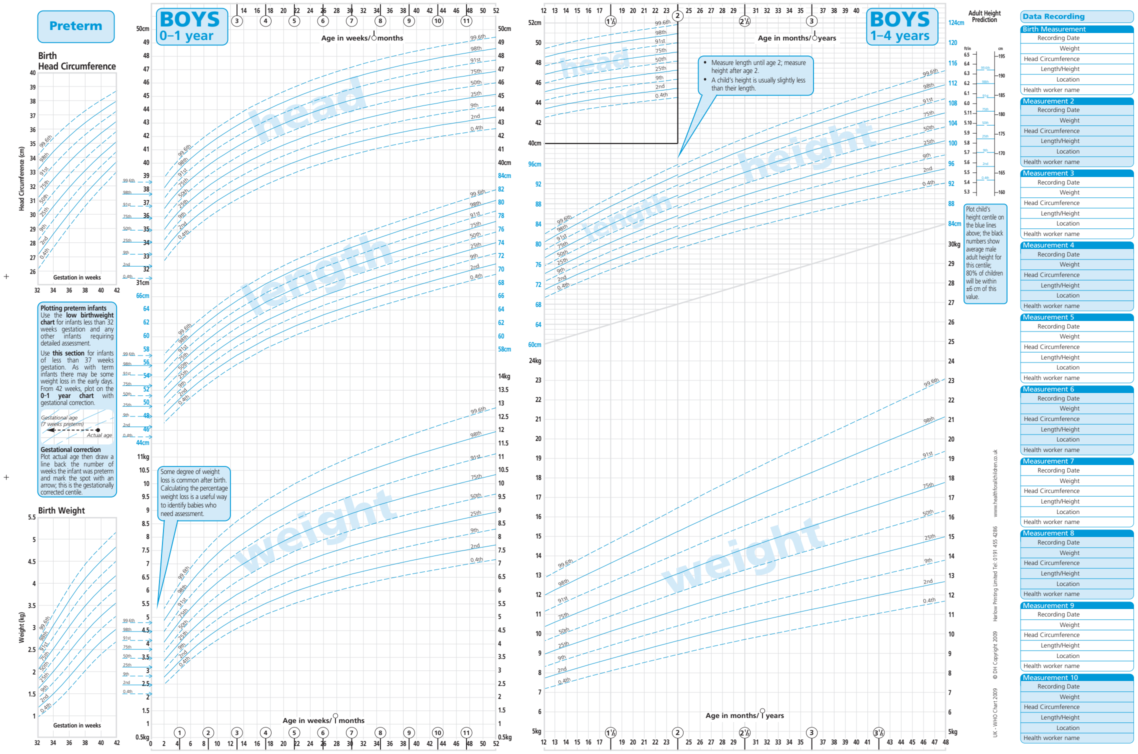
$$\text{BMI}_2 = \frac{\text{Weight in pound}}{(\text{Height in inches})^2} \times 703.$$

Other growth measurements

In addition to the raw measurements mentioned above, there are other raw measurement used to assess child growth development, such as head circumference-for-age, arm circumference-for-age, subscapular skinfold-for-age, triceps skinfold-for-age, and more (WHO Multicentre Growth Reference Study Group & de Onis, 2006). Different measurements serve distinct purposes in addressing specific scientific questions related to child growth development.

2.3.2 Standardised growth measurements

Standardised growth measurements, often referred to as Z-scores or standard deviation scores (SDS), are derived from raw growth measurements, as described above, but are adjusted for a child's age and sex (Cole, 1998). For example, when derived from weight-for-age or height-for-age, they are usually abbreviated as WAZ or HAZ, respectively. These measurements express a child's anthropometric data in relation to a reference population, typically represented as the mean (average) and standard deviation (variability) of that population's measurements for a given age and sex. Standardised growth measurements allow us to assess whether a child's growth falls within expected ranges and to identify deviations from typical growth patterns when compared with the population.



Plotting preterm infants
Use the **low birthweight chart** for infants less than 32 weeks gestation and any other infants requiring detailed assessment.

Use **this section** for infants of less than 37 weeks gestation. As with term infants there may be some weight loss in the early days. From 42 weeks, plot on the **0-1 year chart** with gestational correction.

Gestational age (7 weeks preterm)
Actual age

Gestational correction
Plot actual age then draw a line back the number of weeks the infant was preterm and mark the spot with an arrow; this is the gestationally corrected centile.

Some degree of weight loss is common after birth. Calculating the percentage weight loss is a useful way to identify babies who need assessment.

- Measure length until age 2; measure height after age 2.
- A child's height is usually slightly less than their length.

Plot child's height centile on the blue lines above; the black numbers show average male adult height for this centile; 80% of children will be within ±6 cm of this value.

Figure 2.1: BOYS UK-WHO growth charts 0-4 years. From *UK-WHO growth charts - 0-4 year*, by Royal College of Paediatrics and Child Health, 2023 (https://www.rcpch.ac.uk/resources/uk-who-growth-charts-0-4-years)

2.4 Child growth charts

Child growth charts are an essential tool for monitoring the physical development of children (Cole, 2012). These charts depict the distribution of specific physical growth measurements, as outlined in Section 2.3, that vary with age. Statisticians construct this distribution as a series of smooth curves with different percentiles at each age, based on child growth data from the sample of children. Generally, the chart format displays approximately nine curves, including the 3rd, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 97th percentiles. However, the composition may vary depending on the reference population (refer to Figure 2.1 and Figure 2.2 for examples). This chart is valuable for healthcare professionals to monitor and track a child's growth. For instance, if a 2-year-old boy's weight falls between the 25th and 75th percentiles for 2-year-olds, it indicates that this child is experiencing normal weight for his age, which is a sign of typical growth in terms of weight.

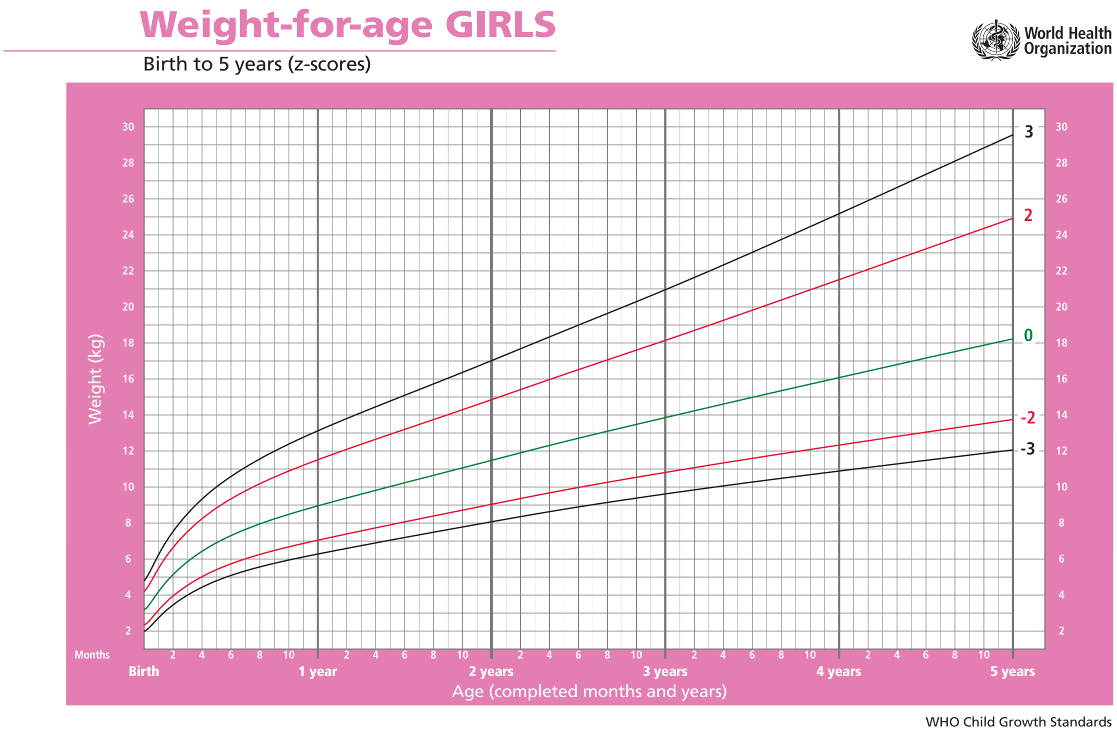


Figure 2.2: Weight-for-age GIRLS charts 0-5 years. From *Child Growth Standards: Weight-for-age*, by World Health Organization, 2023 (https://www.who.int/tools/child-growth-standards/standards/weight-for-age)

In the literature, a variety of statistical approaches are available for constructing this type of chart. These approaches often vary based on study designs or data structures. One commonly employed child growth study is the cross-sectional study, as outlined in Section 2.2. Among the statistical methods used with this data, the Lambda-Mu-Sigma (LMS) method stands out as the most widely recognised for fitting and summarising growth standards (Cole, 1988, 2012; Cole & Green, 1992). The acronym LMS (Lambda-Mu-Sigma) represents three key components of this method: 1) "L" denotes the changing skewness of the growth distribution, 2) "M" represents the median curve, and 3) "S" signifies the coefficient of variation of the growth distribution that changes across ages. These components are used to define the individual centile curves. The fundamental assumption underlying this method is the removal of skewness (the "L" curve) from the growth measurements at each age through power transformation methods, such as the Box-Cox power transformation, to achieve a normal distribution. However, cross-sectional growth charts have limitations; they are unable to account for dynamic or temporal features as each sample's growth data is collected only at a specific single-age point (Cole, 1994).

There are two types of child growth charts: growth standard and growth reference (Cameron & Bogin, 2012; Cole, 2012; Khadilkar & Khadilkar, 2011; WHO Multicentre Growth Reference Study Group & de Onis, 2006).

2.4.1 Attained size and growth terms

Before discussing child growth charts, it is important to clarify two key terms: "Attained Size" and "Growth". These terms relate to the growth measurements or growth metrics in child development and growth charts. "Attained Size" or "Size Attained" refers to the static measurements of a child's body dimensions at a specific point in time (Cameron & Bogin, 2012), typically collected in cross-sectional studies. It represents the child's physical development at that particular age or stage and is often compared to standard growth charts to assess whether the child's growth is within the expected range for their age and sex, commonly referred as *distance curves.* Nevertheless, attained size does not indicate the child's health status at the time of measurement.

On the other hand, "Growth" refers to the dynamic process of increasing in physical size and mass over time (Cameron & Bogin, 2012). It involves changes in growth measurements such as height, weight, and other body dimensions, representing continuous development and reflecting the child's health information. Therefore, it is a growth metric measured in longitudinal studies. Growth can be measured and tracked using growth charts to ensure that a child is developing appropriately according to standardized patterns, commonly referred to as *growth velocity.*

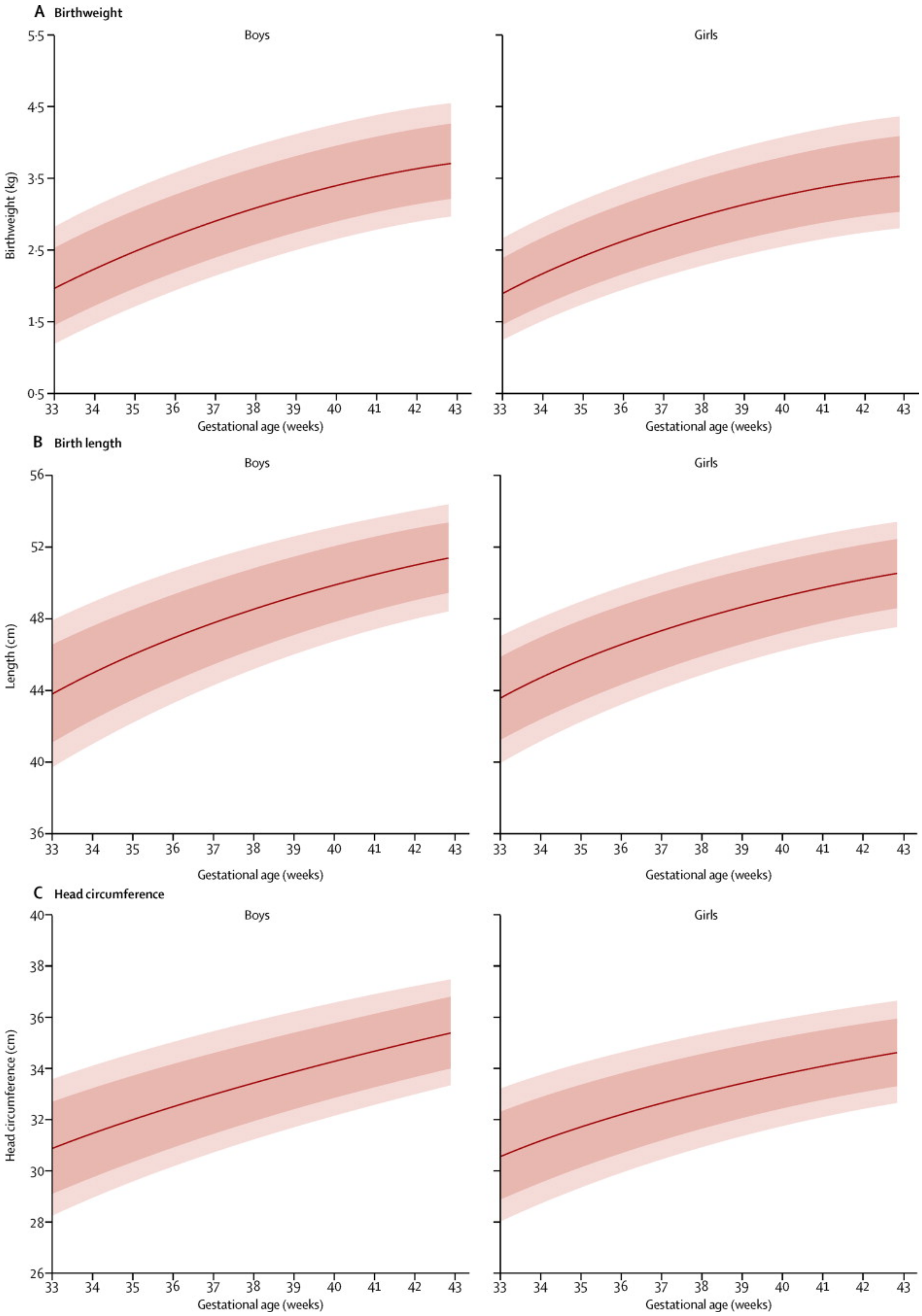


Figure 2.3: Standard growth charts: (A) Birthweight-for-age (weeks), (B) Birth length-for-age (weeks), and (C) Head circumference-for-age (weeks) from Villar et al. (2014)

2.4.2 Growth standard chart

A growth standard chart depicts the growth patterns of a specific population of children who are deemed healthy and experience optimal growth conditions. Such a chart is typically developed using data from a well-nourished and healthy population. Thus, it portrays a healthy growth pattern, and the standard indicates how children should ideally grow, as opposed to how they do in reality (Bertino et al., 2007). The most widely recognised growth standard charts are developed and published by the World-Health Organisation (WHO) (WHO Multicentre Growth Reference Study Group & de Onis, 2006) (see Figure 2.2 for an example). Furthermore, there are international standard charts developed specifically for certain populations, such as fetuses, newborn infants (see Figure 2.3), and the postnatal growth period of preterm infants (Villar et al., 2015; Villar et al., 2014), both implemented by the INTERGROWTH-21st Project.

2.4.3 Growth reference chart

A growth reference chart, while similar to a growth standard, represents the growth patterns of a wider population. This includes children from diverse backgrounds and with varying health statuses, rather than focusing solely on a healthy population (Cole, 2012). Typically, it is constructed using a reference group of children, often chosen to represent a particular geographic region and time period, such as a specific country in a given year. For instance, a recent growth reference chart in the UK is the UK-WHO growth charts, constructed and published in 2010 (Cole et al., 2011; Wright et al., 2010) (see Figure 2.1 for an example). A growth reference chart offers insights into typical child development and can be used to assess whether the measurements of individual children align with those of the reference group.

2.5 Risk factors against child/adolescent growth and development

Child growth and development depend on several factors, including biological, nutritional and environmental influences, social and emotional dynamics, behavioural traits, healthcare access and utilisation, and cultural or societal elements (Berk, 2013; Black et al., 2008; Bronfenbrenner, 1996; Denham, 2006; Rogoff, 2003; Shonkoff et al., 2000; Starfield et al., 2005). A deficiency in any one or more of these factors can impact a child's growth and development, preventing it from progressing as anticipated. Numerous experts have sought to investigate and address these risk factors to enhance their applicability, especially in preventive measures. For example, these risk factors can be divided into two primary categories: community or ecological risk factors and individual risk factors (Ali, 2013). The former encompasses risks directly related to the child, such as low birth weight, household size, undernutrition, and family economics, among others. In contrast, the latter pertains to external factors like poor sanitation, famine, endemic violence, and lack of accessible services, among other (see Table 1 in Ali (2013)). Meanwhile, the WHO classifies risk factors impacting child health into two groups based on child growth development stages: younger and older children (World Health Organization, n.d.). For younger children, there are two sub-stages: prior to birth and at the time of birth, with risks primarily involving low birth weight, malnutrition, breastfeeding, overcrowded conditions, unsafe drinking water, food, and poor hygiene practices. For the older children, the predominant risk factors affecting child health are a combination of environmental pollution (e.g. air pollution), malnutrition and limited physical activity.

The risk factors mentioned above can be directly and generally associated with child growth and development. Nonetheless, certain factors might exert an indirect influence. For example, diverse geographical areas can pose unique challenges in child development (Allel et al., 2021; Lu et al., 2016). Specifically, low-income countries might face more risk factors than middle- or high-income countries. Consequently, differences in geographical locations may results in varying environmental exposures, such as pollution and other environmental hazards, which can be considered as risk factors associated with child growth and development. Recognising this, several countries or regions have compiled risk factors specific to their contexts. In the UK, certain studies have catalogued the risk factors tied to child development to provide insights aimed at improving children's growth, such as the research by Sabates and Dex (2015). A significant contribution of this paper is its mapping of multiple risk factors to variables collected, such as parent–child interactions, family–child interactions, and the home environment in a longitudinal child growth study. Therefore, in this thesis, I will adapt this study as a basis for selecting variables associated with child growth. Details on this approach will be elaborated upon in Section 2.6.3.

2.6 Growing up in Scotland study

"Growing up in Scotland" (GUS) is a substantial longitudinal child growth study measuring children in Scotland born in 2002/03, 2004/05 and 2010/11. Directed and overseen by the Scottish Government, it is carried out by the Scottish Centre for Social Research. The GUS's primary aim is to gather new insights into children and their families to enrich our understanding of children and their families in various domains and also guide policy decisions aiming to improve the lives of children and their families. Moreover, this data is available for academic purposes in various sectors. The official website of the GUS study is: https://growingupinscotland.org.uk.

2.6.1 Study design

The GUS study consists of three sub-cohorts studies: Birth Cohort 1, Birth Cohort 2, and the Child Cohort. The first one is the largest cohort, comprising 5,217 children who were born between June 2004 and May 2005. The second cohort contains approximately 6,127 children born between March 2010 and February 2011. The last one includes 2,858 children born between June 2002 and May 2003. In the sample selection process, families were randomly chosen from the Child Benefit records provided by the Department for Work and Pensions (DWP) and HM Revenue and Customs. These families were invited to join the study through a letter. Data for the three cohorts were collected following the schedule outlined in Table 2.2.

Table 2.2: An overview of data collection for three cohorts of GUS

Child's age	Sweep (SW)	Cohort/Year of data collection			Weight	Length/Height
		Birth Cohort 1	Birth Cohort 2	Child Cohort	collected	collected
10 months	1	2005/06	2010/11	-	Yes	-
Age 2	2	2006/07	-	-	-	-
Age 3	3	2007/08	2013/14	2005/06	-	-
Age 4	4	2008/09	-	2006/07	Yes	Yes
Age 5	5	2009/10	2015/16	2007/08	-	-
Age 6	6	2010/11	-	2008/09	Yes	Yes
Age 8	7	2012/13	-	-	Yes	Yes
Age 10	8	2014/15	-	-	Yes	Yes
Age 12	9	2017/18	-	-	Yes	Yes
Age 14	10	2019/20	-	-	Yes	-
Age 17	11	2021/23	-	-	Yes	-

Note: Adapted from *Study design and timeline*, by Growing up in Scotland, 2023 (https://growingupinscotland.org.uk/study-design-and-timeline)

2.6.2 Data

In this thesis, only the Birth Cohort 1 is under consideration. As shown in Table 2.2, this cohort was designed to collect data in eleven sweeps, corresponding to the child's age at 10 months, 2 years, 3 years, 4 years, 5 years, 6 years, 8 years, 10 years, 12 years, 14 years and 17 years. It is crucial to emphasise that the initial analysis within this thesis covers data from Sweeps 1 to 10, in accordance with the request for access to data until the year 2020. At the end of the tenth sweep, the study included a total of 5719 children. However, no physical growth measurements, specifically raw weight, were collected during the second, third, or fifth sweep due to their focus on the child's primary caregiver and on two cognitive assessments, respectively. Consequently, the data used in this thesis consist of measurements from only seven sweeps. Some children were excluded due to entirely missing data over time in physical growth variables and other variables. Additionally, in particular raw height measurement, children whose raw height measurements decreased compared to previous sweeps were assumed to have human measurement errors and were also discarded. After removing data from these three sweeps and conducting data cleaning, 4563 children (2,326 males and 2,237 females) were included in the dataset for analysis. Table 2.3 shows the number of children across these seven sweeps. It is evident that there were children who dropped out of the study throughout the sweeps.

Several variables were collected in Birth Cohort 1, with the majority gathered through questionnaires. These variables can be categorised into various groups, including Household Information, Pregnancy and birth (only Sweep 1), Parental Support, Parenting Style, Childcare, Child Health and Development, Respondent's Employment, Partner's Employment, Household Income, Ethnicity and Religion, Household Accommodation and Circumstances, and Area-level Variables. In preparing the dataset for analysis, variables with a high percentage of missing values (> 50%) were excluded.

Table 2.3: Number of children across seven sweeps, with percentages in parentheses

Sweep	**All**	**Male**	**Female**
1	4018	2071 (51.5)	1947 (48.5)
4	2603	1315 (32.7)	1288 (32.1)
6	2409	1214 (30.2)	1195 (29.7)
7	2401	1216 (30.3)	1185 (29.5)
8	2352	1179 (29.3)	1173 (29.2)
9	1972	982 (24.4)	990 (24.6)
10	1634	788 (19.6)	846 (21.1)

Raw weight and raw length/height were categorised under Child Health and Development. The former was not collected during the second, third, or fifth sweeps, while the latter was only obtained during Sweeps 5 to 9. Additionally, to obtain the standardised measurements (Z-scores), the raw measurements need to be converted using a growth reference population. These raw and standardised measurements are the primary focus in the analytical context of this thesis.

In this thesis, the process to obtain the Z-score is as follows. The UK-WHO growth reference, which excludes preterm data, was employed as the reference growth population (Cole & Green, 1992; Cole et al., 2011; Wright et al., 2010). This reference, introduced in 2009, replaced the previous UK1990 growth reference charts and is used to assess the growth development of children in the UK. It was fitted using the LMS (Lambda-Mu-Sigma) method, as mentioned in Section 2.4, and is based on three datasets: 1) birth data from the British 1990 growth reference, 2) the WHO growth standard from 2 postnatal weeks to 4 years, and 3) the British 1990 reference from 4 to 20 years. The conversion process was performed using the `LMS2z` function available in the `sitar` package for R

(Cole, 2022). In this function, the Z-scores are calculated by the formula:

$$Z = \frac{(X/M)^L - 1}{L \cdot S},$$

where X is the raw growth measurement (e.g. weight, height), M is the median of the distribution for the specific age (years) and sex, L is the power in the Box-Cox transformation, and S is the generalised coefficient of variation. It is important to note that the data used for the conversion must include children with at least one measurement of weight, and there should be no missing data for age (in years) or sex variables. The interpretation of the Z-score is straightforward. For instance, if the Z-score equals zero, it indicates that the child's measurement is typical or average compared to the UK-WHO reference population. In contrast, if this score is above zero, it means that the child is larger or taller (depending on the measurement) or compared to the average child of the same age and sex in the UK-WHO reference population. Conversely, if the Z-score is below zero, it implies that the child is smaller or shorter (depending on the measurement) compared to the average child.

2.6.3 Mapping variables in GUS as potential risk factors

As previously mentioned, Birth Cohort 1 contains numerous variables. In this thesis, I adopt the framework outlines in "Multiple Risk Factors in Young Children's Development" (MRFYD) by Sabates and Dex (2015) to select potential variables likely associated with child development. This framework utilises an ecological model of child development, helping to pinpoint risk factors associated with child development. In this framework, it categorised variables from the UK Millennium Cohort Study (MCS) (Connelly & Platt, 2014) into three main dimensions: "Proximal family processes", "Distal family variables" and "characteristics". The first dimension includes variables detailing parent-child interactions. The second encompasses variables connected to parental characteristics that influence the child mainly through family-child interactions. The final dimension relates to variables associated with the child's environment during their growing up. Additionally, each variable from the MSC study was classified into ten risk types: Depression, Physical Disability, Substance Misused, Alcohol, Domestic Violence, Financial Stress, Worklessness, Teenage Parenthood, Basic Skills, and Overcrowding.

The MRFYD framework offers several key benefits related to understanding and addressing risk factors that affect young children's development. Firstly, it provides a comprehensive analysis of risk factors, detailing various elements that can impact a child's development, including socioeconomic status, parental education, family structure, and health conditions. Secondly, its grounding in robust data and methodological rigor ensures that the

Table 2.4: Variables in the GUS study mapped to the multiple potential risk indicators following Sabates and Dex (2015)

Type of risk	Variable from the GUS study	Dimension
Depression	1. In general, would you say your health is excellent, very good, good, fair, or poor?	Proximal
Physical Disability	1. Do you have any health problems or disabilities that have lasted or are expected to last more than a year?	Proximal
Substance Misuse	1. During your pregnancy with child did you smoke cigarettes?	Proximal
	2. Thinking back to when you were pregnant with child, which of these best describes how often you usually drank then?	Proximal
Alcohol	1. How often (current) do you have an alcoholic drink?	Distal
Domestic Violence	-	-
Financial Stress	1. Scottish Index of Multiple Deprivation (SIMD) 2006 Quintiles	Distal
	2. Equivalised income	Distal
Worklessness	1. Do you currently have a job, either as an employee or self-employed?	Distal
Teenage Parenthood	1. Age of mother at 1st child's birth	Distal
Basic Skills	-	
Overcrowding	1. Number of people in household	Distal

framework provides evidence-based findings, making it easier for policymakers, educators, and healthcare professionals to implement effective interventions. Importantly, the framework was studied using longitudinal data, which is the same approach used in this thesis, providing insights into the long-term effects of early risk factors. Following this framework, I applied their mapping method to select potential risk variables from the GUS study for analysis in Chapter 6. Variables were selected based on similar or closely related meanings. However, only eight of these risk types could be mapped using the GUS study variables. Some variables were excluded due to a high number of missing values. Table 2.4 presents the final set of variables from the GUS study, alighted with the "Multiple Risk Factors in Young children's Development" framework. Although, this framework offers several benefits as mentioned previously, it has some limitations. For instance, an important factor such as "birth spacing" (the interval between the births of siblings within a family) is not included in this framework. This is because its primary focus is on broader socioeconomic and family structure variables, including parental education, family income, and health conditions.

Table 2.5: Final set of variables from the GUS study after mapping with MRFYD, used for analysis in this thesis

Symbol	Short variable name	Variable	Type of variable
X_1	Age	Age of person 1 - study child	Continuous
X_2	Birth weight	Birth weight in grams	Continuous
X_3	Low birth weight	Low birth weight	Categorical
X_4	Ethnicity of a child	Ethnicity of a child	Categorical
X_5	Child's health in general	How is child's health in general?	Categorical
X_6	Number of accidents or injuries of child	Number of accidents or injury for which child has been taken to the doctor, health centre, or hospital	Continuous
X_7	Child's birth order	Study child's birth order	Continuous
X_8	Mother's marital status	Marital status of Mother	Categorical
X_9	Urban-rural classification	Urban-rural classification	Categorical
X_{10}	Household size	Number of people in household	Continuous
X_{11}	Mother's age at first child's birth	Age of mother at 1st child's birth	Categorical
X_{12}	Respondent's alcoholic drink	How often (current) do you (respondent) have an alcoholic drink?	Categorical
X_{13}	Respondent's current health	In general, would you say your health is excellent, very good, good, fair, or poor?	Categorical
X_{14}	Smoking cigarettes while pregnant	During your pregnancy with child did you smoke cigarettes?	Categorical
X_{15}	Drinking alcohol while pregnant	Thinking back to when you (mother) were pregnant with child, which of these best describes how often you usually drank then?	Categorical
X_{16}	Respondent's health problem in a year	Do you have any health problems or disabilities that have lasted or are expected to last more than a year?	Binary
X_{17}	Equivalised income	Equivalised income†	Continuous
X_{18}	Deprivation	Scottish Index of Multiple Deprivation (SIMD) 2006 Quintiles‡	Categorical
X_{19}	Respondent's current job	Do you currently have a job, either as an employee or self-employed?	Binary

Note: † In this context, the "equivalised income" is a measure used to account for the differing needs of households of various sizes and compositions. In Scotland, it is calculated by adjusting a household's total income to reflect the number of occupants and their specific needs.

Note: ‡ In the GUS data, data zones are grouped into quintiles, indicating that residents whose postcodes falls into the fifth quintile live in one of the 20% most deprived areas in Scotland.

Hence, in this thesis, the variables listed in Table 2.4 are included as potential variables or potential risk factors that will be used to study their association with child growth measurements. Furthermore, I have considered other potential variables, including child demographic variables and those related to the mother (e.g. Birth weight, Low weight, Marital status of Mother etc.). Table 2.5 presents all these variables and their types.

2.6.4 Data exploration

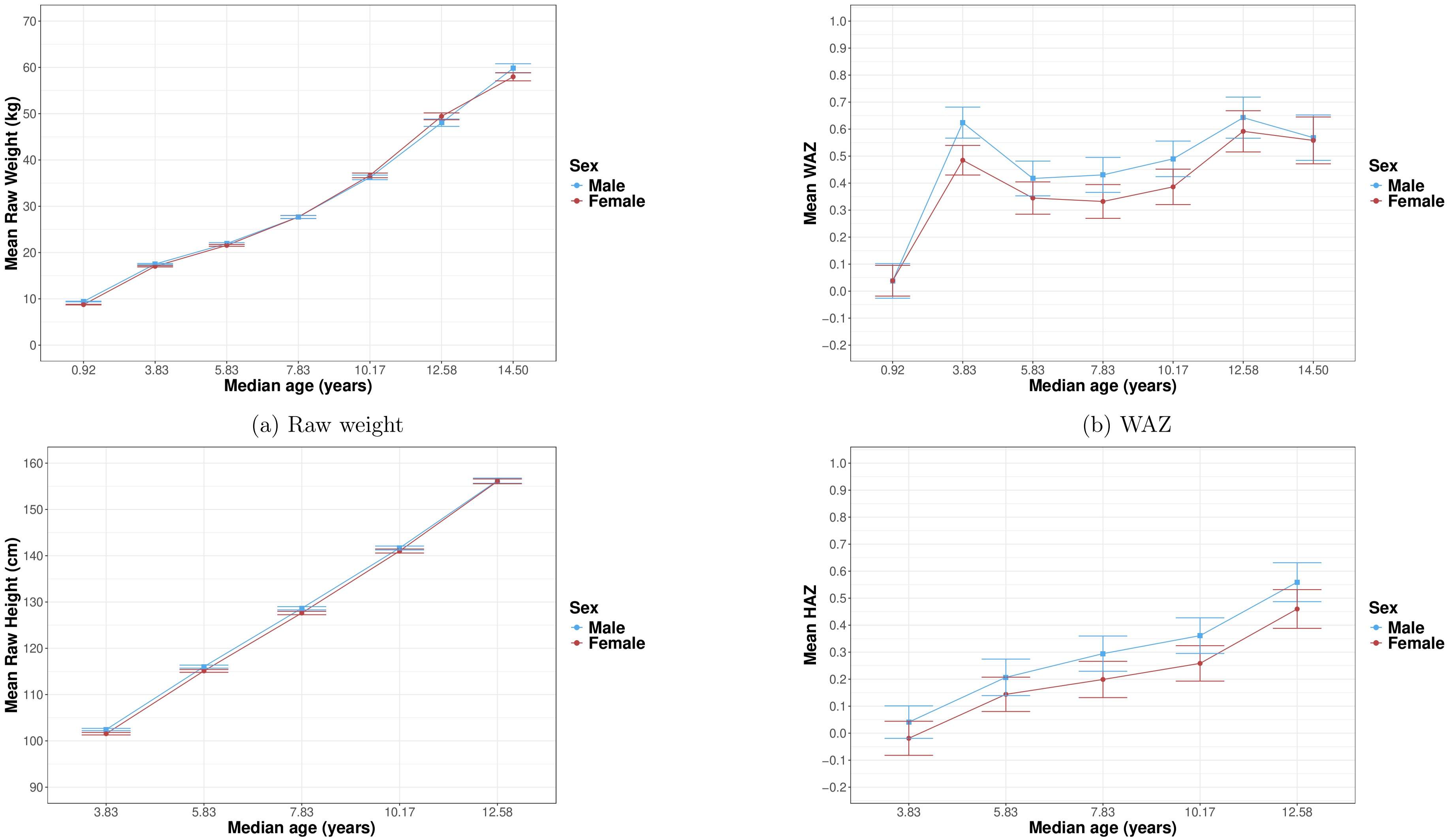
In this section, the GUS dataset is explored. Given the distinct differences in growth patterns and rates between boys and girls, the data will be summarised separately by sex. Table 2.6 provides a summary of the four physical growth measurements for each follow-up sweep, for males and females, respectively. In general, males appear to have slightly higher averages and medians for the two raw measurements (weight and height) compared to females. Meanwhile, the difference between males and females is obvious in the case of the standardised scales (WAZ and HAZ). Therefore, these summaries suggest that the differences in growth patterns and rates between boys and girls persist within this population.

When considered in the mean profile plots (refer to Figure 2.4), the male and female groups exhibit similar trends of mean weight measurements (see Figure 2.4 (a)) throughout seven sweeps but with a slightly different pattern. On average, during the first year of life, males tend to be slightly heavier than females, but the difference seems not to be usually significant. Between the ages of 2 to 6 years, growth rates slow down compared to the first year, and females seem to start catching up with males in weight during this stage. Weight gain continues at a more gradual pace between ages 7 to 10 years. During ages 11 to 14 years, males appear to experience a more noticeable increase in weight, while females also experience growth, which it tends to be less dramatic in terms of weight gain compared to males.

The plot of Weight-for-Age Z-scores (WAZ) for both males and females, when considering the mean profile (see Figure 2.4 (b)), shows a parallel trend with similar mean WAZ scores at the first sweep (corresponding to the child's age at 10 months) and the last sweep (corresponding to the child's age at 14 years). This means that, on average, both males and females present similar WAZ scores at these two time points, but there may be variations in between. However, this observation is somewhat unexpected; although WAZ values are normalised by sex, which typically results in similar scores for males and females, the data may suggest variations that warrant further investigation.

Table 2.6: Summary of physical growth measurements, the GUS dataset

SW	Mean age (SD)	Median age (IQR)	Males								Females							
			Mean Weight (kg) (SD)	**Median Weight (IQR)**	**Mean Height (cm) (SD)**	**Median Height (IQR)**	**Mean WAZ (SD)**	**Median WAZ (IQR)**	**Mean HAZ**	**Median HAZ (IQR)**	**Mean Weight (kg) (SD)**	**Median Weight (IQR)**	**Mean Height (cm) (SD)**	**Median Height (IQR)**	**Mean WAZ**	**Median WAZ (IQR)**	**Mean HAZ**	**Median HAZ (IQR)**
1	0.88	0.92	9.42	9.36	-	-	0.04	0.05	-	-	8.75	8.67	-	-	0.04	0.06	-	-
	(0.05)	(0.83 - 0.92)	(1.29)	(8.70 - 10.09)			(1.14)	(-0.63 - 0.81)			(1.20)	(8.00 - 9.49)			(1.08)	(-0.58 - 0.72)		
4	3.85	3.83	17.52	17.30	102.45	102.50	0.62	0.61	0.04	0.06	17.03	16.80	101.55	101.60	0.48	0.49	-0.02	-0.03
	(0.04)	(3.83 - 3.83)	(2.25)	(16.10 - 18.70)	(4.23)	(99.90 - 105.20)	(0.98)	(-0.01 - 1.22)	(1.02)	(-0.58 - 0.73)	(2.32)	(15.50 - 18.20)	(4.55)	(99.00 - 104.40)	(0.93)	(-0.10 - 1.04)	(1.07)	(-0.65 - 0.65)
6	5.85	5.83	21.93	21.50	116.04	116.20	0.42	0.37	0.21	0.22	21.55	21.10	115.12	115.00	0.34	0.34	0.14	0.14
	(0.04)	(5.83 - 5.83)	(3.23)	(20.00 - 23.40)	(5.25)	(113.00 - 119.58)	(1.04)	(-0.20 - 1.04)	(1.09)	(-0.44 - 0.93)	(3.28)	(19.30 - 23.20)	(5.00)	(111.90 - 118.50)	(0.98)	(-0.30 - 0.97)	(1.04)	(-0.52 - 0.85)
7	7.86	7.83	27.69	26.80	128.65	128.80	0.43	0.40	0.29	0.30	27.68	26.80	127.63	127.60	0.33	0.30	0.20	0.21
	(0.06)	(7.83 - 7.92)	(5.02)	(24.50 - 29.70)	(5.67)	(125.10 - 128.80)	(1.04)	(-0.19 - 1.04)	(1.05)	(-0.38 - 1.02)	(5.10)	(24.00 - 30.20)	(5.89)	(123.60 - 131.40)	(1.01)	(-0.39 - 1.00)	(1.08)	(-0.53 - 0.21)
8	10.20	10.17	36.23	34.80	141.67	141.60	0.49	0.47	0.36	0.37	36.67	35.40	141.02	140.90	0.39	0.40	0.26	0.25
	(0.30)	(10.00 - 10.42)	(7.78)	(31.00 - 39.60)	(6.64)	(137.50 - 146.15)	(1.03)	(-0.16 - 1.18)	(1.04)	(-0.30 - 1.09)	(8.02)	(31.00 - 40.90)	(6.96)	(136.20 - 145.70)	(1.04)	(-0.35 - 1.09)	(1.05)	(-0.47 - 0.96)
9	12.60	12.58	48.04	46.40	156.21	156.15	0.64	0.65	0.56	0.56	49.43	48.20	156.07	156.20	0.59	0.58	0.46	0.48
	(0.31)	(12.33 - 12.75)	(11.44)	(39.55 - 54.00)	(8.21)	(150.50 - 161.50)	(1.10)	(-0.11 - 1.41)	(1.05)	(-0.16 - 1.26)	(10.71)	(41.80 -55.00)	(7.52)	(151.00 - 161.40)	(1.10)	(-0.12 - 1.28)	(1.05)	(-0.21 - 1.18)
10	14.50	14.50	59.81	58.65	-	-	0.57	0.57	-	-	57.98	56.20	-	-	0.56	0.48	-	-
	(0.28)	(14.33 - 14.75)	(12.51)	(51.08 - 65.90)			(1.06)	(-0.11 - 1.21)			(11.65)	(50.10 - 63.40)			(1.15)	(-0.23 - 1.23)		



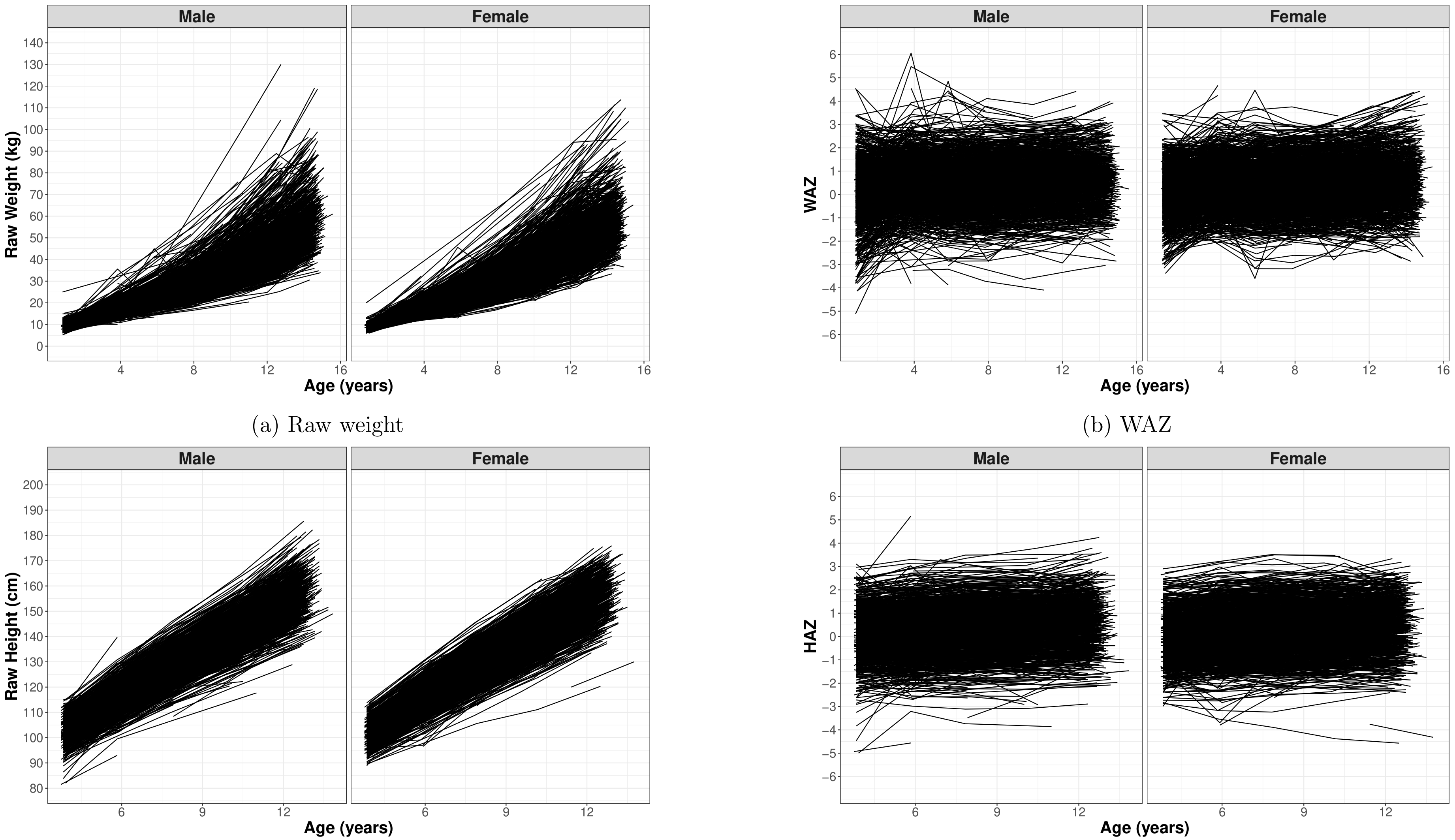
(a) Raw weight

(b) WAZ

(c) Raw height

(d) HAZ

Figure 2.4: Mean growth measurement profiles with 95% CI bars over the median age (years) by sex, GUS



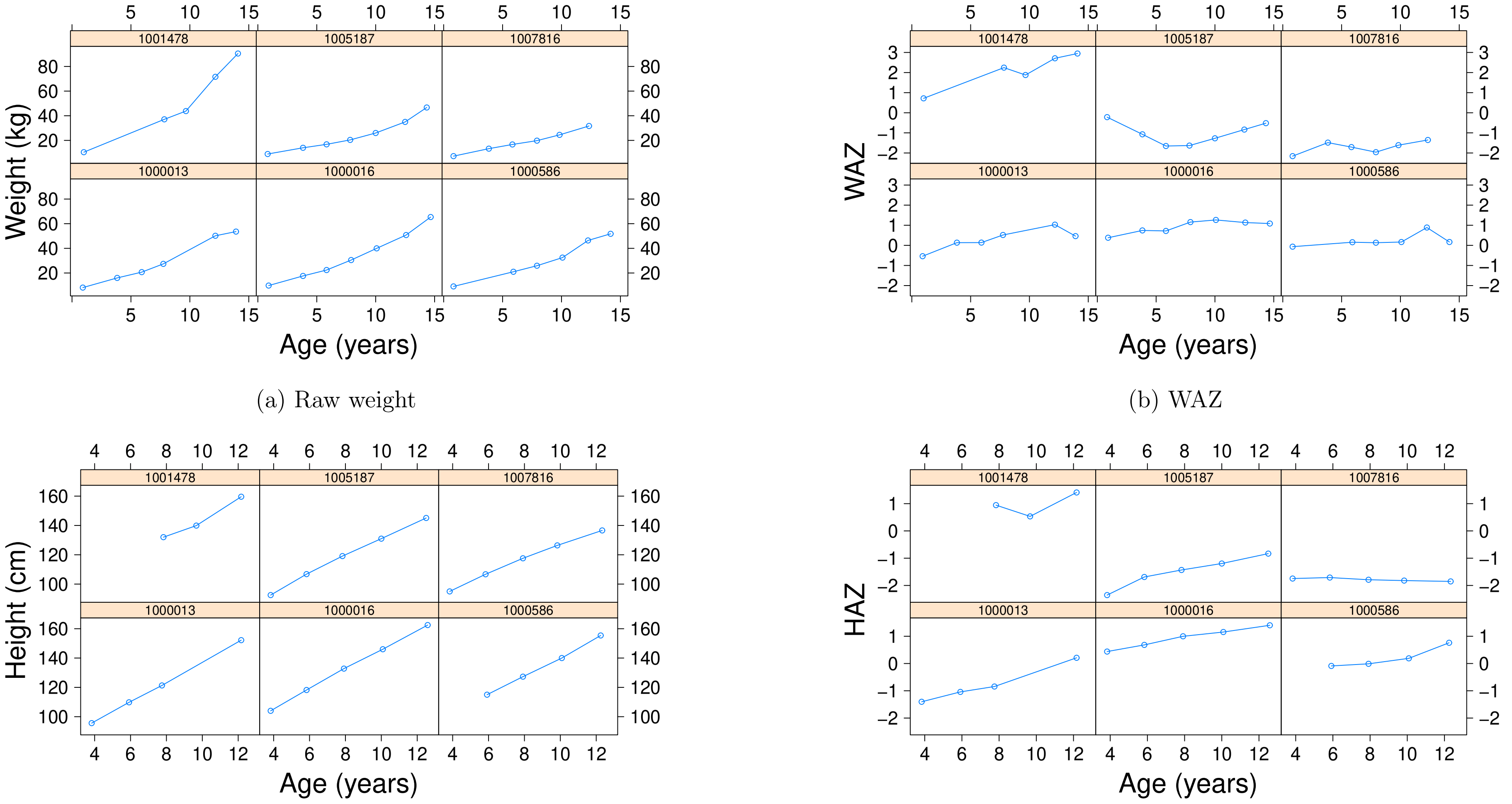
(a) Raw weight

(b) WAZ

(c) Raw height

(d) HAZ

Figure 2.5: Individual, Raw Weight, Raw Height, WAZ, and HAZ trajectories from the GUS data by sex



(a) Raw weight

(b) WAZ

(c) Raw height

(d) HAZ

Figure 2.6: Individual, Raw Weight, Raw Height, WAZ, and HAZ trajectories from the GUS data by sex

Regarding the measurements related to height, both males and females exhibit a similar growth trend. On average, males appear to have slightly higher height growth patterns than females between the ages of 4 to 10 years but seem to be very similar in ages of 12 years (see Figure 2.4 (c)). Notably, both groups have different HAZ scores throughout five sweeps (ages 4 to 12 years). Such divergence in HAZ scores is unexpected, analogous to the findings with WAZ.

Figure 2.5 presents plots illustrating how each child's growth measurement points change over time. These plots also demonstrate that a child's growth in the four growth measurements follows a non-linear pattern as they age. They reveal variations in these measurements both between and within the children. In particular, the between-individual variability in both raw measurements was small in early childhood and increased with the child's age (see Figures 2.5 (a) and (b)), while the standardised scales show a contrasting trend (see Figures 2.5 (b) and (d)). This suggests that the GUS data explicitly show heterogeneous variability as a common feature of longitudinal data.

In Figure 2.6, the trajectories of six random children are plotted. It is evident that each child has their own growth trajectory in the four growth measurements, and these trajectories tend to follow a non-linear pattern. Furthermore, it shows that each child may exhibit differences in two key features: the observation times and the number of observations per individual (or individual size).

To clarify the observation times feature, time observed values were plotted for each scheduled age (Figure 2.7). It is evident that time observed values (represented by black points) closely align with the scheduled age (years - vertical red line) during the initial ages. However, as age increases, there is a tendency for observation times to become more variable.

Another common feature of longitudinal data that should be investigated is the serial correlation in repeated measurements on an individual. An empirical variogram is a well-known plot that is helpful for this purpose. This plot is used to assess spatial dependence or variability between data points at different time points or locations. In this case, the empirical variogram $\gamma(k)$ is defined as one-half of the squared differences between pairs of measurements, representing variability in the differences, associated with the corresponding time distances (or lag distance, k) (Diggle, 2005):

$$\gamma_{ij}(k) = \frac{1}{2}\Big(r_i(j) - r_i(j+k)\Big)^2,$$

where $r_i(j)$ represents the value of the process at time j for individual i, and $r_i(j+k)$ is the value at a time a distance k away from j for individual i. The variogram is plotted

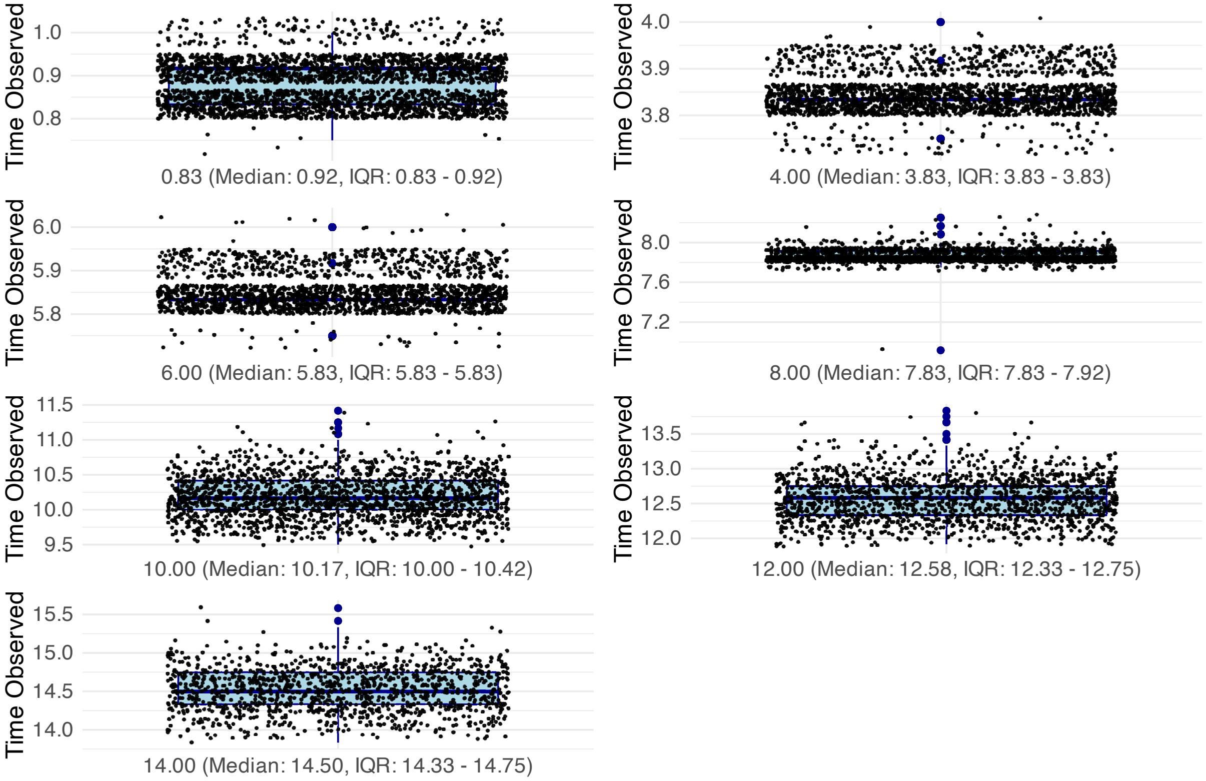


Figure 2.7: Box-plots of time observed values varies by scheduled ages at 0.83, 4.00, 6.00, 8.00, 10.00, 12.00, and 14.00. Note that labels on the x-axis represent a scheduled age along with the median and IQR.

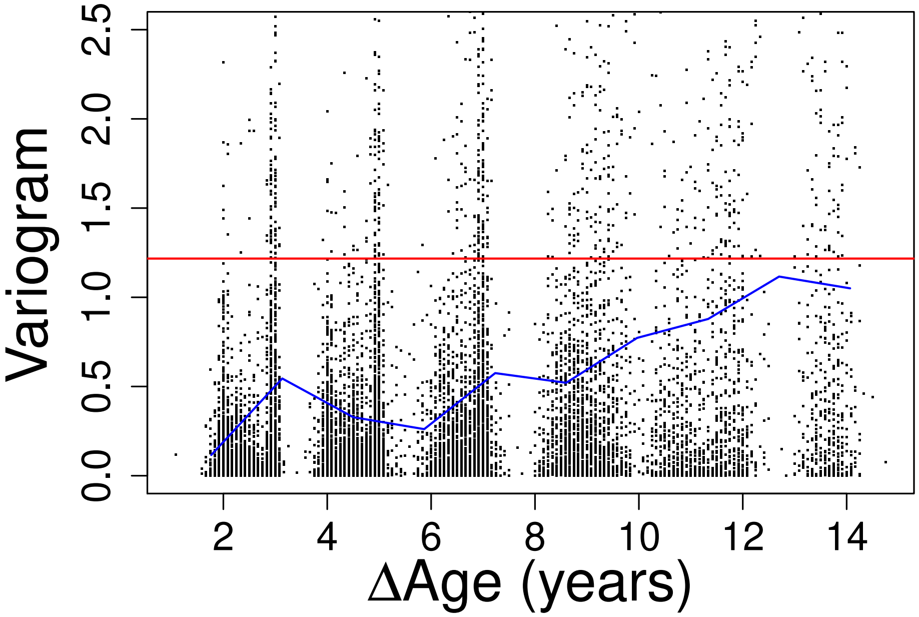


Figure 2.8: Empirical variogram for WAZ in males, with the *red horizontal line* representing the maximum level (or sill) of variance observed and the *blue line* representing the averages of half-squared differences observed at each time lag between pairs of measurements (*black dots*).

for averages of each k for $\gamma_{ij}(k)$ for all i. To interpret this plot, if the variogram values (y-axis) approach the sill (the horizontal line) as the lag distance (x-axis) increases, it suggests strong spatial or temporal dependence. Figure 2.8 presents an example of the empirical variogram plot for WAZ in males, illustrating that the averages of half-squared differences observed ($\gamma_{ij}(k)$) increases with lag k. Therefore, it suggests that this dataset provides evidence of serial correlation.

In addition, I have explored the variables mentioned in Section 2.6.3. In longitudinal data, variables can be classified into two groups. The first group comprises "time-invariant variables", which do not vary with time and maintain consistent values across different time points. The second group consists of variables whose values change over time, often referred to as "time-varying variables". Therefore, at this point, each variable will be summarised by these two groups.

Table 2.7 presents the time-varying and time-invariant continuous variables for the GUS data, both in aggregate and stratified by sex. On average, as anticipated, household size appeared to increase as the sweep number increased, while the number of child accidents or injuries remained relatively steady. There were increases in average equivalised incomes between Sweep 1 to Sweep 7, followed by a drop at Sweep 8 and another increase afterward. Males appeared to have a higher birth weight than females on average.

Table 2.7: Summary of time-varying and time-invariant continuous variables for the GUS data

Variable			**SW1**	**SW4**	**SW6**	**SW7**	**SW8**	**SW9**	**SW10**
Household size	All	Mean (SD)	3.69 (0.99)	3.90 (0.88)	4.01 (0.90)	4.26 (0.97)	4.38 (1.03)	4.37 (1.02)	4.37 (1.01)
	Male	Mean (SD)	3.67 (0.97)	3.91 (0.89)	4.04 (0.88)	4.25 (0.93)	4.41 (1.00)	4.39 (0.97)	4.39 (0.96)
	Female	Mean (SD)	3.70 (1.02)	3.89 (0.88)	3.99 (0.91)	4.26 (0.99)	4.36 (1.06)	4.35 (1.06)	4.35 (1.05)
Number of accidents or injuries of child	All	Mean (SD)	0.11 (0.36)	0.20 (0.52)	0.17 (0.47)	0.30 (0.69)	0.23 (0.58)	0.22 (0.53)	0.22 (0.54)
	Male	Mean (SD)	0.13 (0.40)	0.22 (0.48)	0.19 (0.46)	0.33 (0.79)	0.24 (0.55)	0.22 (0.54)	0.22 (0.54)
	Female	Mean (SD)	0.09 (0.32)	0.18 (0.56)	0.14 (0.47)	0.27 (0.58)	0.23 (0.61)	0.22 (0.51)	0.22 (0.53)
Equivalised income[a]	All	Mean (SD)	21.36 (12.60)	26.64 (11.71)	27.19 (11.69)	29.81 (16.61)	27.44 (12.01)	26.61 (12.09)	33.04 (16.37)
	Male	Mean (SD)	21.24 (12.49)	26.47 (11.60)	26.87 (11.64)	29.89 (16.75)	27.29 (11.89)	26.76 (12.06)	32.92 (15.99)
	Female	Mean (SD)	21.48 (12.73)	26.81 (11.82)	27.52 (11.73)	29.73 (16.47)	27.60 (12.13)	26.46 (12.13)	33.15 (16.72)
Birth weight (grams)	All	Mean (SD)	3425.48 (582.80)						
	Male	Mean (SD)	3499.82 (575.03)						
	Female	Mean (SD)	3349.91 (580.95)						

Note: [a] All values are expressed in thousands.

In Table 2.8, a summary of time-invariant categorical variables is presented. Within this dataset, the proportion of males was slightly higher than females. As expected, the proportion of White children was higher than that of other ethnicities. There were a few children with low birth weight, accouting for 6.1% of the total. Additionally, more than half of the mothers had their first child at the age of over 30 years old. When considering the time of pregnancy, it is noteworthy that 22.3% of mothers had smoked, and approximately 27% had consumed alcohol.

Table 2.9 shows a summary of time-varying categorical variables for the entire GUS dataset. The summary indicates that the majority of children were in very good health consistently across sweeps, with only 3.1% to 6.5% of children found to have bad or very bad health. The majority of mothers had a marital status categorised as "Married and living with husband". About 60% of children lived in urban areas, whether large or other urban. Parents mostly consumed alcohol once a month or less, with some consuming it once a week. The current health of parents appeared to be good to excellent, with only 9% to 13.5% reporting fair or poor health. When considering only the past year, over 80% of parents did not report any health problems. The majority of parents still had their jobs. Additionally, it is noteworthy that approximately 60% of the children lived in the top three least deprived quintiles.

When stratified by sex, the summaries showed a similar trend compared to aggregated data, and both summaries are presented in Tables 2.10 and 2.11. The percentages for each group across different variables remained consistent over the sweeps in both male and female datasets. However, the percentage of respondents with no current job was higher in the first sweep (SW1) compared to the subsequent sweeps. This trend was observed in both the male and female datasets, indicating that after 2005/06 (the data collection year for SW1), respondents were more likely to have a job.

Table 2.8: Time-invariant categorical variables for GUS data

Variable	Level	Children (%)	Male (%)	Female (%)
Sex		4563 (100.0)	2326 (51.0)	2237 (49.0)
Ethnicity of child	White (reference)	4418 (96.8)	2249 (50.9)	2169 (49.1)
	Other	145 (3.2)	77 (53.1)	68 (46.9)
Low birth weight	Yes	280 (6.1)	132 (47.1)	148 (52.9)
	No (reference)	4283 (93.9)	2194 (51.2)	2089 (48.8)
Mother's age at first child's birth	$<$ 20 years old	243 (5.3)	133 (54.7)	110 (45.3)
	20 - 29 years old (reference)	1789 (39.2)	911 (50.9)	878 (49.1)
	$\geq$ 30 years old	2531 (55.5)	1282 (50.7)	1249 (49.3)
Smoking cigarettes while pregnant	Yes	1017 (22.3)	514 (50.5)	503 (49.5)
	No (reference)	3546 (77.7)	1812 (51.1)	1734 (48.9)
Drinking alcohol while pregnant	Everyday, 3 - 6 times a week	32 (0.7)	17 (53.1)	15 (46.9)
	1 - 2 times a week	147 (3.2)	74 (50.3)	73 (49.7)
	2 - 3 times a month	216 (4.7)	109 (50.5)	107 (49.5)
	<once a month	850 (18.6)	453 (53.3)	397 (46.7)
	Never - did not drink at all (reference)	3318 (72.7)	1673 (50.4)	1645 (49.6)

Note: The term "reference" (indicated within the parentheses) represents the reference categories which will be utilised in the model categories.

Table 2.9: Summary of time-varying categorical variables for GUS data

Variables	**Level/Measurements**	**SW1**	**SW4**	**SW6**	**SW7**	**SW8**	**SW9**	**SW10**
		Children (%)	Children (%)	Children (%)	Children (%)	Children (%)	Children (%)	Children (%)
Child's health in general	Very good (reference)	3014 (75.0)	1908 (73.3)	1865 (77.4)	1906 (79.4)	1934 (82.2)	1515 (76.8)	1216 (74.4)
	Good	785 (19.5)	556 (21.4)	461 (19.1)	417 (17.4)	353 (15.0)	387 (19.6)	340 (20.8)
	Bad & Very bad	219 (5.5)	139 (5.3)	83 (3.4)	78 (3.2)	65 (2.8)	70 (3.5)	78 (4.8)
Mother's marital status	Single, that is, never married	1504 (37.4)	744 (28.6)	612 (25.4)	554 (23.1)	468 (19.9)	382 (19.4)	246 (15.1)
	Married and living with husband (reference)	2257 (56.2)	1703 (65.4)	1622 (67.3)	1649 (68.7)	1635 (69.5)	1351 (68.5)	1159 (70.9)
	Other	257 (6.4)	156 (6.0)	175 (7.3)	198 (8.2)	249 (10.6)	239 (12.1)	229 (14.0)
Urban-rural classification	Large urban (reference)	1395 (34.7)	882 (33.9)	773 (32.1)	778 (32.4)	763 (32.4)	633 (32.1)	519 (31.8)
	Other urban	1271 (31.6)	859 (33.0)	799 (33.2)	815 (33.9)	767 (32.6)	646 (32.8)	529 (32.4)
	Small, accessible towns	412 (10.3)	255 (9.8)	247 (10.3)	239 (10.0)	236 (10.0)	215 (10.9)	178 (10.9)
	Small, remote towns	135 (3.4)	83 (3.2)	75 (3.1)	84 (3.5)	89 (3.8)	57 (2.9)	60 (3.7)
	Accessible rural	588 (14.6)	350 (13.4)	341 (14.2)	313 (13.0)	326 (13.9)	275 (13.9)	246 (15.1)
	Remote rural	217 (5.4)	174 (6.7)	174 (7.2)	172 (7.2)	171 (7.3)	146 (7.4)	102 (6.2)
Respondent's alcoholic drinks	Everyday	58 (1.4)	38 (1.5)	32 (1.3)	27 (1.1)	27 (1.1)	24 (1.2)	24 (1.5)
	4 - 6 times a week	167 (4.2)	118 (4.5)	112 (4.6)	119 (5.0)	121 (5.1)	97 (4.9)	85 (5.2)
	2 - 3 times a week	571 (14.2)	401 (15.4)	389 (16.1)	398 (16.6)	398 (16.9)	333 (16.9)	294 (18.0)
	Once a week	822 (20.5)	555 (21.3)	514 (21.3)	504 (21.0)	496 (21.1)	415 (21.0)	335 (20.5)
	2- 3 times a month	632 (15.7)	432 (16.6)	397 (16.5)	402 (16.7)	390 (16.6)	322 (16.3)	259 (15.9)
	Once a month or less	1106 (27.5)	713 (27.4)	661 (27.4)	641 (26.7)	620 (26.4)	523 (26.5)	421 (25.8)
	Not in the last year	202 (5.0)	114 (4.4)	103 (4.3)	106 (4.4)	104 (4.4)	87 (4.4)	69 (4.2)
	Do not drink at all (reference)	460 (11.4)	232 (8.9)	201 (8.3)	204 (8.5)	196 (8.3)	171 (8.7)	147 (9.0)
Respondent's current health	Excellent (reference)	802 (20.0)	517 (19.9)	434 (18.0)	485 (20.2)	489 (20.8)	487 (24.7)	343 (21.0)
	Very Good	1602 (39.9)	1037 (39.8)	973 (40.4)	1001 (41.7)	932 (39.6)	843 (42.7)	705 (43.1)
	Good	1093 (27.2)	771 (29.6)	745 (30.9)	679 (28.3)	680 (28.9)	469 (23.8)	415 (25.4)
	Fair or poor	521 (13.0)	278 (10.7)	257 (10.7)	236 (9.8)	251 (10.7)	173 (8.8)	171 (10.5)
Respondent's health problem(s) in a year	Yes	618 (15.4)	431 (16.6)	406 (16.9)	382 (15.9)	412 (17.5)	357 (18.1)	310 (19.0)
	No (reference)	3400 (84.6)	2172 (83.4)	2003 (83.1)	2019 (84.1)	1940 (82.5)	1615 (81.9)	1324 (81.0)
Respondent's current job	Yes (reference)	2562 (63.8)	2480 (95.3)	2325 (96.5)	2288 (95.3)	2266 (96.3)	1905 (96.6)	1590 (97.3)
	No	1456 (36.2)	123 (4.7)	84 (3.5)	113 (4.7)	86 (3.7)	67 (3.4)	44 (2.7)
Deprivation	Quintile 1 (least deprived) (reference)	844 (21.0)	635 (24.4)	603 (25.0)	619 (25.8)	646 (27.5)	548 (27.8)	452 (27.7)
	Quintile 2	853 (21.2)	624 (24.0)	585 (24.3)	599 (24.9)	595 (25.3)	506 (25.7)	422 (25.8)
	Quintile 3	836 (20.8)	555 (21.3)	506 (21.0)	520 (21.7)	509 (21.6)	405 (20.5)	323 (19.8)
	Quintile 4	690 (17.2)	394 (15.1)	392 (16.3)	367 (15.3)	354 (15.1)	299 (15.2)	270 (16.5)
	Quintile 5 (most deprived)	795 (19.8)	395 (15.2)	323 (13.4)	296 (12.3)	248 (10.5)	214 (10.9)	167 (10.2)

Note: The term "reference" (indicated within the parentheses) represents the reference categories which will be utilised in the model categories.

Table 2.10: Summary of time-varying categorical variables for GUS males data

Variables	**Level/Measurements**	**SW1**	**SW4**	**SW6**	**SW7**	**SW8**	**SW9**	**SW10**
		Children (%)	Children (%)	Children (%)	Children (%)	Children (%)	Children (%)	Children (%)
Child's health in general	Very good (reference)	1484 (71.7)	931 (70.8)	929 (76.5)	939 (77.2)	965 (81.8)	748 (76.2)	598 (75.9)
	Good	452 (21.8)	304 (23.1)	238 (19.6)	232 (19.1)	178 (15.1)	196 (20.0)	157 (19.9)
	Bad & Very bad	135 (6.5)	80 (6.1)	47 (3.9)	45 (3.7)	36 (3.1)	38 (3.9)	33 (4.2)
Mother's marital status	Single, that is, never married	777 (37.5)	365 (27.8)	307 (25.3)	275 (22.6)	243 (20.6)	183 (18.6)	106 (13.5)
	Married and living with husband (reference)	1166 (56.3)	876 (66.6)	824 (67.9)	836 (68.8)	815 (69.1)	680 (69.2)	573 (72.7)
	Other	128 (6.2)	74 (5.6)	83 (6.8)	105 (8.6)	121 (10.3)	119 (12.1)	109 (13.8)
Urban-rural classification	Large urban (reference)	740 (35.7)	457 (34.8)	399 (32.9)	404 (33.2)	393 (33.3)	321 (32.7)	254 (32.2)
	Other urban	653 (31.5)	415 (31.6)	383 (31.5)	380 (31.3)	368 (31.2)	297 (30.2)	249 (31.6)
	Small, accessible towns	187 (9.0)	124 (9.4)	119 (9.8)	117 (9.6)	110 (9.3)	102 (10.4)	81 (10.3)
	Small, remote towns	60 (2.9)	40 (3.0)	37 (3.0)	46 (3.8)	49 (4.2)	34 (3.5)	30 (3.8)
	Accessible rural	305 (14.7)	179 (13.6)	176 (14.5)	169 (13.9)	160 (13.6)	140 (14.3)	121 (15.4)
	Remote rural	126 (6.1)	100 (7.6)	100 (8.2)	100 (8.2)	99 (8.4)	88 (9.0)	53 (6.7)
Respondent's alcoholic drinks	Everyday	28 (1.4)	16 (1.2)	16 (1.3)	13 (1.1)	13 (1.1)	11 (1.1)	13 (1.6)
	4 - 6 times a week	92 (4.4)	69 (5.2)	65 (5.4)	69 (5.7)	69 (5.9)	53 (5.4)	50 (6.3)
	2 - 3 times a week	285 (13.8)	212 (16.1)	194 (16.0)	205 (16.9)	200 (17.0)	165 (16.8)	145 (18.4)
	Once a week	440 (21.2)	278 (21.1)	267 (22.0)	265 (21.8)	254 (21.5)	218 (22.2)	165 (20.9)
	2- 3 times a month	339 (16.4)	227 (17.3)	204 (16.8)	211 (17.4)	204 (17.3)	165 (16.8)	140 (17.8)
	Once a month or less	568 (27.4)	355 (27.0)	333 (27.4)	316 (26.0)	314 (26.6)	262 (26.7)	194 (24.6)
	Not in the last year	98 (4.7)	52 (4.0)	48 (4.0)	50 (4.1)	49 (4.2)	37 (3.8)	30 (3.8)
	Do not drink at all (reference)	221 (10.7)	106 (8.1)	87 (7.2)	87 (7.2)	76 (6.4)	71 (7.2)	51 (6.5)
Respondent's current health	Excellent (reference)	385 (18.6)	262 (19.9)	220 (18.1)	238 (19.6)	243 (20.6)	229 (23.3)	173 (22.0)
	Very Good	820 (39.6)	519 (39.5)	473 (39.0)	489 (40.2)	474 (40.2)	438 (44.6)	331 (42.0)
	Good	586 (28.3)	398 (30.3)	379 (31.2)	362 (29.8)	330 (28.0)	227 (23.1)	207 (26.3)
	Fair or poor	280 (13.5)	136 (10.3)	142 (11.7)	127 (10.4)	132 (11.2)	88 (9.0)	77 (9.8)
Respondent's health problem(s) in a year	Yes	340 (16.4)	222 (16.9)	221 (18.2)	210 (17.3)	198 (16.8)	179 (18.2)	139 (17.6)
	No (reference)	1731 (83.6)	1093 (83.1)	993 (81.8)	1006 (82.7)	981 (83.2)	803 (81.8)	649 (82.4)
Respondent's current job	Yes (reference)	1313 (63.4)	1258 (95.7)	1172 (96.5)	1170 (96.2)	1140 (96.7)	952 (96.9)	767 (97.3)
	No	758 (36.6)	57 (4.3)	42 (3.5)	46 (3.8)	39 (3.3)	30 (3.1)	21 (2.7)
Deprivation	Quintile 1 (least deprived) (reference)	437 (21.1)	343 (26.1)	305 (25.1)	324 (26.6)	333 (28.2)	288 (29.3)	231 (29.3)
	Quintile 2	433 (20.9)	315 (24.0)	290 (23.9)	295 (24.3)	294 (24.9)	245 (24.9)	192 (24.4)
	Quintile 3	431 (20.8)	267 (20.3)	251 (20.7)	258 (21.2)	253 (21.5)	199 (20.3)	155 (19.7)
	Quintile 4	341 (16.5)	196 (14.9)	198 (16.3)	187 (15.4)	178 (15.1)	147 (15.0)	127 (16.1)
	Quintile 5 (most deprived)	429 (20.7)	194 (14.8)	170 (14.0)	152 (12.5)	121 (10.3)	103 (10.5)	83 (10.5)

Note: The term "reference" (indicated within the parentheses) represents the reference categories which will be utilised in the model categories.

Table 2.11: Summary of time-varying categorical variables for GUS females data

Variables	**Level/Measurements**	**SW1**	**SW4**	**SW6**	**SW7**	**SW8**	**SW9**	**SW10**
		Children (%)	Children (%)	Children (%)	Children (%)	Children (%)	Children (%)	Children (%)
Child's health in general	Very good (reference)	1530 (78.6)	977 (75.9)	936 (78.3)	967 (81.6)	969 (82.6)	767 (77.5)	618 (73.0)
	Good	333 (17.1)	252 (19.6)	223 (18.7)	185 (15.6)	175 (14.9)	191 (19.3)	183 (21.6)
	Bad & Very bad	84 (4.3)	59 (4.6)	36 (3.0)	33 (2.8)	29 (2.5)	32 (3.2)	45 (5.3)
Mother's marital status	Single, that is, never married	727 (37.3)	379 (29.4)	305 (25.5)	279 (23.5)	225 (19.2)	199 (20.1)	140 (16.5)
	Married and living with husband (reference)	1091 (56.0)	827 (64.2)	798 (66.8)	813 (68.6)	820 (69.9)	671 (67.8)	586 (69.3)
	Other	129 (6.6)	82 (6.4)	92 (7.7)	93 (7.8)	128 (10.9)	120 (12.1)	120 (14.2)
Urban-rural classification	Large urban (reference)	655 (33.6)	425 (33.0)	374 (31.3)	374 (31.6)	370 (31.5)	312 (31.5)	265 (31.3)
	Other urban	618 (31.7)	444 (34.5)	416 (34.8)	435 (36.7)	399 (34.0)	349 (35.3)	280 (33.1)
	Small, accessible towns	225 (11.6)	131 (10.2)	128 (10.7)	122 (10.3)	126 (10.7)	113 (11.4)	97 (11.5)
	Small, remote towns	75 (3.9)	43 (3.3)	38 (3.2)	38 (3.2)	40 (3.4)	23 (2.3)	30 (3.5)
	Accessible rural	283 (14.5)	171 (13.3)	165 (13.8)	144 (12.2)	166 (14.2)	135 (13.6)	125 (14.8)
	Remote rural	91 (4.7)	74 (5.7)	74 (6.2)	72 (6.1)	72 (6.1)	58 (5.9)	49 (5.8)
Respondent's alcoholic drinks	Everyday	30 (1.5)	22 (1.7)	16 (1.3)	14 (1.2)	14 (1.2)	13 (1.3)	11 (1.3)
	4 - 6 times a week	75 (3.9)	49 (3.8)	47 (3.9)	50 (4.2)	52 (4.4)	44 (4.4)	35 (4.1)
	2 - 3 times a week	286 (14.7)	189 (14.7)	195 (16.3)	193 (16.3)	198 (16.9)	168 (17.0)	149 (17.6)
	Once a week	382 (19.6)	277 (21.5)	247 (20.7)	239 (20.2)	242 (20.6)	197 (19.9)	170 (20.1)
	2- 3 times a month	293 (15.0)	205 (15.9)	193 (16.2)	191 (16.1)	186 (15.9)	157 (15.9)	119 (14.1)
	Once a month or less	538 (27.6)	358 (27.8)	328 (27.4)	325 (27.4)	306 (26.1)	261 (26.4)	227 (26.8)
	Not in the last year	104 (5.3)	62 (4.8)	55 (4.6)	56 (4.7)	55 (4.7)	50 (5.1)	39 (4.6)
	Do not drink at all (reference)	239 (12.3)	126 (9.8)	114 (9.5)	117 (9.9)	120 (10.2)	100 (10.1)	96 (11.3)
Respondent's current health	Excellent (reference)	417 (21.4)	255 (19.8)	214 (17.9)	247 (20.8)	246 (21.0)	258 (26.1)	170 (20.1)
	Very Good	782 (40.2)	518 (40.2)	500 (41.8)	512 (43.2)	458 (39.0)	405 (40.9)	374 (44.2)
	Good	507 (26.0)	373 (29.0)	366 (30.6)	317 (26.8)	350 (29.8)	242 (24.4)	208 (24.6)
	Fair or poor	241 (12.4)	142 (11.0)	115 (9.6)	109 (9.2)	119 (10.1)	85 (8.6)	94 (11.1)
Respondent's health problem(s) in a year	Yes	278 (14.3)	209 (16.2)	185 (15.5)	172 (14.5)	214 (18.2)	178 (18.0)	171 (20.2)
	No (reference)	1669 (85.7)	1079 (83.8)	1010 (84.5)	1013 (85.5)	959 (81.8)	812 (82.0)	675 (79.8)
Respondent's current job	Yes (reference)	1249 (64.1)	1222 (94.9)	1153 (96.5)	1118 (94.3)	1126 (96.0)	953 (96.3)	823 (97.3)
	No	698 (35.9)	66 (5.1)	42 (3.5)	67 (5.7)	47 (4.0)	37 (3.7)	23 (2.7)
Deprivation	Quintile 1 (least deprived) (reference)	407 (20.9)	292 (22.7)	298 (24.9)	295 (24.9)	313 (26.7)	260 (26.3)	221 (26.1)
	Quintile 2	420 (21.6)	309 (24.0)	295 (24.7)	304 (25.7)	301 (25.7)	261 (26.4)	230 (27.2)
	Quintile 3	405 (20.8)	288 (22.4)	255 (21.3)	262 (22.1)	256 (21.8)	206 (20.8)	168 (19.9)
	Quintile 4	349 (17.9)	198 (15.4)	194 (16.2)	180 (15.2)	176 (15.0)	152 (15.4)	143 (16.9)
	Quintile 5 (most deprived)	366 (18.8)	201 (15.6)	153 (12.8)	144 (12.2)	127 (10.8)	111 (11.2)	84 (9.9)

Note: The term "reference" (indicated within the parentheses) represents the reference categories which will be utilised in the model categories.

2.7 Chapter summary

The chapter offers a concise overview of child growth and development, emphasising the Growing up in Scotland study (GUS), which provides data for thesis analysis. The literature reviewed in the first section reveals that human growth has been studied for centuries, aiming to understand the progression from birth to adulthood. Childhood, recognised as a pivotal stage, comprises sub-periods determined by age, including prenatal, infancy, early childhood, middle childhood, and adolescence. Growth, whether in childhood or adolescence, during this period is crucial as it impacts further life stages. Therefore, it is important for both family and community units to be concerned their children. Physical growth, a vital facet of child development, reflects overall health and well-being, and is utilised to track or monitor child and adolescent growth and development. The subsequent section delves into child growth studies, crucial for understanding the intricacies of children's physical development. Such studies illuminate growth patterns, risk factors, and the formulation of growth standards and charts for health and public health purposes. Child growth studies primarily utilise two methodologies: cross-sectional and longitudinal. The latter can offer detailed insight into growth, encompassing dynamic or periodic aspects, while the former cannot.

The third section outlines the diverse measurements employed in modern child physical growth studies. These measurements can broadly be categorised into two types: raw and z-scale (standardised) growth measurements. The former measurements are traditional metrics that provide direct, unadjusted data on aspects like height, weight, and other physical attributes of children at specific ages. In contrast, the latter measurements, such as Z-scores or standard deviation scores (SDS), derive from raw measurements but are adjusted for age and sex. These scores enable comparison of a child's growth data against a reference population's average and standard deviation, assessing a child's growth against expected norms and helps identify deviations from typical growth patterns.

Next, child growth charts are discussed. These charts are vital tools for monitoring children's physical development, illustrating the distribution of growth measurements that vary with age. Constructed from child growth data, they typically display nine centile curves, although this can vary based on the reference population. Healthcare professionals use them to monitor and gauge typical growth. Among various methods, the LMS method is particularly notable for creating these charts. There are two primary types: the growth standard chart, which represents the ideal growth patterns of healthy children (e.g. those by WHO), and the growth reference chart, which presents growth patterns of a broader population, such as the UK-WHO charts from 2010. The latter provides insights into general child development.

Additionally, the risk factors associated with child growth and development are summarised in Section 2.5. There is evidence that children's experiences of growth and development can be positively or negatively influenced by several factors, whether direct or indirect. Experts have sought to summarise these factors for intervention and prevention purposes, especially by location or country. Such summaries are beneficial for policy formulation.

In the final section, the Growing up in Scotland (GUS) study is introduced. This study comprises three sub-cohorts: Birth Cohort 1, Birth Cohort 2, Child Cohort. This thesis primarily focuses on Birth Cohort 1. Data up to the year 2020 (sweeps 1 to 10) were considered. Emphasis was placed on raw weight and height data, which were then standardised using the UK-WHO growth reference. The framework by Sabates and Dex (2015) was employed to select variables associated with child physical growth measurements. During data exploration, growth patterns were observed to vary between males and females. Typically, males exhibited slightly higher averages in weight and height than females. Growth patterns tended to follow a non-linear trajectory. Notable variation in growth measurements was observed both among and within children over time. Observation times for children varied, particularly as they aged. The dataset provides evidence of serial correlation. Lastly, both time-varying and time-invariant variables were summarised.

Chapter 3

Statistical background

In the analysis of longitudinal child growth data (LCGD), it is essential to use statistical methodologies that can adeptly address inherent complexities, such as the correlation structures within the data and the observed diverse growth patterns. While the literature does not prescribe a single statistical method, multiple techniques are frequently amalgamated for a comprehensive analysis. This chapter elucidates the statistical methodologies, as described in existing literature, used for LCGD analysis. These methods will be further employed in the subsequent chapters. The initial section describes mixed-effects models, detailing their definition, general framework, parameter estimation procedures, and expanded models. Section 3.2 introduces the flexible mixed-effects models. In Section 3.3, correlation models are reviewed as another common model utilised for monitoring child growth via longitudinal child growth data. Section 3.4 outlines quantile regression (QR), introducing its definition, properties and two parameter estimation methods. Specifically, Bayesian quantile regression is dissected in Section 3.5. Section 3.6 reviews quantile regression models in longitudinal data. Sections 3.7 to 3.12 present methodologies for modelling non-linear dependencies in regression analysis. The final section provides a brief summary of the chapter.

3.1 Mixed-effects models

Various statistical approaches have been proposed in the literature to analyse LCGD. Among these, *conditional models* (conditional on the random effect), also known as mixed-effects models, are particularly prominent. These methods are known by various names depending on the area of application or statistical method, such as random-effects models (Laird & Ware, 1982), multilevel models (Goldstein, 1989), and hierarchical models (Bryk & Raudenbush, 1987). The core framework of these models involves incorporating random individual effects into the regression model to account for the dependency between observations from the same subject taken on different occasions. In essence, these random effects delineate individual variation and capture the correlation structure in repeated observations. Therefore, this section elucidates the principles of mixed-effects models, laying the groundwork for understanding all mixed-effects model variants.

3.1.1 Definition of Mixed-effects model

As previously mentioned, mixed-effects models are a type of model that incorporates random individual effects into the regression model. To introduce this kind of model, consider the simple model form: let y_{ij} denote the outcome measured on individual i at time t_{ij}, for $i = 1, \ldots, N$ individuals and $j = 1, \ldots, n_i$ time points,

$$y_{ij} = \underbrace{\beta_0 + \beta_1 t_{ij}}_{\text{Regression model}} + \underbrace{b_{0i} + b_{1i} t_{ij}}_{\text{Random individual effects}} + \underbrace{\epsilon_{ij}}_{\text{Random errors}}, \tag{3.1}$$

where β_0 is the overall intercept or initial level, β_1 is the overall slope (representing linear change across time), and b_{0i} and b_{1i} represent the random individual effects relating to intercept and slope, respectively. Thus, the model (3.1) comprises two main components plus the random error term (ϵ_{ij}): (1) the regression model and (2) the random individual effects model. Typically, the first component generates the mean model (or population-averaged model) and does not involve any individual, characterising it as a “fixed model”. Conversely, the second part, termed the “individual model”, illustrates how each individual deviates from the mean model.

Fixed effects

The term “fixed effects” pertains to any observed effects (e.g. covariates, factors) included in the fixed model. In essence, it delineates effects that influence only the mean values of the response across the entire population. Consequently, the interpretation of these effects is confined to the population level.

Random effects

The term “random effects” refers to any unobserved effect that account for variations in the data that cannot be explained by the fixed effects. These effects capture the random variability in the data attributable to individual differences or differences between levels of a grouping factor. This variation is inherently unpredictable because each individual or group may have unique and unobservable characteristics affecting the outcome.

3.1.2 Linear mixed-effects models

The mixed-effects models for LCGD commence with the simple linear mixed-effects models (LMMs), proposed by Laird and Ware (1982). These models tackle both between-individual and within-individual variability by formulating a two-level hierarchical model. The first-level model estimates each individual's trajectory over time, accounting for its variability. The second-level model estimates change across individuals, addressing the between-individual variability. Both levels are modelled simultaneously, allowing for the analysis of all individual data in a single analysis.

Let y_{ij} be the growth measurement corresponding to the jth ($j = 1, \ldots, n_i$) measurement of the ith ($i = 1, \ldots, N$) child, and let t_{ij} represent a continuous time variable (e.g. age). The model framework is

Level 1 Repeated observation models: $$y_{ij} = a_{0i} + a_{1i}t_{ij} + \epsilon_{ij}, \tag{3.2}$$

Level 2 Individual model: $$a_{0i} = \beta_{00} + u_{0i}, \tag{3.3}$$
$$a_{1i} = \beta_{10} + u_{1i}, \tag{3.4}$$

Mixed model: $$y_{ij} = \beta_{00} + \beta_{10}t_{ij} + u_{0i} + u_{1i}t_{ij} + \epsilon_{ij}, \tag{3.5}$$

Mean model: $$\mu_{ij} = \beta_{00} + \beta_{10}t_{ij}. \tag{3.6}$$

In model (3.2), a_{0i} and a_{1i} represent the random intercepts and slopes of individuals, respectively. These random coefficients describe the change in an individual i's trajectory over time within a linear model, analogous to being linear in parameters.

In model (3.3), β_{00} represents the mean intercept of a_{0i}, analogous to the Level-1 intercepts. This coefficient is often used to describe the mean of the entire population and is typically assumed to be a fixed effect in the model.

In model (3.4), β_{10} represents the mean slope of a_{1i}, analogous to the Level-1 slopes. This coefficient describes the average change in growth over time and is also assumed to be fixed effect.

Here, u_{0i} and u_{1i} are random individual effects for the intercept and slope, respectively, analogous to individual coefficients. These effects are employed to capture the differences when an individual's growth trajectory deviates from the mean model (3.6).

The mixed model (3.5) can be represented in matrix notation in the following single form:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, N. \tag{3.7}$$

Here, $\mathbf{y}_i$ is a known $n_i \times 1$ vector of observations for the ith subject at the jth measurement time ($j = 1, \dots, n_i$), given by

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix}.$$

$\mathbf{X}_i$ is a known $n_i \times 2$ design matrix for fixed effects, which contains a column of 1 and the observed values of the time variable (t) at the individual level,

$$\mathbf{X}_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}.$$

$\boldsymbol{\beta}$ is a 2×1 vector of unknown fixed effects coefficients,

$$\boldsymbol{\beta}_i = \begin{pmatrix} \beta_{00} \\ \beta_{10} \end{pmatrix}.$$

Additionally, $\mathbf{Z}_i$ is a known $n_i \times 2$ design matrix for random effects, equivalent to $\mathbf{X}_i$. In this case, $\mathbf{u}_i$ is a 2×1 vector of unknown random effect coefficients where

$$\mathbf{u}_i = \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right),$$

σ_0^2, σ_1^2, and σ_{01} are elements of the variance-covariance matrix $\mathbf{G}$, representing variances of random intercepts and random slopes, and their covariance, respectively. $\boldsymbol{\epsilon}_i$ is an $n_i \times 1$ vector of random errors where

$$\boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{n_i}\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{R}_i = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \right), \tag{3.8}$$

and σ^2 is the residual variance. In the above assumption, each random error is assumed to be independent, with a homogeneous residual variance for all repeated observations.

However, assumption (3.8) may be unrealistic for LCGD due to the correlation in repeated observations. As a result, $\mathbf{R}_i$ can be specified with a different structure to capture that characteristic (Littell et al., 2000). A popular option is a **first-order autoregressive structure**, or **AR(1)** (Jacobs & Lewis, 1978), where covariances between observations on the same child decay over time:

$$\mathbf{R}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho^{n_i-1} \\ \rho & 1 & \cdots & \rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \cdots & 1 \end{bmatrix},$$

where ρ represents the correlation coefficient in repeated observations. Furthermore, other structures can alternatively specify the R matrix; for example, a **compound symmetry (CS)** structure. The earliest explicit discussions of this structure were by Votaw (1948). The structure can be represented as follows:

$$\mathbf{R}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}.$$

The CS specifies that the covariances of each pair of observations are identical. However, CS may be unsuitable for LCGD because covariances are typically unequal as age increases. In contrast, there are two other structures that allow each covariance to be different, namely, **Toeplitz (TOEP)** (Toeplitz, 1911):

$$\mathbf{R}_i = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n_i} \\ \rho_1 & 1 & \cdots & \rho_{n_i-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n_i} & \rho_{n_i-1} & \cdots & 1 \end{bmatrix},$$

and an **unstructured (UN)** structure:

$$\mathbf{R}_i = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n_i} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_i1} & \sigma_{n_i2} & \cdots & \sigma_{n_i}^2 \end{bmatrix}.$$

In the TOEP structure, the covariance matrix exhibits patterns wherein the covariance between two points depends is determined solely by the lag between them, rather than their specific temporal or spatial positions. Conversely, the UN structure does not adhere any specific pattern or rule.

Note that the above specification considers the case with only one covariate (t) and two random individual random effects (u_{0i} and u_{1i}). Generally, the model can be expanded to include p covariates and q random effects. Thus, the matrix $\mathbf{X}$ and the vector $\mathbf{u}$ would take the general forms:

$$\mathbf{X}_i = \begin{pmatrix} 1 & t_{i1} & x_{2i1} & \cdots & x_{pi1} \\ 1 & t_{i2} & x_{2i2} & \cdots & x_{pi1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & x_{2in_i} & \cdots & x_{pin_i} \end{pmatrix},$$

and

$$\mathbf{u}_i = \begin{pmatrix} u_{0i} \\ u_{1i} \\ \vdots \\ u_{qi} \end{pmatrix} \sim \mathcal{N}_q \left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \cdots & \sigma_{0q} \\ \sigma_{01} & \sigma_1^2 & \cdots & \sigma_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{0q} & \sigma_{1q} & \cdots & \sigma_q^2 \end{bmatrix} \right).$$

For all individuals, the mixed-effects model (3.7) can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{3.9}$$

where

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_N' \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_N \end{pmatrix}, \text{ and } \mathbf{u} = \begin{pmatrix} \mathbf{u}_1' \\ \mathbf{u}_2' \\ \vdots \\ \mathbf{u}_N' \end{pmatrix}.$$

The variance-covariance of $\mathbf{y}$, $\text{Var}(\mathbf{y}) = \mathbf{V}$, can be defined by

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}.$$

Thus, the model (3.9) can be written with normally distributed random variables in marginal form as

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}). \tag{3.10}$$

3.1.3 Maximum likelihood (ML) estimation

The marginal log-likelihood of $\mathbf{y}$ under the model (3.10) (Demidenko, 2013; Pinheiro & Bates, 2000) is given by

$$l(\boldsymbol{\beta}, \mathbf{V}) = -\frac{1}{2}\Big\{\log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\Big\} - \frac{N}{2}\log(2\pi). \tag{3.11}$$

For any fixed $\mathbf{V}$, the log-likelihood (3.11) is maximised over $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

For the variance-covariance ($\mathbf{V}$), $\hat{\boldsymbol{\beta}}$ can be substituted back into (3.11), thereby obtaining the profile log-likelihood:

$$l_p(\mathbf{V}) = -\frac{1}{2}\Big\{\log|\mathbf{V}| + \mathbf{y}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}\Big\}. \tag{3.12}$$

Subsequently, the function (3.12) can be maximised for the parameters in $\mathbf{V}$.

3.1.4 Restricted maximum likelihood (REML) estimation

As is well known (Demidenko, 2013; Laird & Ware, 1982), the ML estimator has some shortcomings. Notably, it does not adjust the estimation of variance components to account for the degrees of freedom used in estimating the fixed effects, which typically results in biased estimators and the overestimation of the variance-covariance parameters. Consequently, to adjust this respect, restricted maximum likelihood estimators (REML) (Corbeil & Searle, 1976; Patterson & Thompson, 1971) are often preferred over ML estimators. The expression for the REML is given as

$$l_R(\mathbf{V}) = -\frac{1}{2}\Big\{\log|\mathbf{V}| + \log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}\Big\}.$$

As the expression above shows, the part related to the estimation of fixed effects is removed from the likelihood function compared to function (3.12). This approach ensures that the degrees of freedom used for estimating the fixed effects are not consumed in the estimation of variance components. Consequently, the modified likelihood function, which involves only the random effect part, is maximised to estimate the variance components of the random effects.

3.1.5 Prediction of the random effects

As previously outlined, the random effects are utilised to account for variation at the individual or group level. In theory, these effects are assumed to originate from a normal distribution with a mean of zero. Unlike the fixed effects, they are treated as random variables, which cannot be directly estimated but are predicted. The prediction of $\mathbf{u}$ is straightforward, using the best linear unbiased prediction (BLUP) (Robinson, 1991),

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

The solution above is based on finding $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ to minimise the prediction error, which results in $\hat{\mathbf{u}}$ being the best prediction. It is also a linear prediction because it is predicted as a linear combination of observed data. Furthermore, given that its expectation, $E(\hat{\mathbf{u}}) = E(\mathbf{u}) = 0$, it is unbiased.

3.1.6 Extended models

In simple LMMs, the relationship between child growth measurements (e.g. weight, height, WAZ, HAZ) and time (or age) is typically assumed to follow a linear trend. However, in reality, growth does not usually follow such a relationship (Anderson, Hafen, et al., 2019; Laird & Ware, 1982). The model outlined above cannot capture other growth patterns, such as non-linear ones. Although LMMs allow for the modelling of non-linear relationship by using higher-order polynomials (e.g. a quadratic trend over time), another issue arises. Specifically, using parametric functions to fit non-linear data tends to yield models with excessive numbers of parameters, complicating the interpretation of those parameters. To overcome this limitation, nonlinear mixed-effects models (NMMs) offer an alternative by describing growth mechanisms with few parameters (Cole et al., 2010; Lawton et al., 1972; Stützle et al., 1980). Despite the challenges associated with interpreting the parameters of higher-order polynomials, such modelling can yield a comprehensive depiction of growth trajectory trends, thus providing insights into how growth changes over time. Moreover, various mixed-effects model variants have been proposed as alternatives in this respect. For instance, piecewise mixed-effects models (PMMs) have been proposed to address growth patterns comprising at least two critical periods (Fitzmaurice et al., 2011; Muggeo et al., 2014; Naumova et al., 2001). Each of these models possesses its own advantages; however, they also exhibit limitations when it comes to accommodating a range of growth patterns. In other words, these models lack the requisite flexibility to accurately capture the data process. For instance, PMMs necessitate numerous parameters to manage the smoothness of growth curves across multiple critical periods.

3.2 Flexible mixed-effects models

Another method of modelling non-linear trajectories is via a “nonparametric” approach, allowing for the modelling of the nonlinear relationship between the response and the time variable using a nonparametric function. Typically, spline techniques (see Section 3.5) are used to construct these functions. Theoretically, such splines can be reformulated as a model comprising two components: parametric and nonparametric. Consequently, the term *semiparametric model* is often used to describe this type of model (Ruppert et al., 2003). Regarding model fitting, upon choosing a spline, the semiparametric model can be fitted by transforming the spline model into a mixed-effects model framework (splines as mixed-effects models) and then employing the penalised least squares method outlined in Section 3.6 to estimate the model's parameters.

Several semiparametric approaches, known as *semiparametric mixed-effects models*, have been widely proposed for modelling longitudinal data across various applications (Aniley et al., 2019; Durbán et al., 2005; Maringwa et al., 2008; Ruppert et al., 2003; Szczesniak et al., 2016; Thilakarathne et al., 2011; Zeger & Diggle, 1994; D. Zhang et al., 1998).

3.2.1 Model specification

The simplest form of semiparametric mixed-effects model based on penalised splines can be expressed by:

$$y_{ij} = f(t_{ij}) + u_{0i} + u_{1i}t_{ij} + \epsilon_{ij}, \quad i = 1, 2, \ldots, N, j = 1, 2, \ldots, n_i \tag{3.13}$$

where

$$f(t_{ij}) = \beta_{00} + \beta_{10}t_{ij} + \sum_{k=1}^{\kappa} \nu_{k0}(t_{ij} - \delta_k)_+,$$

Here, $f(t_{ij})$ is a smooth function representing the population mean curve over time. In this case, this smooth function is defined as truncated linear splines. However, another basis such as truncated polynomial splines or B-splines can be used (Durbán et al., 2005).

The model (3.13) can be written as a mixed-effects model framework by:

$$y_{ij} = \underbrace{\beta_{00} + \beta_{10}t_{ij}}_{\mathbf{X}\boldsymbol{\beta}} + \underbrace{\sum_{k=1}^{\kappa} \nu_k(t_{ij} - \delta_k)_+ + u_{0i} + u_{1i}t_{ij}}_{\mathbf{Z}\mathbf{u}^*} + \underbrace{\epsilon_{ij}}_{\boldsymbol{\epsilon}}, \tag{3.14}$$

and the model (3.14) can be expressed in the matrix notation form as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}^* + \boldsymbol{\epsilon}, \tag{3.15}$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)'$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)'$, $\boldsymbol{\beta} = (\beta_{00}, \beta_{10})'$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N)'$,

$$\mathbf{X}_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} \mathbf{B}_1 & \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{B}_2 & \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_N & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_N \end{pmatrix},$$

$$\mathbf{B}_i = \begin{pmatrix} (t_{i1} - \delta_1)_+ & \cdots & (t_{i1} - \delta_\kappa)_+ \\ (t_{i2} - \delta_1)_+ & \cdots & (t_{i1} - \delta_\kappa)_+ \\ \vdots & \ddots & \vdots \\ (t_{in_i} - \delta_1)_+ & \cdots & (t_{in_i} - \delta_\kappa)_+ \end{pmatrix},$$

$$\mathbf{u}^* = (\nu_1, \nu_2, \ldots, \nu_k, u_{01}, u_{02}, \ldots, u_{0N}, u_{1N})'.$$

The model (3.15) has its essential assumptions as follows:

$$\mathbf{u} = \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim \mathsf{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right), \nu_k \sim \mathsf{N}(0, \sigma_\nu^2), \text{ and } \epsilon_{ij} \sim \mathsf{N}(0, \sigma_\epsilon^2),$$

and

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{u}^*) = \begin{pmatrix} \sigma_\nu^2 \mathbf{I} & 0 \\ 0 & \underset{1 \le i \le N}{\text{block-diagonal}}\, \mathbf{G} \end{pmatrix}. \tag{3.16}$$

Here, ν_k are treated as random effects, analogous to u_{0i} and u_{1i}. These random effects are presumed to follow a normal distribution with zero mean and finite variance ($\sigma_\nu^2 > 0$). Adhering to this assumption serves as a form of regularisation, averting the overfitting of the population mean curve. It imposes a penalty on the coefficients, encouraging shrinkage towards zero and resulting in a more stable and interpretable model. However, this condition can apply different penalties, for example, $\nu_k \sim \mathsf{Laplace}(0, \sigma_\nu^2)$ (Wand, 2003).

3.2.2 Estimation

Assuming $\sigma_0^2, \sigma_1^2, \sigma_{01}, \sigma_\nu^2$ and σ_ϵ^2 are known, the estimates of $(\boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{u}_i)$ are obtained by minimising the penalised least squares:

$$\sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \left\{ y_{ij} - f(t_{ij}) - h_i(t_{ij}) \right\}^2 + \sigma_\epsilon^2 \boldsymbol{u}_i' \mathbf{G}^{-1} \boldsymbol{u}_i \right] + \lambda \boldsymbol{\nu}' \boldsymbol{\nu}, \tag{3.17}$$

where $\lambda = \sigma_\epsilon^2 / \sigma_\nu^2$, which represents the smoothing parameter. In expression (3.17), a term multiplied by the smoothing parameter is a penalty term that yields smoother fitted curves. Given all the above assumptions, the penalised spline smoother corresponds to the optimal predictor in a mixed-effects models framework (Durbán et al., 2005; Heckman et al., 2013; Wu & Zhang, 2006). Therefore, all parameters in the model can be estimated via the restricted maximum likelihood method of the mixed-effect model framework,

$$l_R(\mathbf{V}) = -\frac{1}{2}\Big\{\log|\mathbf{V}| + \log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}\Big\}, \tag{3.18}$$

where $\mathbf{V} = \mathbf{Z\Sigma Z}' + \mathbf{R}$, $\mathbf{R} = \sigma_\epsilon^2\mathbf{I}$ and $\mathbf{\Sigma}$ is defined in (3.16). The maximisation of (3.18) yields,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{3.19}$$

and

$$\hat{\mathbf{u}^*} = \mathbf{\Sigma Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{3.20}$$

Existing mixed-effects model software, i.e. the function gamm in the mgcv package (Wood, 2017a) of the R environment, can be used to estimate (3.19) and (3.20). In child growth modelling, Durbán et al. (2005) leveraged these advantages and used the equivalence between a penalised spline smoother and the optimal predictor in a mixed-effects model to form the semiparametric mixed-effects models.

3.3 Correlation models

Another common approach for modelling longitudinal child growth data is through *Correlation models*, particularly in the area of child health monitoring. The concept behind this approach is to identify metrics that capture the relationship between a child's growth measurements at multiple time points. These metrics are then utilised to make inferences and predict the child's growth measurements in the future. This can be exemplified by considering the correlation between growth measurements (e.g. height or weight) of a child at ages t_1 and t_2. With this information, we can infer the growth measurement at age t_2 for another child, given his or her growth measurement at t_1 years.

Cole (1995) applied this concept to propose an analysis of correlation models for assessing weight gain in British infants. Weight gain was defined as a standard deviation (Z-score) of a child's weight compared with the average weight of the reference population of children at the same two ages. Hence, it can be expressed mathematically as follows: Let z_1 and z_2 be weight gain measured at t_1 and t_2, respectively, and ρ_{12} represents the correlation between these two time points. The conditional weight gain can be expressed as

$$z_{2|1} = \frac{z_2 - \rho_{12}z_1}{\sqrt{1-\rho_{12}^2}} \tag{3.21}$$

The idea behind the conditional weight gain mentioned above is derived from simple linear regression, expressed as $E(z_2|z_1) = \rho_{12}z_1$ with standard deviation $\sqrt{1-\rho_{12}^2}$. Consequently, this conditional weight gain dependents on the correlation between them, ρ_{12}, rather than on the time points t_1 and t_2. This leads to the approach being a regression to the mean. By accurately estimating the correlation term for all pairs of possible time points, we can predict other weight gains. Therefore, the main point of this approach lies in calculating the correlations of all pairs. In practice, these calculations are computed based on a series of age groups, which are constructed by discretising continuous time. The estimation method depends on the assumption of the correlation process. For example, Cole (1995) utilised Fisher's transformation, $z = \frac{1}{2}\log\big((1+\rho_{12})/(1-\rho_{12})\big)$, to address this aspect for (3.21), so that $\rho_{12} = (\exp(2z)-1)/(\exp(2z)+1)$.

Since the proposal of Cole (1995), there have been other correlation models developed in the area of modelling child growth data, particularly focusing on proposing correlation functions and estimation methods. Argyle et al. (2008) proposed a particular two-parameter Markov form to represent the correlation function in infancy, with its parameters estimated via likelihood methods. Feng et al. (2020) employed several correlation functions, such as the exponential function, the exponential function with a nugget effect term, Markovian (following the approach of Argyle et al. (2008)), Markovian with nugget effect, and two nonparametric correlation functions (with functional data analysis), to monitor fetal growth. Moreover, Anderson, Xiao, et al. (2019) developed a child growth correlation matrix by pooling data from multiple studies with varying age ranges using a two-stage approach. This approach involves constructing a raw correlation (incomplete) matrix derived by univariate meta analyses in the first stage, which was then smoothed in the second stage to yield a complete and valid correlation matrix.

Although correlation models offer advantages in monitoring child growth, they have several limitations. Firstly, they require collecting data at consistent ages or at least at ages close to each other for all children (W. Johnson, 2015). Hence, they may not be suitable for longitudinal child data with timing differences between subjects, such as the GUS data. Nonetheless, this issue can be addressed by discretising continuous age differences into a series of age groups, but the loss of granularity and information may be a concern. This loss of precision can impact the ability of correlation models to capture subtle changes in growth trajectories and may result in less accurate predictions or interpretations of child growth patterns. Secondly, correlation models usually consider conditioning only on previous outcomes defined for just two ages (Van Buuren, 2023), whereas more than two ages should be considered in order to assess growth curves comprehensively. While multiple calculations can be utilised in this case, multiple testing issues must be noted. To address these issues, Van Buuren (2023) proposed a method to calculate conditional gain for multiple time points. However, this approach still focuses solely on previous outcomes and does not consider any predictors in the model, such as parental anthropometry, genetic predisposition, environmental factors. This narrow focus may lead to incomplete or biased assessments of child growth and development. Thirdly, correlation models do not take into account other characteristics of longitudinal child growth data, such as the heterogeneity of children. As a result, they may overlook important factors that contribute to the diversity of growth patterns observed in children. In this thesis, therefore, this kind of model is not considered as an approach.

3.4 Quantile regression

In this section, quantile regression (QR) is introduced. Similar to mean regression, QR can elucidate the relationship between the response variable and its covariates. The distinguishing feature of QR, however, is its focus on the impact of covariates across the entire conditional distribution of the response. This allows for an examination of changes in the distribution's location, scale, and shape, conditional on the covariates. For illustration, consider a child growth study where being overweight—a health issue influenced by numerous factors or covariates—is of interest. To analyse the determinants specifically affecting the upper tail of the weight distribution, QR is especially pertinent. Traditional methods, such as the ordinary least squares (OLS) regression, primarily focus on effects that influence the mean of the response distribution. Relying exclusively on the mean might not provide a comprehensive understanding of the distribution. Some factors may impact the mean but not other measures, such as the upper or lower quantiles of the response. QR provides a solution to this limitation. Within the literature, QR models are principally categorised based on parameter estimation techniques: *Distribution-free methods* and *Likelihood-based methods*.

3.4.1 A definition of quantiles

The term "quantiles" denotes specific location points within a dataset or distribution that segment the data into intervals of equal probability. In other words, it determines the number of observations in a distribution that are either above or below a specified limit. Figure 3.1 provides an example of quantiles for the raw weight in males within the GUS data.

Percentiles and quantiles

Percentiles are another version of quantiles, but they are represented in a different index. Percentiles are indexed by sample percentage rather than by sample fractions (the ratio of sample size to population size, equivalent to probability). For example (refer to Figure 3.1)),

- The 25th percentile is known as the 0.25th quantile or the *lower quartile*.
- The 50th percentile is known as the 0.50th quantile or the *median*.
- The 75th percentile is known as the 0.75th quantile or the *upper quartile*.

3.4.2 Quantile of random variable

Let $F_Y(y)$ be the cumulative distribution function of a continuous random variable Y, defined as:

$$F_Y(y) = \Pr(Y \leq y).$$

Then, for $\tau \in (0, 1)$, the τth quantile of the random variable Y is defined as:

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

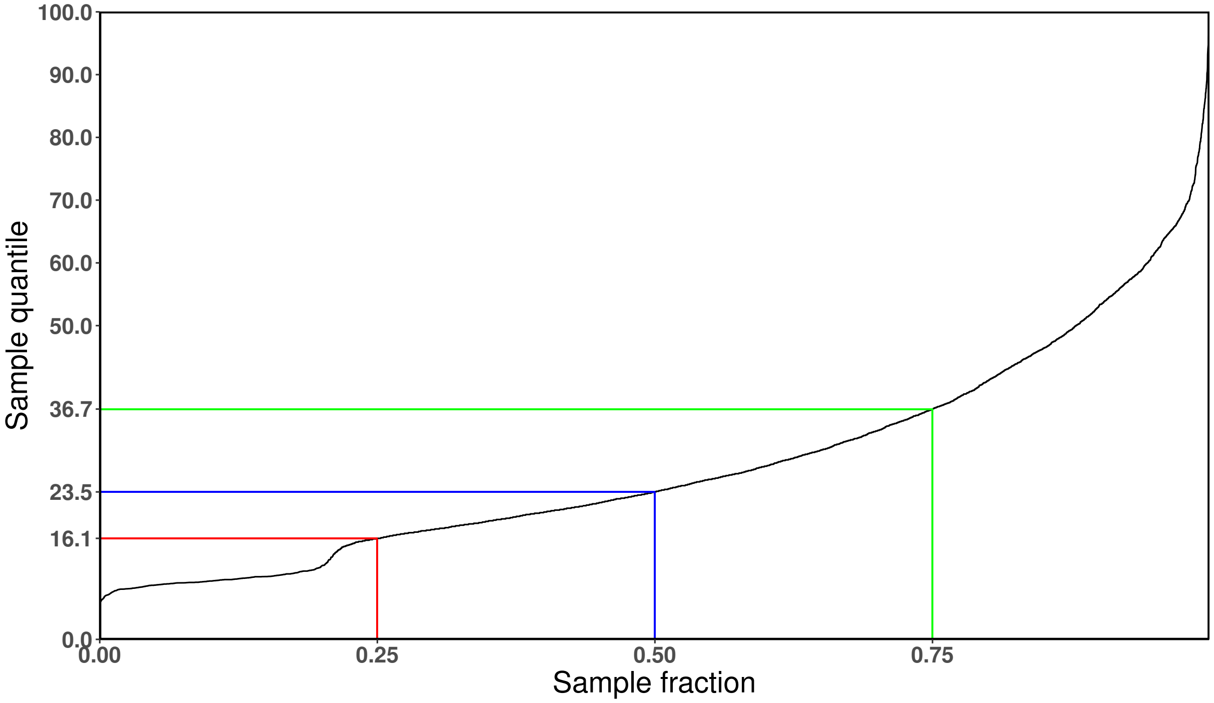


Figure 3.1: Quantile plots for the raw weight in males within the GUS data

3.4.3 Linear quantile regression

The basic concept of quantiles, as previously outlined, can be applied to construct a regression form for the specific τth quantile of Y conditional on covariates X. Given that Y is a continuous response variable and $\boldsymbol{x}$ is a $p \times 1$ vector of known covariates, the distribution of Y conditional on $\boldsymbol{x}$ can be denoted as $F_Y(y|\boldsymbol{x}) = \Pr(Y \leq y|\boldsymbol{x})$. Figure 3.2 illustrates conditional quantiles in a continuous response (e.g. raw weight) over a covariate (e.g. Age in years). Consequently, the τth quantile of Y conditional on $\boldsymbol{x}$, for $\tau \in (0, 1)$, is defined as:

$$Q_\tau(Y|\boldsymbol{x}) = \inf\{y : F_Y(y|\boldsymbol{x}) \geq \tau\}. \tag{3.22}$$

From this, equation (3.22) can be viewed as representing the linear form of the quantile regression model for a sample of n independent observations $\{y_i, \boldsymbol{x}_i\}$ from $i = 1$ to n, expressed as:

$$Q_\tau(Y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau, \tag{3.23}$$

or equivalently,

$$Q_\tau(Y_i|\boldsymbol{x}_i) = \beta_{0,\tau} + \beta_{1,\tau}x_1 + \cdots + \beta_{p,\tau}x_p.$$

Alternatively, the model in (3.23) can be conveyed through the linear model as

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_\tau + \epsilon_i, \tag{3.24}$$

where ϵ_i is a random error term based on the assumption $Q_\tau(\epsilon_i|\mathbf{x}_i) = 0$. Here, $\boldsymbol{\beta}_\tau \in \mathbb{R}^p$ is the vector of unknown regression coefficients, which can be interpreted as the marginal change in the τth quantile resulting from a marginal change in $\boldsymbol{x}$. Different settings of τ may yield coefficients that vary both in magnitude and sign. Each quantile must be increasing in τ and not cross each other.

3.4.4 Quantile function properties

Monotonicity

In quantile terms, if $\tau_1 < \tau_2$, then $Q_{\tau_1}(Y) \leq Q_{\tau_2}(Y)$. This implies that the quantile $Q_\tau(Y)$ is monotone in τ. In the context of linear quantile regression, $Q_\tau(Y|\boldsymbol{x})$ should increase with τ for any given value of $\boldsymbol{x}$, consistent with the properties of quantiles.

Equivariance

The equivariance is a measure of how alterations in the response variable, such as scaling or reparameterisation, influence the regression estimates in quantile regression. Such understanding aids in the interpretation of statistical results. Let $\hat{\beta}_\tau(y, X)$ represent the estimator for the τth regression quantile based on observations (y, X). Suppose A is any $p \times p$ non-singular matrix, γ belongs in $\mathbb{R}^p$, and $a > 0$. Then, for any τ in the interval $[0, 1]$:

1. Scale equivariance: $\hat{\beta}_\tau(ay, X) = a\hat{\beta}_\tau(y, X)$ and $\hat{\beta}_\tau(-ay, X) = -a\hat{\beta}_{1-\tau}(y, X)$

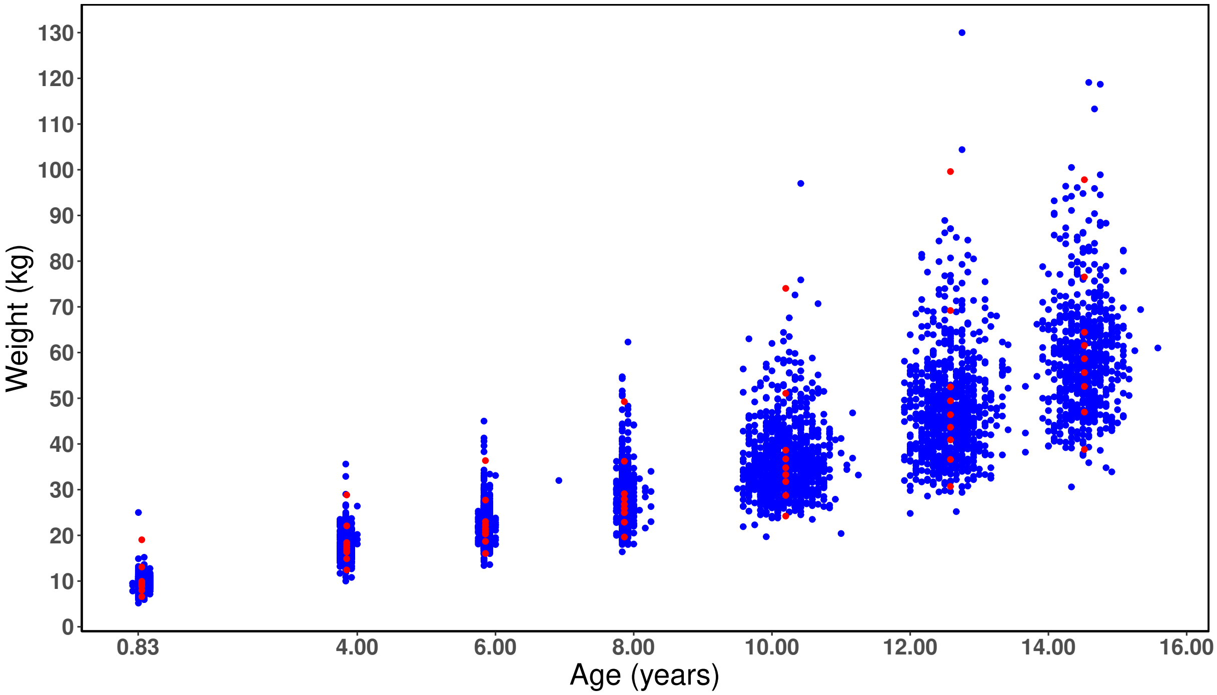


Figure 3.2: Conditional quantiles in raw weight over age (years). The red dots represent quantiles at τ values of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

2. Regression shift: $\hat{\beta}_\tau(y + X\gamma, X) = \hat{\beta}_\tau(y, X) + \gamma$

3. Reparameterisation of design: $\hat{\beta}_\tau(y + XA) = A^{-1}\hat{\beta}_\tau(y, X)$.

Equivariance to monotone transformation

Let $h(\cdot)$ be an increasing function on $\mathbb{R}$. For any variable Y,

$$Q_\tau(h(Y|\boldsymbol{x})) = h(Q_\tau(Y|\boldsymbol{x})).$$

As an illustration, consider $\log(Y) = \boldsymbol{x}'\boldsymbol{\beta}$. This implies that $Q_\tau(\log(Y)|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$. If our interest lies in $Q_\tau(Y|\boldsymbol{x})$, utilising this property allows us to deduce that

$$Q_\tau(\log(Y|\boldsymbol{x})) = \log(Q_\tau(Y|\boldsymbol{x})) = \boldsymbol{x}'\boldsymbol{\beta}.$$

Consequently, $Q_\tau(Y|\boldsymbol{x}) = \exp(\boldsymbol{x}'\boldsymbol{\beta})$.

Interpolation

Quantile regression is characterised by its ability to fit a model such that, for a given quantile τ, exactly τ percent of the data points will lie on or below the model's prediction for that quantile. To illustrate that with a simple example, consider the median regression, $\tau = 0.50$. The regression line should pass through the median of the data points, meaning that approximately 50% of the data points will be above and 50% below the regression line. Consequently, the residuals for the data points that the regression line fits exactly will be zero, while the remaining residuals will be evenly split between 50% positive and 50% negative. In the general case, this results in $n(\tau)$ positive residuals and $n(1-\tau)$ negative residuals, where n is the total number of observations (Koenker, 2005).

3.4.5 Distribution-free method

In classical quantile regression, to determine the coefficient estimates for the model, the following optimisation problem must be solved:

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}), \tag{3.25}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the quantile loss function and $I(\cdot)$ is the indicator function (see Figure 3.3). Typically, the solution of (3.25) is obtained by using linear programming algorithms (Koenker, 2005).

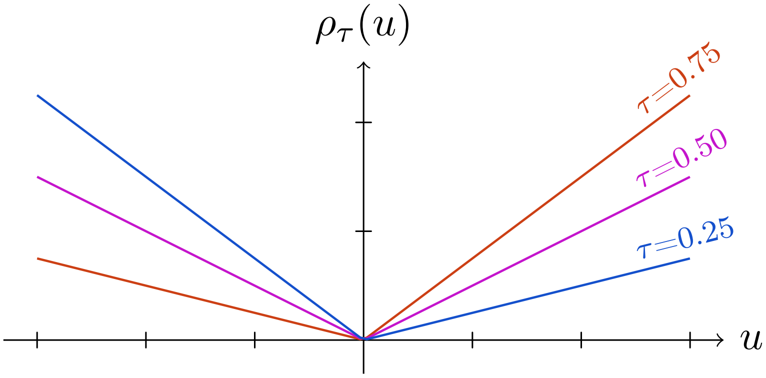


Figure 3.3: Quantile loss function, $\rho_\tau(u)$

Consider the linear quantile regression model given by (3.24),

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau + \epsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau + (u_i - v_i), \tag{3.26}$$

where $u_i = \epsilon_i I(\epsilon_i > 0)$ and $v_i = |\epsilon_i| I(\epsilon_i < 0), i = 1, \ldots, n$. The variables, u_i and v_i, are introduced to represent the positive and negative components of the residual vector. As such, equation (3.26) can be reformulated using these variables within a linear programming framework as:

$$\min_{(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{v}) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}} \left\{ \tau 1_n' \boldsymbol{u} + (1-\tau) 1_n' \boldsymbol{v} | \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{u} - \boldsymbol{v} = \boldsymbol{y} \right\}. \tag{3.27}$$

The minimisation problem in (3.27) can be addressed using the simplex algorithm in linear programming method (refer Koenker and D'Orey (1987) for additional details). This estimation approach does not depend on any distribution assumptions regarding the response. Consequently, this method is based on a *distribution-free* approach and is also referred to as *classical quantile regression.*

At present, implementation this method is straightforward using the quantreg package in R (Koenker, 2019). However, this method can be computationally challenging when the sample size is large. To mitigate this problem, two alternative approaches have been proposed: the Frish-Newton interior-point method (Portnoy & Koenker, 1997) and the sparse regression quantile fitting (Koenker & Ng, 2003). The former addresses the challenges in the simplex algorithm for larger sample sizes by traversing the interior of the feasible region. In contrast, the latter utilises sparse matrices to enable efficient computation, especially when the covariates include many factors.

Extended models

Classical QR models can be extended to accommodate various statistical frameworks. Examples include nonlinear quantile regression (Koenker & Park, 1996), nonparametric quantile regression such as local polynomial quantile regression (Chaudhuri, 1991), quantile smoothing splines (Koenker et al., 1994), and quantile regression splines (He & Ng, 1999; He & Shi, 1994; Hendricks & Koenker, 1992). There are also studies on penalised quantile regression splines, as highlighted by Muggeo et al. (2013), Ng and Maechler (2007), and Pratesi et al. (2009).

3.4.6 Likelihood-based method

Koenker and Machado (1999a) discovered that maximising the likelihood function of the residual error, ϵ, (assumed to follow an asymmetric Laplace (AL) distribution) yields a result equivalent to (3.25). Consequently, the parameter estimation for the quantile regression model can alternatively be determined using the method of maximum-likelihood estimation.

Asymmetric Laplace distribution

The AL distribution is a continuous probability distribution in which the random variable, Y, has a density characterised by three specific parameters: the location parameter $\mu \in \mathbb{R}$, the scale parameter $\sigma > 0$, and the skew parameter $\tau \in (0, 1)$ (Yu & Zhang, 2005). The density function is represented as:

$$f(y;\tau,\mu,\sigma)=\frac{\tau(1-\tau)}{\sigma}\text{exp}\Bigg\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\Bigg\},\quad -\infty<y<\infty, \tag{3.28}$$

where $\rho_\tau(u)$ is the quantile loss function. Figure 3.4 displays the density function of this distribution. The distribution is denoted briefly as $\mathcal{AL}(\mu,\sigma,\tau)$. Its distribution function (see Figure 3.5) and quantile function (see Figure 3.6) are respectively given by:

$$F(y;\tau,\mu,\sigma)=\begin{cases}\tau\text{exp}\Big\{\frac{1-\tau}{\sigma}(y-\mu)\Big\} & ,y\leq\mu\\ 1-(1-\tau)\text{exp}\Big\{-\frac{\tau}{\sigma}(y-\mu)\Big\} & ,y>\mu\end{cases}$$

and

$$F^{-1}(p;\tau,\mu,\sigma)=\begin{cases}\mu+\frac{\sigma}{1-\tau}\text{log}\Big(\frac{p}{\tau}\Big) & ,0\leq p<\tau\\ \mu-\frac{\sigma}{\tau}\text{log}\Big(\frac{1-p}{1-\tau}\Big) & ,\tau<p\leq 1.\end{cases}$$

For the aforementioned quantile function, the pth-quantile of the random variable y is equivalent to the location parameter μ when $p=\tau$, that is,

$$F^{-1}(p;\tau,\mu,\sigma)|_{p=\tau}=\mu.$$

To estimate μ using the method of maximum-likelihood estimation, let us consider an illustrative case where $\tau=1/2$. With this assumption, the density function can be reformulated as:

$$f(y;1/2,\mu,\sigma)=\frac{1}{4\sigma}\text{exp}\Bigg\{-\frac{|y-\mu|}{2\sigma}\Bigg\}.$$

Assuming σ is known while μ remains unknown, the log-likelihood function becomes:

$$\mathcal{L}(\mu,\sigma)\propto-\frac{1}{2\sigma}\sum_{i=1}^{n}|y_i-\mu|.$$

Subsequently, taking the partial derivative of $\mathcal{L}(\mu,\sigma)$ with respect to μ yields:

$$\frac{\partial\mathcal{L}(\mu,\sigma)}{\partial\mu}=\frac{1}{2\sigma}\sum_{i=1}^{n}\text{sgn}(y_i-\mu)$$

To maximise the likelihood function, μ needs to be considered such that:

$$\frac{1}{2\sigma}\sum_{i=1}^{n}\text{sgn}(y_i-\mu)=0 \tag{3.29}$$

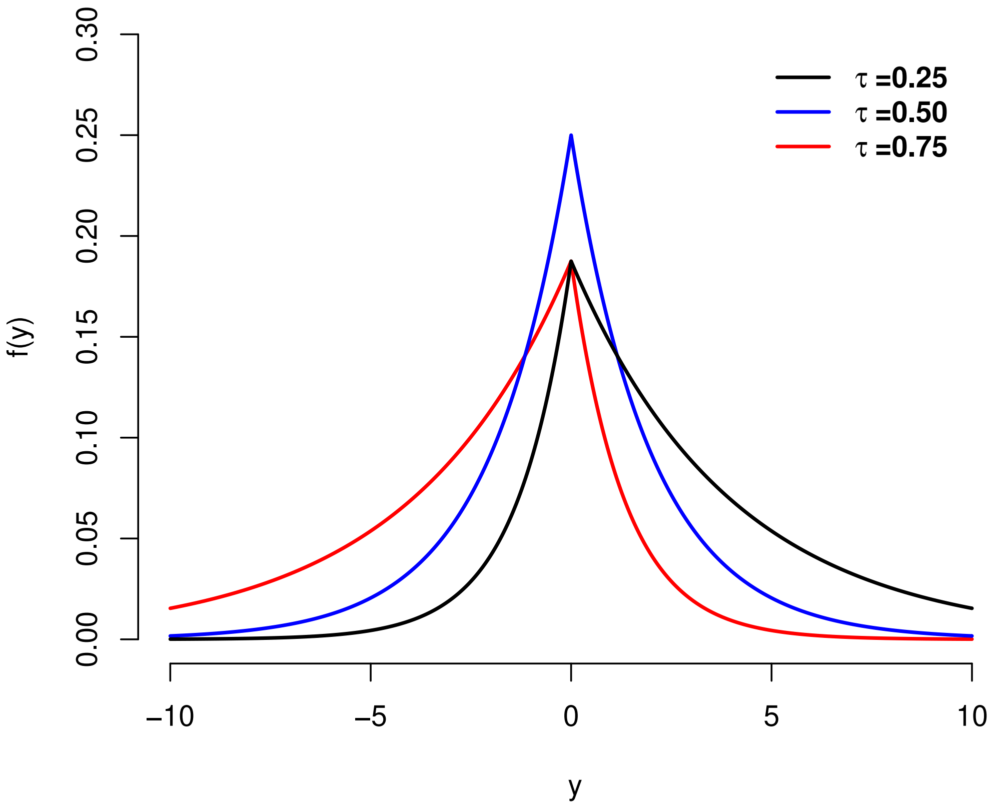


Figure 3.4: AL density function with $\mu = 0, \sigma = 1$, and $\tau \in (0.25, 0.50, 0.75)$.

Considering equation (3.29), there are two cases based on the number n, either odd or even. These cases determine the maximum likelihood estimator of μ as follows:

1. For an odd number: One y_i from $(y_1, \ldots, y_n)$ can be selected to be μ in order to satisfy equation (3.29). A way to address this is by defining this value as:

$$\hat{\mu} = \text{median}(y_1, \ldots, y_n).$$

   This results in $(n-1)/2$ instances where $(y_i - \mu) > 0$ and for the remaining $(n-1)/2$ instances where $(y_i - \mu) < 0$. It is evident that $\hat{\mu}$ satisfies (3.29). Hence, it can be concluded that $\hat{\mu}$ is the maximum likelihood estimator for μ.

2. For an even number: It is not possible to select a single y_i to satisfy (3.29). However, it can be minimised by arranging the observations as $y_1 \leq y_2 \leq \ldots \leq y_n$ and then choosing either $y_{n/2}$ or $y_{(n+1)/2}$.

Therefore, the maximum likelihood estimator of μ is $\hat{\mu} = \text{median}(y_1, \ldots, y_n)$. Alternatively, it can be expressed as:

$$\hat{\mu}_{\tau=0.5} = \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} \frac{1}{2}|y_i - \mu| = \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} \rho_{0.5}(y_i - \mu),$$

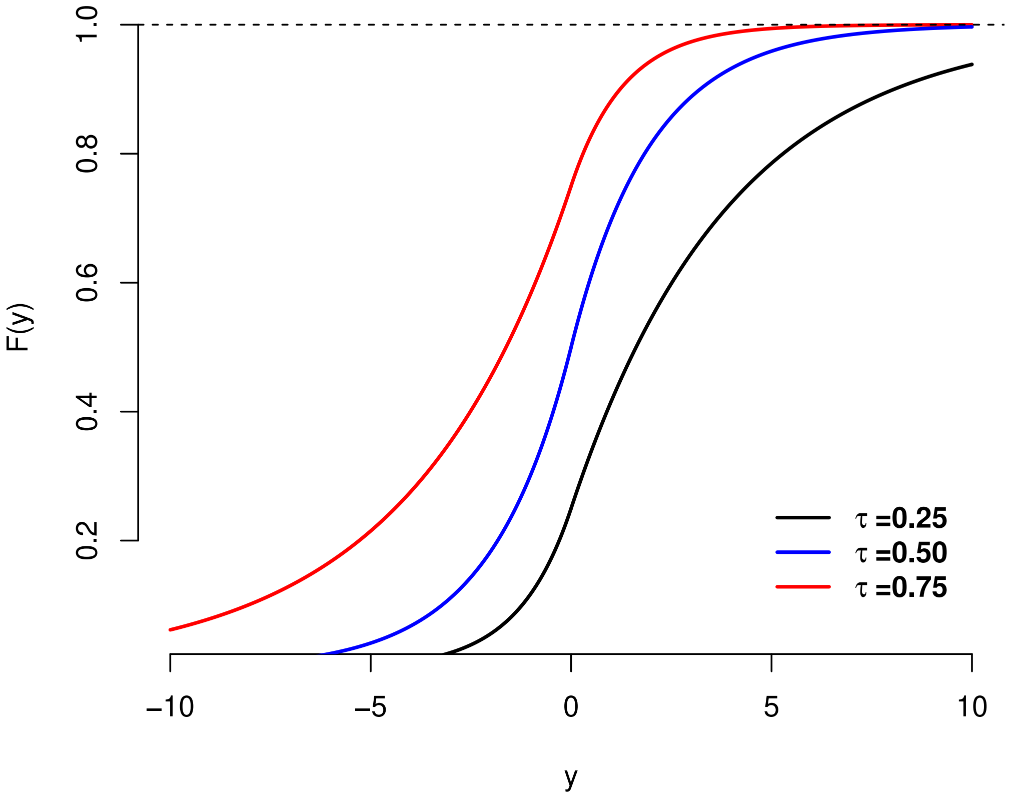


Figure 3.5: AL distribution function with $\mu = 0, \sigma = 1$, and $\tau \in (0.25, 0.50, 0.75)$.

where $\rho_{0.5}(u) = u(0.5 - I(u < 0))$ and $I(\cdot)$ represents an indicator function. For any given τ, the MLE of μ_τ can be described as:

$$\hat{\mu}_\tau = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \mu).$$

The AL distribution to quantile regression

Assuming that each independent response variable y_i, given $\mathbf{x}_i$ and $\boldsymbol{\beta}$, follows the AL distribution with $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}_\tau$, σ, and τ:

$$y_i|\mathbf{x}_i, \boldsymbol{\beta} \sim \mathcal{AL}(\mu_i, \sigma, \tau), \quad i = 1, \ldots, n. \tag{3.30}$$

The likelihood function can be expressed as

$$\mathcal{L}(\boldsymbol{\beta}, \sigma, \tau) = \left[\frac{\tau(1-\tau)}{\sigma}\right]^n \exp\left\{ - \sum_{i=1}^{n} \rho_\tau \left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) \right\}. \tag{3.31}$$

In this case, when τ is known, maximising the logarithm of (3.31) with respect to $\boldsymbol{\beta}$ yields $\hat{\boldsymbol{\beta}}_\tau$, which is equivalent to (3.25).

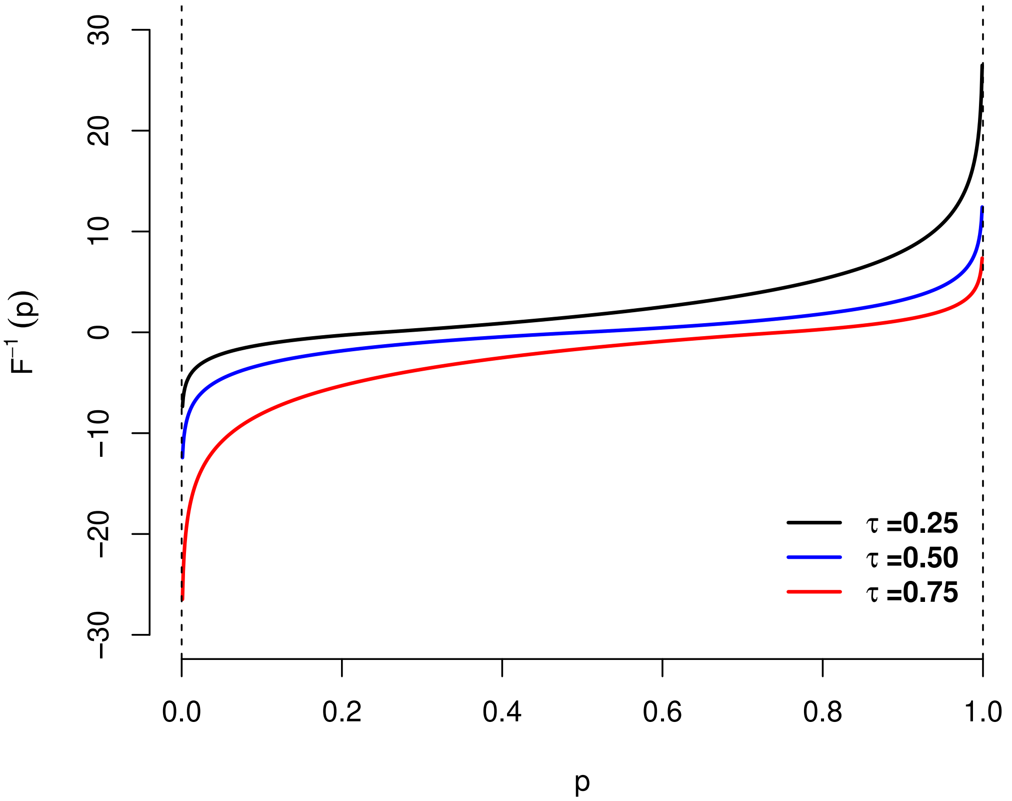


Figure 3.6: AL inverse (quantile) function with $\mu = 0, \sigma = 1$, and $\tau \in (0.25, 0.50, 0.75)$.

In its application, Yu and Moyeed (2001) leveraged this property to develop a Bayesian framework for the linear quantile regression model. This approach has since motivated numerous researchers to refine and apply this distribution to form quantile regression within various statistical frameworks. For example, Tsionas (2003) and Kozumi and Kobayashi (2011) demonstrated that the AL distribution could be formulated as a mixture of normals representation. This formulation facilitates the use of the Gibbs sampling algorithm to estimate model parameters. Furthermore, Geraci and Bottai (2007) employed this distribution to formulate the likelihood function of clustered or longitudinal data, enabling the estimation of the quantile functions.

3.5 Bayesian quantile regression

The Bayesian method offers an alternative statistical approach for estimating parameters based on underlying posterior information. This approach combines prior knowledge of unknown parameters with observable data to calculate the "posterior probability" using Bayes' theorem (Bayes & Price, 1763). This theorem is fundamentally defined for events

A and B as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where $P(A|B)$ is the conditional probability of event A occurring given that event B has occurred, and $P(A)$, $P(B)$ are the probabilities of events A and B occurring, respectively.

Let $\mathbf{y}$ be the observed data, and $\boldsymbol{\theta}$ denote the parameters of interest. According to Bayes's theorem, the posterior probability distribution, $p(\boldsymbol{\theta}|\mathbf{y})$, can be written as

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})},$$

where $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function of the observed data, $p(\boldsymbol{\theta})$ is the known prior probability distribution of the parameters, and $p(\mathbf{y})$ is the marginal probability distribution of the observed data. As it does not depend on $\boldsymbol{\theta}$, the posterior probability distribution can instead be written, up to proportionality, as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Therefore, this posterior is theoretically derived by multiplying the likelihood function of the data with the prior probability density of the relevant parameters (Gelman, 2014).

3.5.1 Prior probability distribution

As is known, the prior probability distribution is one of the two cores of the Bayesian method. In practice, there are two choices in this respect, depending on user specifications:

Noninformative prior: a type of prior commonly used due to the lack of knowledge about the parameters of interest. Hence, users seeks priors that exert minimal influence on the inference, essentially "letting the data speak for themselves" (Congdon, 2006, p.113). Moreover, the utilisation of this prior often offers a broad and diffuse range of possible values for the parameter. The Jeffrey's prior (Jeffreys, 1946) and the uniform prior distribution (Gelman, 2014) are two common types of noninformative priors widely used in Bayesian analysis.

Informative prior: a prior that is chosen based on substantive knowledge about the parameter of interest, without considering the current data. Typically, this prior knowledge or strong beliefs about the parameter come from previous research, expert opinion, or other relevant information sources (Gelman, 2014, p.34). As a result, this prior is often subjective, depending on the analyst's perspective. For instance, if an analyst believes that a continuous parameter varies around a certain value with some degree of variation, a normal distribution with its mean and standard deviation might be used as an informative prior to represent this belief.

3.5.2 Markov Chain Monte Carlo

In practice, the posterior probability distribution is usually complex and defined over high-dimensional spaces. Consequently, approximation methods, such as Markov Chain Monte Carlo (MCMC) algorithms, are often employed in Bayesian statistics to estimate the posterior distribution of a parameter (Gelman, 2014, page: 275-288). Both the Gibbs sampler and the Metropolis-Hastings algorithm are types of MCMC often used in this area.

Gibbs sampler

The Gibbs sampler, described by Geman and Geman (1984), was named after the physicist Josiah Willard Gibbs. The idea behind this sampler is to sequentially generate posterior samples from a set of conditional distributions rather than sampling directly from a joint distribution. The sampling is performed on only one variable at a time, conditional on the current values of all the other variables. This is particularly useful for sampling from high-dimensional probability distributions. The basic outline of the Gibbs sampler algorithm for drawing r samples from the posterior distribution is as follows.

1. Set initial values for a distribution of p parameters, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p$, defined as $\boldsymbol{\theta}_1^{(0)}, \boldsymbol{\theta}_2^{(0)}, \ldots, \boldsymbol{\theta}_p^{(0)}$.

2. For each iteration r, where $r = 1, 2, \ldots, R$, follow the steps below for sampling each parameter $i = 1, \ldots, p$:

   (a) Sample $\boldsymbol{\theta}_1^{(r)}$ from $p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(t-1)}, \boldsymbol{\theta}_3^{(t-1)}, \ldots, \boldsymbol{\theta}_p^{(t-1)}, \mathbf{y})$.

   (b) Sample $\boldsymbol{\theta}_2^{(r)}$ from $p(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_3^{(t-1)}, \ldots, \boldsymbol{\theta}_p^{(t-1)}, \mathbf{y})$.

   (c) Continue this process for the remaining parameters. This process implies for the ith parameter: Sample $\boldsymbol{\theta}_i^{(r)}$ from $p(\boldsymbol{\theta}_i | \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_{i-1}^{(t)}, \boldsymbol{\theta}_{i+1}^{(t-1)}, \ldots, \boldsymbol{\theta}_p^{(t-1)}, \mathbf{y})$.

   (d) Sample $\boldsymbol{\theta}_p^{(r)}$ from $p(\boldsymbol{\theta}_p | \boldsymbol{\theta}_1^{(r)}, \boldsymbol{\theta}_2^{(r)}, \ldots, \boldsymbol{\theta}_{p-1}^{(r)}, \mathbf{y})$.

As is known, the Gibbs sampler algorithm performs sampling from conditional distributions. These conditional distributions are required to be valid probability distribution from which samples can be drawn. In other word, they must be proper distributions, Hence, the Gibbs sampler has limitations in this respect. Consequently, other methods should be considered instead in cases where the conditional distribution is not straightforward to sample from, such as the Metropolis-Hastings algorithm.

Metropolis-Hastings

The Metropolis-Hastings algorithm is another MCMC technique, proposed by Hastings (1970). This algorithm generates a sequence of sample values from a chosen proposal distribution to match a target distribution (the distribution from which we want to sample) based on an acceptance criterion. Following this principle, the Metropolis-Hastings algorithm is more flexible and can be used even when the conditional distribution are not known or not straightforward to sample from. The Metropolis–Hastings algorithm can thus be written as follows:

1. Choose arbitrary values $\boldsymbol{\theta}_1^{(0)}, \boldsymbol{\theta}_2^{(0)}, \ldots, \boldsymbol{\theta}_p^{(0)}$ as the initial observations.

2. For each iteration $r = 1, 2, \ldots, R$, follow the steps below for sampling $\boldsymbol{\theta}_i^{(r)}$, where $i = 1, \ldots, p$:

   (a) From the current position $\boldsymbol{\theta}_i^{(r-1)}$, generate a set of candidate parameter values $\boldsymbol{\theta}_i^{*}$ from the proposal distribution $q(\boldsymbol{\theta}_i^{*}|\boldsymbol{\theta}_i^{(r-1)})$.

   (b) Compute the acceptance ratio α, given by:

   $$\alpha = \min\left(1, \frac{p(\boldsymbol{\theta}_i^{*}|\mathbf{y})q(\boldsymbol{\theta}_i^{(r-1)}|\boldsymbol{\theta}_i^{*})}{p(\boldsymbol{\theta}_i^{(r-1)}|\mathbf{y})q(\boldsymbol{\theta}_i^{*}|\boldsymbol{\theta}_i^{(r-1)})}\right),$$

   where $p(\boldsymbol{\theta}_i)$ is the target distribution, and $q(\boldsymbol{\theta}_i^{*}|\boldsymbol{\theta}_i)$ is the proposal distribution.

   (c) If acceptance occurs, set $\boldsymbol{\theta}_i^{(r)} = \boldsymbol{\theta}_i^{*}$. Otherwise, set $\boldsymbol{\theta}_i^{(r)} = \boldsymbol{\theta}_i^{(r-1)}$.

3.5.3 Convergence, mixing, acceptance rates, thinning and model diagnostics and effective sample size of the MCMC algorithm

In the MCMC algorithm, the objective is to construct a Markov chain whose sequence of sampled states converges to the target posterior distribution after iterating a sufficient number of times. The *convergence* is a crucial aspect, indicating that the Markov chain has the property to reach a stationary distribution that accurately represents the target distribution. In the event that the chain has converged, subsequent samples provide reliable estimates of the parameters or quantities of interest in the target distribution.

The question then arises: how is it determined whether the simulated Markov chain has converged in distribution to the target distribution? The basic tool widely used to diagnose convergence in this respect is the trace plot (see Figure 3.7). This plot visually presents the sampled values of a parameter or variable of interest across iterations. Typically, the

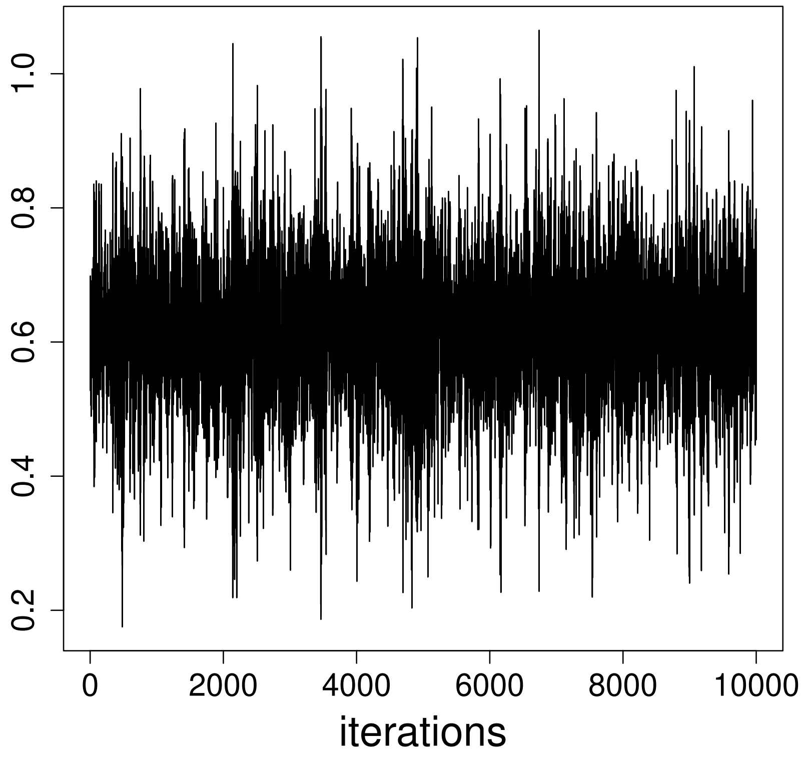


Figure 3.7: An example of a trace plot.

horizontal axis denotes the iteration number, while the vertical axis represents the sample values of the parameter or variable. Each point on the plot corresponds to a single sampled value obtained from the MCMC algorithm at particular iteration. The interpretation of the trace plot is straightforward, with flat and stable traces usually indicating convergence.

Another diagnostic tool used to assess convergence in MCMC simulations is the Gelman-Rubin statistic, also known as the potential scale reduction factor (Gelman & Rubin, 1992). This quantity measures convergence by comparing the variability between multiple chains to the variability within each individual chain. The underlying idea is that if multiple chains are sampling from the same distribution and have converged to the target distribution, then their variances should be similar. The Gelman-Rubin statistic is computed by

$$\hat{R}_{GR} = \sqrt{\frac{\hat{V}}{W}}.$$

where W is the average of the variances within each chain and $\hat{V}$ is the estimated variance of the target distribution. Both terms can be computed by

$$W = \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{R_m - 1} \sum_{r=1}^{R_m} (\theta_r^m - \hat{\theta}_m)^2 \right),$$

and

$$\hat{V} = \frac{R-1}{R} W + \frac{1}{R} B,$$

where M is the number of chains, R is the number of iterations, $\hat{\theta}_m$ is the (posterior) mean of the m-th chain, B is the between-chain variance defined as $B = \frac{R}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta})^2$, and $\hat{\theta}$ is the overall mean of all chains defined as $\hat{\theta} = \frac{1}{M} \sum_{m}^{M} \hat{\theta}_m$. Ideally, if the M chains have converged to the target distribution, the value of $\hat{R}_{GR}$ should approach 1 as the number of iteration, R, increases. In practice, $\hat{R}_{GR} \leq 1.1$ typically indicates that the Markov chain has converged. Values greater than 1.1 may suggest non-convergence. This is considered a more stringent criterion for $\hat{R}_{GR}$ (Brooks & Gelman, 1998). In this thesis, convergence will be assessed by examining parameter trace plots as well as by using the Gelman-Rubin statistic.

Another point that needs to be considered in the MCMC algorithm is how effectively the Markov chain explores the space of possible states. This refers to a term called *mixing*. In MCMC, when the chain mixes well, it means that the exploration of the space of possible parameter values can move quickly from one value to another. This ensures that the samples generated by the chain accurately represent the target distribution. In contrast, poor mixing can lead to slow exploration of the space and sometimes result in getting trapped in certain regions, providing inefficient sampling and potentially biased or inaccurate estimates. In practice, trace plots for individual parameters are typically used to investigate mixing.

As mentioned previously, mixing refers to the ability of the MCMC algorithm to transition effectively from one state to another within the state space. One factor that closely influences the efficiency and effectiveness of the algorithm in exploring the target distribution is the *acceptance rate*. This rate is defined as the proportion of proposed moves that are accepted during the simulation. A too high acceptance rate may force the algorithm to accept too many moves within a limited region of the state space, potentially causing inefficient exploration of other regions, while a too low acceptance rate can result in poor mixing because proposed moves are frequently rejected.

In addition, samples generated from MCMC algorithms typically exhibit autocorrelation because each sample depends on the previous one. This leads to increased uncertainty associated with the estimation of posterior quantities of interest (e.g. mean, variance, and quantiles), resulting in wider confidence intervals, less precision, and inflated standard errors (Link & Eaton, 2012). To address this issue, a set of samples is selected by taking only every kth sample from the posterior distribution and discarding all others (Johansen, 2010; Link & Eaton, 2012), a process known as *thinning*. However, thinning may unnecessarily discard information when summarising the Markov chain (Link & Eaton, 2012). In this thesis, thinning is not considered for this reason. To mitigate the impact of autocorrelation in MCMC algorithms, it is essential to ensure that the *effective sample size* (ESS) is sufficiently large to obtain reliable information about the (target) posterior distributions. The ESS is the approximate number of independent samples for any parameter of interest from the MCMC chain (Gelman, 2014), and can be defined as

$$n_{\text{ESS}} = \frac{n}{1 + 2\sum_{k}^{K} \rho_k},$$

where n is the total number of MCMC samples, ρ_k is the autocorrelation at lag K, and K is the lag at which the autocorrelation drops below the threshold.

The performance and efficiency of the algorithms depend on many aspects. The initial values can significantly impact in this, in both the Gibbs sampler and the Metropolis-Hastings algorithm, if they are not selected to be consistent with the true parameter values (Turner et al., 2013; Van Ravenzwaaij et al., 2018). For instance, initial values that are either too large or too small compared to the mode of the target distribution tend to result in slower convergence. Therefore, the selection of initial values should be done carefully. Another factor is the length of the "burn-in" period, an initial phase of sampling which is discarded. This is because the burn-in period allows for consideration of only those chains that yield sample values representative of the equilibrium distribution (Johansen, 2010). Having too long a burn-in period may be costly in terms of computational resources, while too short a period may retain non-equilibrium samples. Furthermore, (this factor relates to the first in terms of) selecting poorly chosen initial values may require longer burn-in periods. In practice, there is no specific rule to determine the length of the burn-in period, but it is usually investigated using trace plots of the parameters. Nevertheless, there are other factors impacting specific algorithms. For instance, in the Metropolis-Hastings algorithm, an inappropriate proposal distribution can affect acceptance rates, leading them to be too low, and result in inefficient exploration of the parameter space (Johansen, 2010).

3.5.4 Posterior summary statistics

In Bayesian statistics, the point estimate, $\hat{\theta}$, is commonly obtained from the mean or median of the posterior distribution, known as the posterior mean and posterior median, respectively. The credible interval represents the interval estimates of parameters of interest in this context. In practice, this interval is commonly obtained by the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior distribution of θ. Note that the credible interval can be interpreted as indicating that there is a $100(1-\alpha)\%$ chance that this random parameter θ belongs to an interval, which is determined based on the observed data.

3.5.5 Bayesian methods for quantile regression

It is crucial to emphasise that both the likelihood function and priors play a central role in the Bayesian method. In the context of quantile regression, Yu and Moyeed (2001) pioneered the use of the asymmetric Laplace distribution, as mentioned in Section 3.4.6, to represent the likelihood function of data. Alternatively, due to computational aspects, the scale mixture of normals of the AL distribution emerges as another viable option (Alhamzawi et al., 2012; Alhamzawi, 2013; Kozumi & Kobayashi, 2011; Tsionas, 2003). For prior specifications, there is no specific standard prior for the quantile regression coefficients. For example, an improper uniform distribution was utilised in the work of Yu and Moyeed (2001), while symmetric prior distributions, such as a normal distribution, usually appeared in Bayesian quantile regression model based on a location-scale mixture representation of the asymmetric Laplace distribution (Alhamzawi, 2013; Kozumi & Kobayashi, 2011). Therefore, this section reviews some principal methods that are relevant to Bayesian quantile regression.

Extending the AL distribution to Bayesian linear quantile regression

Assume each Y_i is independent and that errors ϵ_i follow the AL distribution with density given in (3.28) when $\mu = 0$, $\sigma = 1$, and $\int f_\tau(\epsilon)d\epsilon = \tau$, with $f_\tau(\cdot)$ denoting the error density. This implies that each response variable, Y_i, given $\boldsymbol{x}_i$ and the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$, follows the AL distribution with $\mu_i = Q_\tau(Y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau$, and a fixed τ (Yu & Moyeed, 2001):

$$Y_i|\boldsymbol{x}_i, \boldsymbol{\beta} \sim \mathrm{AL}(\boldsymbol{x}_i'\boldsymbol{\beta}_\tau, \sigma = 1, \tau), \quad i = 1, \ldots, n. \tag{3.32}$$

This leads to the formation of the linear quantile regression model as:

$$Y_i = \mu_i + \epsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau + \epsilon_i.$$

Based on the model (3.32), we can form the likelihood of $Y_1, \ldots, Y_n$ as follows:

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{x}_i, \boldsymbol{\beta}, \tau) = \tau^n(1-\tau)^n \exp\left\{ -\sum_{i=1}^{n} \rho_\tau \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right) \right\}.$$

Therefore, we can employ the above likelihood function to form the Bayesian framework:

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})p(\boldsymbol{\beta}).$$

For easier understanding, we illustrate the simple linear median regression:

$$Y_i = \beta_{0,\tau} + \beta_{1,\tau}x_i + \epsilon_i,$$

where $\tau = 1/2$. Hence, each response variable, Y_i, given $\boldsymbol{x}_i$ and the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, has the distribution as:

$$Y_i|\boldsymbol{x}_i, \boldsymbol{\beta} \sim \mathsf{AL}(\boldsymbol{x}_i'\boldsymbol{\beta}_\tau, \sigma = 1, \tau = 0.5), \quad i = 1, \ldots, n.$$

In this case, the likelihood function used to represent the probability distribution of data can be expressed as follows:

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}|\boldsymbol{x}_i, \boldsymbol{\beta}) &= \left(\frac{1}{4}\right)^n \left\{ -\sum_{i=1}^{n} \rho_{1/2} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right) \right\} \\ &= \left(\frac{1}{4}\right)^n \left\{ -\sum_{i=1}^{n} \frac{1}{2}\Big|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\Big| \right\}. \end{aligned}$$

In the next step, we need to determine the joint prior distribution of regression parameters, $p(\boldsymbol{\beta}) = p(\beta_0, \beta_1)$. Yu and Moyeed (2001) pointed out that there is limited knowledge about this prior information in the area of quantile regression. They suggested employing a noninformative prior to handle this situation. Additionally, they proved that an improper uniform prior distribution $p(\boldsymbol{\beta}_\tau) \propto 1$, can yield the proper joint posterior distribution, $p(\boldsymbol{\beta}_\tau|\boldsymbol{y}, \boldsymbol{X})$ (for more details, see Yu and Moyeed (2001)). Owing to this, the joint posterior distribution of the parameter of interest appears to be proportional to the likelihood surface. Consequently, this leads to providing the posterior distribution with an unknown form. Given this fact, the model parameters can be estimated via the Metropolis-Hastings algorithm because the full conditional for each parameter does not correspond to any known standard distribution (Benoit & Poel, 2017).

Location-scale mixture of normals representation of the AL distribution

Tsionas (2003) demonstrated that the AL distribution can be formulated as a mixture of normals representation. This discovery can lead to the use of the Gibbs sampling algorithm for more effective estimation of the model parameters compared to the Bayesian quantile regression proposed by Yu and Moyeed (2001). Furthermore, Kozumi and Kobayashi (2011) applied this evidence to enhance the Gibbs sampling, which can yield more efficient results than the previous method.

Following this representation, the quantile regression model (3.24) can be rewritten as:

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau + \underbrace{\theta v_i + \omega\sqrt{\sigma v_i}u_i}_{\epsilon_i},$$

where

$$\theta = \frac{1-2\tau}{\tau(1-\tau)}, \quad \omega^2 = \frac{2}{\tau(1-\tau)},$$

and $v_i = \sigma z_i$. Both $z_i \sim \text{Exp}(1)$ and $u_i \sim N(0,1)$ are mutually independent random variables. This leads to each response variable, Y_i, given z_i, having the normal distribution with mean $\boldsymbol{x}_i'\boldsymbol{\beta}_\tau + \theta v_i$ and variance $\omega^2\sigma v_i$.

Let $\boldsymbol{y} = (y_1,\ldots,y_n)'$ and $\boldsymbol{v} = (v_1,\ldots,v_n)'$. Hence, the Bayesian hierarchical quantile model can be expressed as:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{v},\boldsymbol{\beta}_\tau &\sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}_\tau + \theta v_i, \omega^2\sigma v_i), \quad i = 1,\ldots,n\\
v_i &\sim \text{Exp}(\sigma),\\
\boldsymbol{\beta}_\tau &\sim \mathcal{N}(\boldsymbol{\beta}_{\tau,0}, \boldsymbol{B}_{\tau,0}),\\
\sigma &\sim \text{InvGamma}\left(\frac{n_0}{2}, \frac{s_0}{2}\right),
\end{aligned}$$

where $\boldsymbol{\beta}_{\tau,0}$ and $\boldsymbol{B}_{\tau,0}$ are the prior mean and covariance of $\boldsymbol{\beta}_\tau$, respectively. Here, $\text{InvGamma}(a,b)$ denotes an inverse Gamma distribution with shape parameter a and scale parameter b. Also, n_0 and s_0 are the prior shape and scale of σ.

To form the Gibbs sampler, $\boldsymbol{\beta}_\tau$, $\boldsymbol{v}$ and σ, are sampled from their conditional distributions. Then, the full conditional distribution of $\boldsymbol{\beta}_\tau$ is given by

$$\boldsymbol{\beta}_\tau|\boldsymbol{y},\boldsymbol{v},\sigma \sim N(\widetilde{\boldsymbol{\beta}}_\tau, \widetilde{\boldsymbol{B}}_\tau),$$

where

$$\widetilde{\boldsymbol{B}}_\tau^{-1} = \sum_{i=1}^n \frac{\boldsymbol{x}_i\boldsymbol{x}_i'}{\omega^2\sigma v_i} + \boldsymbol{B}_{\tau,0}^{-1} \quad \text{and} \quad \widetilde{\boldsymbol{\beta}}_\tau = \widetilde{\boldsymbol{B}}_\tau\left\{\sum_{i=1}^n \frac{\boldsymbol{x}_i(y_i-\theta v_i)}{\omega^2\sigma v_i} + \boldsymbol{B}_{\tau,0}^{-1}\boldsymbol{\beta}_{\tau,0}\right\}.$$

The full conditional distribution of each v_i is then a generalised inverse Gaussian (GIG) distribution,

$$v_i|\boldsymbol{y},\boldsymbol{\beta}_\tau,\sigma \sim \text{GIG}\left(\frac{1}{2},\delta_{1i},\delta_{2i}\right),$$

where

$$\delta_{1i}^2 = \frac{(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2}{\omega^2\sigma} \quad \text{and} \quad \delta_{2i}^2 = \frac{2}{\sigma} + \frac{\theta^2}{\omega^2\sigma}.$$

The full conditional distribution of σ is an inverse Gamma distribution, given by

$$\sigma|\boldsymbol{y},\boldsymbol{\beta}_\tau,\boldsymbol{v} \sim \text{InvGamma}\left(\frac{\widetilde{a}}{2},\frac{\widetilde{b}}{2}\right),$$

where

$$\widetilde{a} = n_0 + 3n \quad \text{and} \quad \widetilde{b} = s_0 + 2\sum_{i=1}^n v_i + \sum_{i=1}^n \frac{(y_i-\boldsymbol{x}_i'\boldsymbol{\beta}_\tau-\theta v_i)^2}{\omega^2 v_i}.$$

This approach can be implemented using the function bayesQR in the bayesQR package (Benoit & Poel, 2017) in R. Note that this package follows the Gibbs sampling algorithm as proposed by Kozumi and Kobayashi (2011).

However, a random variable based on the AL distribution can be conveniently expressed as a scale mixture of normals in an alternative manner (Alhamzawi & Yu, 2013; Alhamzawi et al., 2012). Assuming that the error term in the quantile regression model follows $\epsilon_i \sim \mathcal{AL}(0,\sigma,\tau)$ and letting $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)'$, the likelihood function can be denoted by

$$l(\boldsymbol{\epsilon}|\sigma) \propto \sigma^{-n}\exp\left\{-\sum_{i=1}^n \frac{|\epsilon_i|+(2\tau-1)\epsilon_i}{2\sigma}\right\}. \tag{3.33}$$

Andrews and Mallows (1974) showed that for any $a,b>0$,

$$\exp\{-|ab|\} = \int_0^\infty \frac{a}{\sqrt{2\pi e}}\exp\left\{-\frac{1}{2}(a^2e+b^2e^{-1})\right\}. \tag{3.34}$$

Applying equation (3.34) to the right-hand side of the function (3.33) by assuming $a =$

$1/\sqrt{2\sigma}$, $b = \epsilon/\sqrt{2\sigma}$ and multiplying by $\exp\{-(2\tau-1)\epsilon/2\sigma\}$, it results in

$$\sigma^{-n}\exp\left\{-\sum_{i=1}^{n}\frac{|\epsilon_i|+(2\tau-1)\epsilon_i}{2\sigma}\right\} \\ = \prod_{i=1}^{n}\int_0^\infty \frac{1}{\sigma\sqrt{4\pi\sigma v_i}}\exp -\left\{\frac{(\epsilon_i-\xi v_i)^2}{4\sigma v_i}-\zeta v_i\right\}dv_i \tag{3.35}$$

where $\xi = (1-2\tau)$, $\zeta = \tau(1-\tau)/\sigma$ and $v_i \sim \mathrm{Exp}(\zeta)$. Further, let $\epsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}_\tau$, this implies that

$$l(\boldsymbol{y}|\boldsymbol{v},\boldsymbol{\beta},\sigma) \propto \prod_{i=1}^{n}\int_0^\infty \frac{1}{\sigma\sqrt{4\pi\sigma v_i}}\exp -\left\{\frac{(y_i-\mathbf{x}_i'\boldsymbol{\beta}_\tau-\xi v_i)^2}{4\sigma v_i}-\zeta v_i\right\}dv_i.$$

Thus, this thesis applies the scale mixture of normals (3.35) in the context of quantile regression model.

3.6 Review of QR models for longitudinal data

This section reviews some QR models applied to longitudinal data. These models serve as an essential foundation for expansion to other QR variants, such as the flexible QR models.

3.6.1 Quantile regression model with fixed-effects

The model proposed by Koenker (2004) is one of the best-known classical quantile regression models for longitudinal data. Koenker added an individual-specific term to the classical QR model, similar to the approach of mixed-effects models. The general form of the τth conditional quantile model of the response Y_{ij} is

$$Q_\tau(y_{ij}|\mathbf{x}_{ij},u_i) = \mathbf{x}_{ij}'\boldsymbol{\beta}_\tau + u_i + \epsilon_{ij}, \quad i=1,\ldots,N, \quad j=1,\ldots,n_i.$$

Here, u_i are treated as the individual pure location shift effects on the conditional quantiles of the response. Consequently, these effects differ from the random individual effects in the standard mixed-effects model framework. The random error ϵ_{ij} is assumed to follow the τth conditional quantile equal to zero. The fixed-effect parameter $\boldsymbol{\beta}$ is permitted only to describe the relationship between the explanatory variable $\mathbf{x}_{ij}$ and the τth quantiles. Parameter estimates are obtained by solving

$$\min_{\boldsymbol{\beta},u}\sum_{k=1}^{q}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\omega_k\rho_{\tau_k}(y_{ij}-u_i-\mathbf{x}_{ij}'\boldsymbol{\beta}), \tag{3.36}$$

where $\rho_{\tau_k}, k = 1, \ldots, q$, is the quantile loss function and ω_k are the weights employed to avoid the crossing quantile when estimating u_i. Koenker (2004) noted that solving problem (3.36) may be impossible when each data dimension (i.e. q, N and n_i) is large.

3.6.2 Penalised quantile regression model with fixed effects

One possible way to address the limitation of (3.36) when dealing with a large sample size is by incorporating a penalty term into the objective function to shrink u_i, as expressed in the equation:

$$\min_{\boldsymbol{\beta},u} \sum_{k=1}^{q} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \omega_k \rho_{\tau_k}(y_{ij} - u_i - \mathbf{x}_{ij}'\boldsymbol{\beta}) + \lambda \sum_{i=1}^{n} |u_i|.$$

In this case, λ is the penalisation parameter used to shrink the location shift parameters u_i to control the variability of the individual-specific effects. The optimal value of this parameter can be chosen via an asymptotic approximation proposed by Lamarche (2010).

The selection of the penalisation parameter, which may impact the estimates, is the primary drawback of this method (Geraci & Bottai, 2007; Marino & Farcomeni, 2015). A poor choice in this parameter introduces bias into the estimates. Another issue is that the number of parameters in the model depends on the number of subjects N, meaning that as the sample size increases, more parameters will need to be estimated (Geraci & Bottai, 2007). Moreover, some studies found that when the number of repeated measurements (n_i) is small, the estimates may exhibit large biases (Kato et al., 2012; Marino & Farcomeni, 2015). In terms of applications, this method is extensively used in economics, but is relatively rare in models for longitudinal child growth data.

3.6.3 Linear quantile mixed-effects models

Geraci and Bottai (2007) formulated the quantile regression model with random intercept effects using the likelihood-based method described in Section 3.4.6. These effects are analogous to the random effects used in mixed-effects models, inducing dependence between observations taken from the same subject on different occasions. Following the likelihood-based method, the authors introduced the AL distribution as a corresponding distribution for random errors, ϵ, in order to estimate the conditional quantile functions. Geraci and Bottai (2014) also extended this model to include more general random effects, such as random slopes.

The linear mixed quantile models of the response Y_{ij} can be defined as

$$Q_\tau(y_{ij}|\mathbf{x}_{ij}, \boldsymbol{u_i}) = \mathbf{x}'_{ij}\boldsymbol{\beta}_\tau + \mathbf{z}'_{ij}\boldsymbol{u}_i + \epsilon_{ij}, \tag{3.37}$$

where $Q_\tau(\cdot)$ is the quantile function of the response y_{ij} conditional on the random effects $\boldsymbol{u}_i$ and covariates $\mathbf{x}_{ij}$. Note that if $\boldsymbol{z}_{ij} = (1, \ldots, 1)$, the model (3.37) simplifies to the random intercepts QR model. Assuming that each y_{ij} given $\boldsymbol{u}_i$ and $\mathbf{x}_{ij}$, follows the AL distribution with location, scale and skewness parameters, which defined by $\mu_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}_\tau + \mathbf{z}'_{ij}\boldsymbol{u}_i$, σ and τ, respectively:

$$f(y_{ij}; \boldsymbol{\beta}, \boldsymbol{u}_i, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma}\exp\left\{ -\rho_\tau\left(\frac{y_{ij}-\mu_{ij}}{\sigma}\right)\right\},$$

where $\tau \in (0, 1)$ is a fixed and known parameter.

To form the likelihood function, each random term must follow three specific assumptions:

- $\boldsymbol{u}_i$ is a random vector independent of the random errors term and distributed to follow $f_u(\boldsymbol{u}_i; \boldsymbol{\Phi})$ where $\boldsymbol{\Phi}$ is a $q \times q$ variance-covariance matrix and must be a real symmetric positive-definite matrix. Note that $\boldsymbol{u}$ depends on τ through $\boldsymbol{\Phi}$. In the original work, Geraci and Bottai (2014) made two choices of assumptions about random effects, i.e. a normal distribution and a symmetric Laplace distribution.
- $\epsilon_{ij} \sim \mathcal{AL}(0, \sigma, \tau)$, and
- $\boldsymbol{u}_i$ and ϵ_{ij} are independent of one another.

The marginal likelihood function can be defined as

$$\mathcal{L}(\boldsymbol{\beta}, \sigma, \tau, \boldsymbol{\Phi}) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(y_{ij}; \boldsymbol{\beta}, \boldsymbol{u}_i, \sigma, \tau) f_u(\boldsymbol{u}_i; \boldsymbol{\Phi}) d\boldsymbol{u}_i. \tag{3.38}$$

It is clear that the concept of this approach is similar to the mixed-effects models for the mean. The main difference is that the random error term is assumed to be the AL distribution to allow for estimating the quantile coefficients. One main difficulty with this approach is that the marginal likelihood function (3.38) has an integral term that does not have a closed-form solution. Thus, it would need to apply some methods to approximate that term, e.g. Gaussian quadrature methods (see more details in Geraci and Bottai (2014)).

3.7 Splines

This section introduces the spline methods used to represent non-linear dependence in the context of regression. The first two subsections introduce the fundamental concept of splines by describing the simplest splines. The last subsection outlines some of the most popular spline methods, which are generally chosen due to their simplicity and attractive numerical properties.

3.7.1 Polynomial splines

Consider a quadratic regression model (a polynomial of degree 2),

$$y_i = g(x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$

A polynomial curve can be fitted by the regression function of the form:

$$g(x_i) = \beta_0 B_0(x_i) + \beta_1 B_1(x_i) + \beta_2 B_2(x_i),$$

where $B_0(x) = 1$, $B_1(x) = x$ and $B_2(x) = x^2$ are called *basis* functions, and β_0, β_1 and β_2 are *basis* coefficients. More generally, for a polynomial of degree d, the regression function can be written as a linear combination of p basis functions,

$$g(x_i) = \sum_{j=0}^{p} B_j(x_i)\beta_j.$$

It can also be represented in matrix form as

$$g(\mathbf{x}) = \mathbf{B}\boldsymbol{\beta},$$

where

$$\mathbf{B} = \begin{pmatrix} B_0(x_1) & B_1(x_1) & \cdots & B_d(x_1) \\ B_0(x_2) & B_2(x_2) & \cdots & B_d(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_0(x_n) & B_3(x_n) & \cdots & B_d(x_n) \end{pmatrix} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^d \end{pmatrix}$$

The example above is the fitted global function, meaning that the fitted function is estimated to follow the form of a particular polynomial function throughout the range of x. Typically, those forms may be too restrictive to capture non-trivial curvature at some local points (see Figure 3.8). As a result, the fitted curve may not reflect the underlying shape of the data. In pursuit of a better alternative, it is critical to consider partitioning the range of x into smaller intervals. Following that, the regression function of each interval is locally fitted. Then, each function is connected, resulting in a smooth curve. This is one

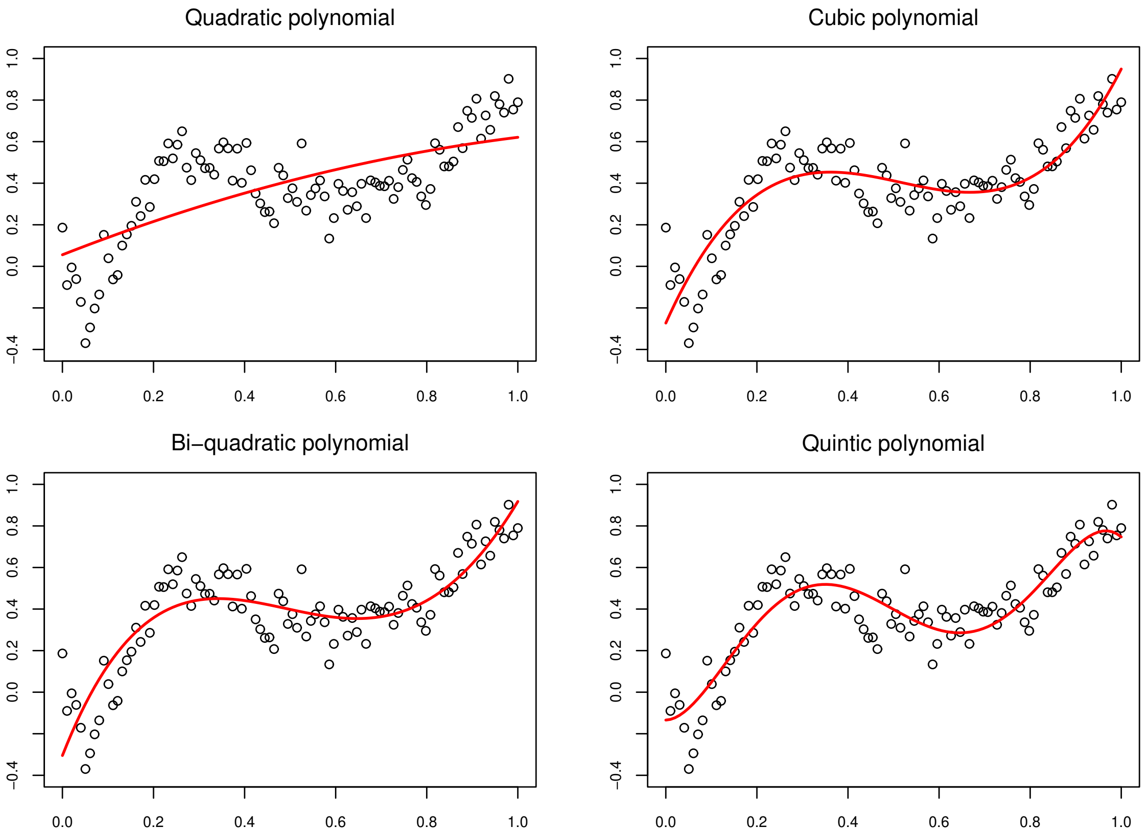


Figure 3.8: Fitting a simulated non-linear data relationship by varying the degree of polynomial

kind of "local" regression. There are several methods available for achieving this, with the projection of the data onto a much smaller set of *locally defined basis functions* being one of the most prominent.

3.7.2 Truncated power series

A popular set of locally defined basis functions is known as *truncated power basis functions* (Ruppert et al., 2003),

$$g(x) = \gamma_0 + \gamma_1 x + \ldots + \gamma_d x^d + \sum_{k=1}^{K} \beta_{dk}(x - \kappa_k)_+^d,$$

where $d \geq 1$ is the degree of the polynomial, $\kappa_1, \ldots, \kappa_K$ represent the position of points, commonly referred to as the *knots*,

$$(u)_+ = \begin{cases} u, & \text{if}\, u > 0 \\ 0, & \text{otherwise} \end{cases}$$

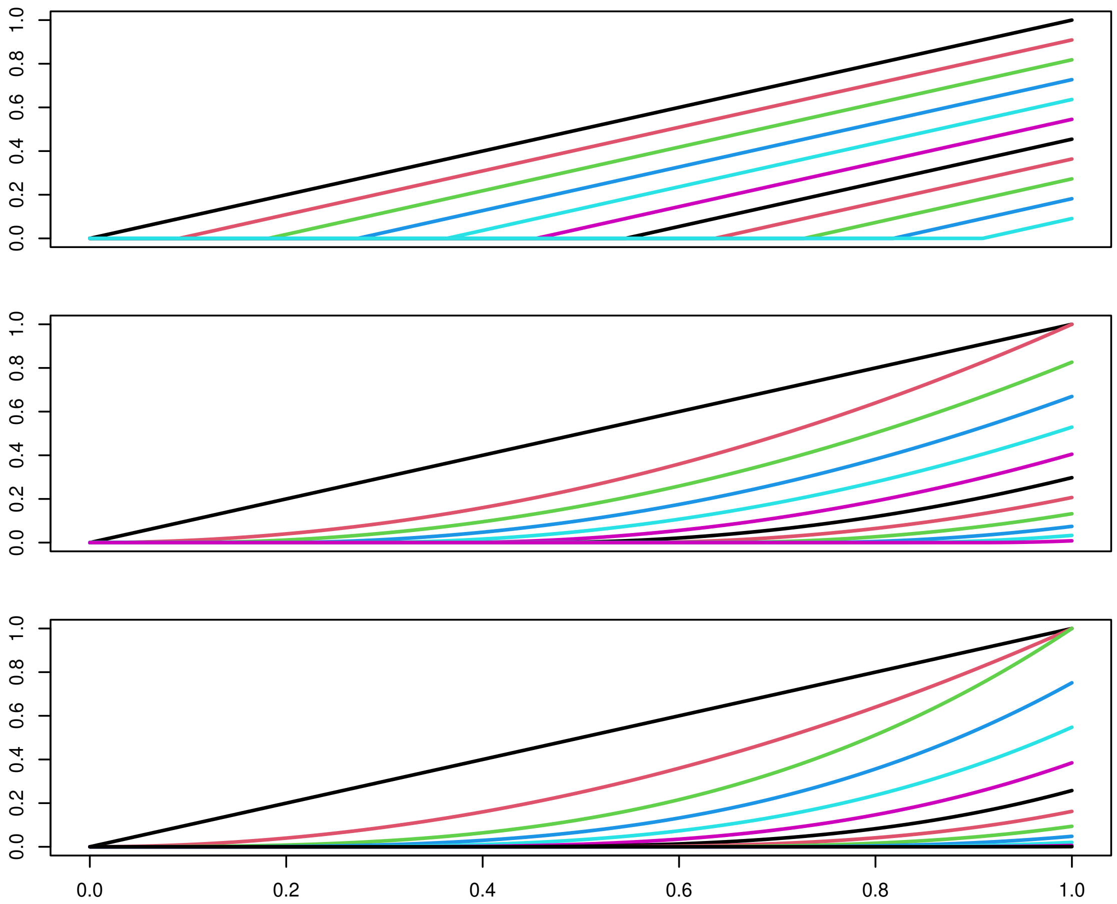


Figure 3.9: Truncated power basis functions of degree 1 (linear), 2 (quadratic) and 3 (cubic) with ten equidistant knots (K=10), respectively.

This basis system includes the basis functions $1, x, \ldots, x^d, (x-\kappa_1)_+^d, \ldots, (x-\kappa_K)_+^d$.

These basis functions offer advantages in terms of their simplicity in construction and interpretation, especially when the truncated line basis is d=1 (Wand, 2003). However, they do not form an orthogonal basis, leading to numerical instability when a large number of knots is defined (Hastie & Tibshirani, 1999). Figure 3.9 shows the truncated power basis functions with different degrees of the polynomial.

3.7.3 Restricted cubic splines

Restricted cubic splines, also referred to as *natural cubic* splines, consist of cubic polynomials constrained by continuity and slope conditions at each knot. Additionally, there is an extra requirement for linearity at the curve's endpoints (also known as boundary knots), typically preceding the first knot and succeeding the final one (Harrell, 2015; Perperoglou et al., 2019). These boundary conditions typically require the second derivative (curvature) to be zero at the endpoints of the spline, which effectively prevents the curve from exhibiting excessive oscillations beyond the range of the data (Hastie & Tibshirani, 1999; Hastie et al., 2009). The restricted cubic splines with K knots $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ and boundary knots $(x < \kappa_1)$ and $(x > \kappa_K)$, can be expressed as:

$$g(x) = \gamma_0 + \gamma_1 x + \sum_{k=1}^{K-2} \beta_k (x - \kappa_k)_*^3,$$

where

$$\begin{aligned}(x - \kappa_k)_*^3 =& (x - \kappa_k)_+^3 - (x - \kappa_{K-1})_+^3 \frac{\kappa_K - \kappa_k}{\kappa_K - \kappa_{K-1}} \\ &+ (x - \kappa_K)_+^3 \frac{\kappa_{K-1} - \kappa_k}{\kappa_K - \kappa_{K-1}}, k = 1, 2, \ldots, K-2.\end{aligned}$$

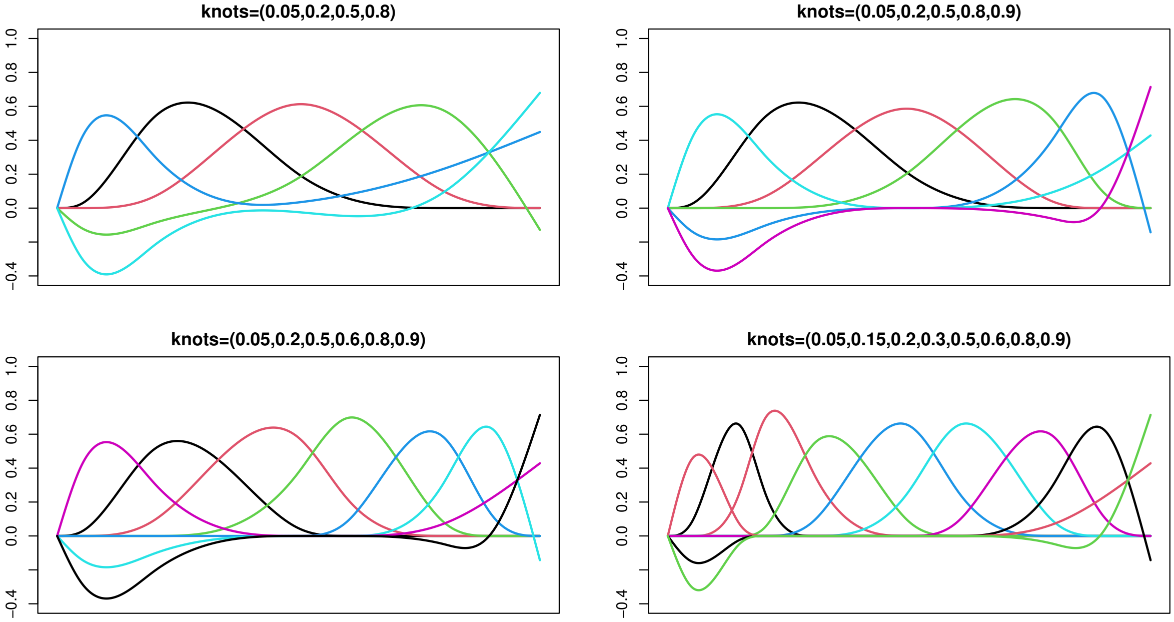


Figure 3.10: Restricted cubic spline basis functions with four knots specified, utilising boundary knots (0,1).

Figure 3.10 illustrates the restricted cubic splines with different knots specified when boundary knots (0,1) were applied.

3.7.4 B-splines

Another popular choice is B-splines (de Boor, 1972). These have the appealing property that any given basis function is only nonzero over a span of a small number of adjacent knots, thus resulting in a sparse design matrix, which is convenient for estimating the regression coefficients. As a result, such splines tend to avoid computational issues. The general form of the unknown function that consists of a set of B-splines basis functions is given as

$$g(x) = \sum_{j=0}^{p} \gamma_j B_{j,d}(x),$$

where $B_{j,d}(x)$ represents the jth basis function with a piecewise polynomial of degree d, and γ_j are the corresponding coefficients. Let $\kappa = (\kappa_0, \kappa_1, \ldots, \kappa_{p+d+1})$ be the knots vector where $\kappa_0 \leq \kappa_1 \leq \cdots \leq \kappa_{p+d+1}$ and p is the number of basis functions. Note that the knot vector in this context includes both internal and external knots. It consists of two knots, typically referred to as boundary knots, usually (but not always) placed at the minimum and maximum of x, which serve to anchor the B-spline basis. For each individual B-spline basis function, $B_{j,d}(x)$, it can be defined recursively, for $d = 0$ as:

$$B_{j,0}(x) = \begin{cases} 1 & \kappa_j \leq x < \kappa_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

and for $d > 0$,

$$B_{j,d}(x) = \frac{x - \kappa_j}{\kappa_{j+d} - \kappa_j} B_{j,d-1}(x) + \frac{\kappa_{j+d+1} - x}{\kappa_{j+d+1} - \kappa_{j+1}} B_{j+1,d-1}(x).$$

Figures 3.11 and 3.12 show two different B-spline basis functions with knots placed uniformly and non-uniformly, respectively, and with different degrees of the polynomial.

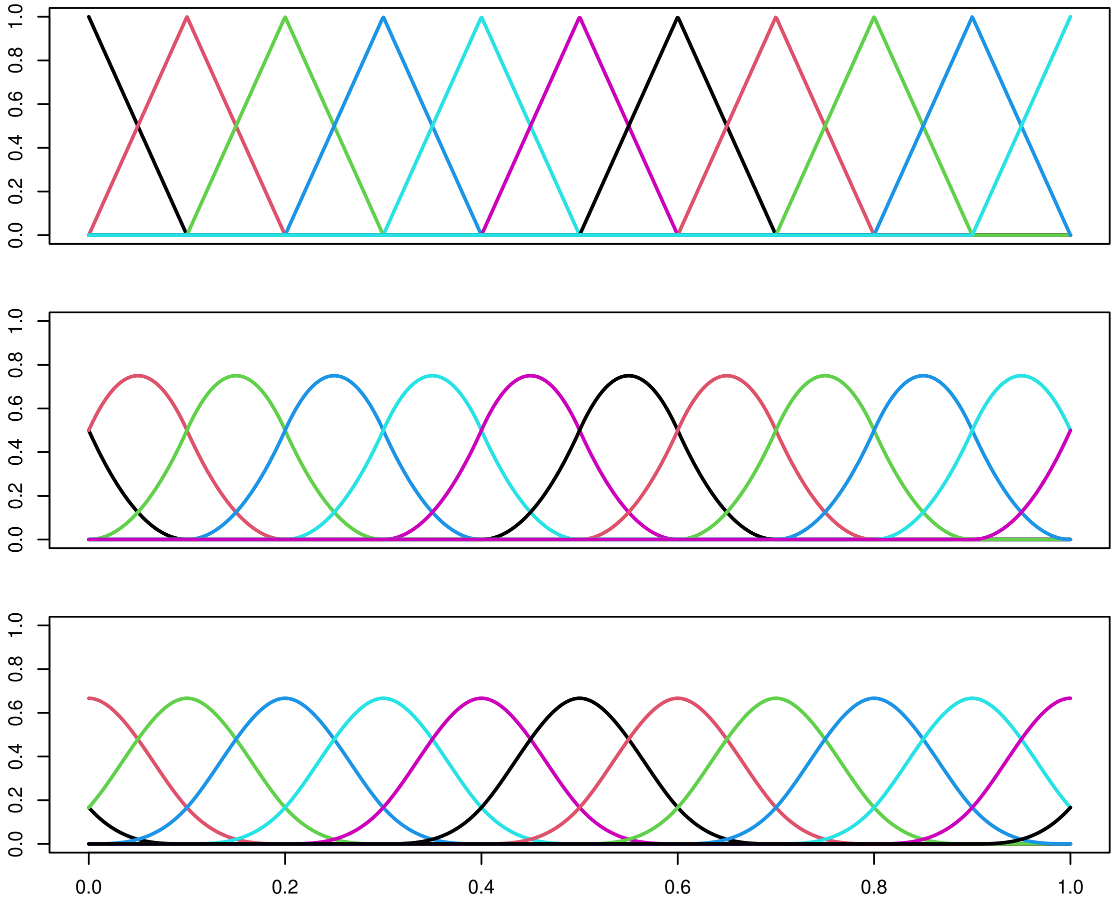


Figure 3.11: Uniform B-spline basis functions of degree 1 (linear), 2 (quadratic) and 3 (cubic) with thirteen knots (eleven internal knots and two external knots) , respectively.

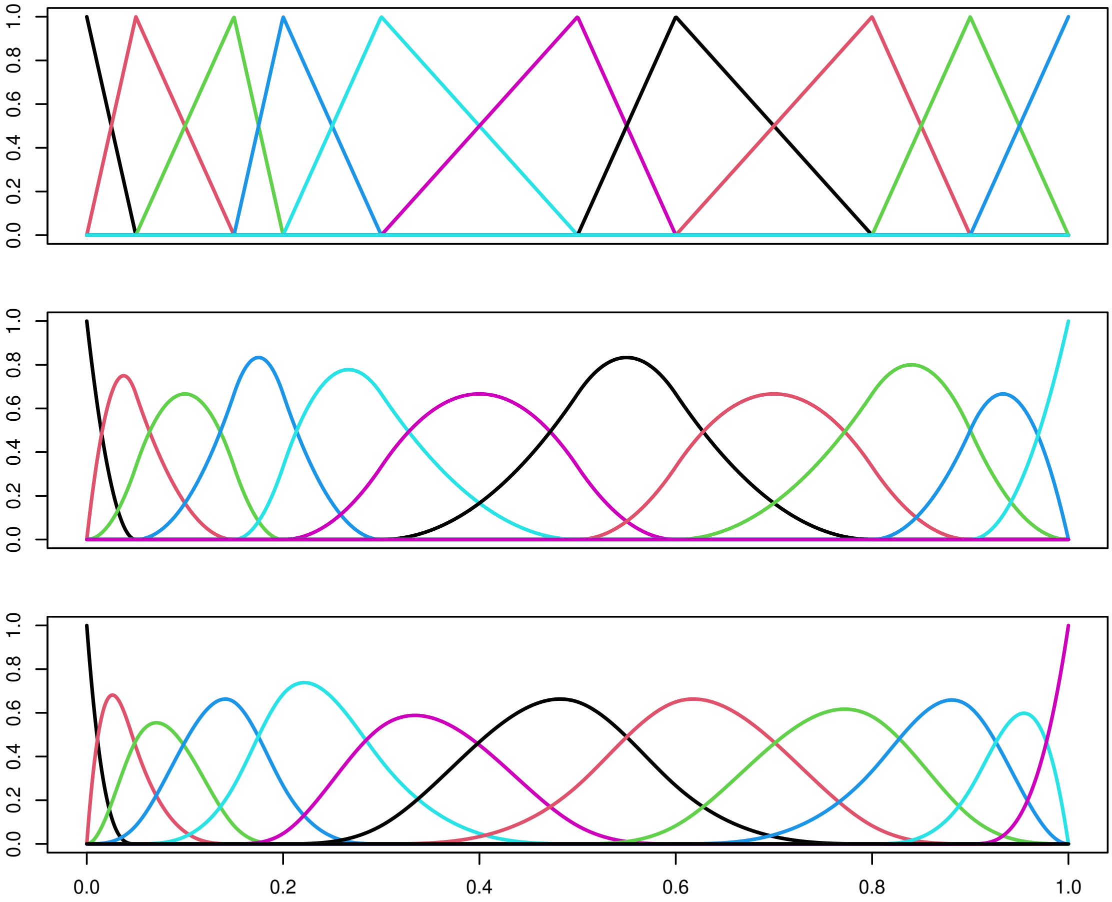


Figure 3.12: Non-uniform B-spline basis functions of degree 1 (linear), 2 (quadratic) and 3 (cubic) with thirteen knots (eleven internal knots and two external knots), respectively.

3.7.5 Choice of number and placement of knots

Knots play a major role in determining the flexibility of curves in splines, particularly in the last three types mentioned above. This is because knots determine the points at which the piecewise polynomial segments of the spline are connected (Perperoglou et al., 2019). Two features of knots can impact this aspect: the number and placement of knots. Both features are fixed and cannot be changed during the model fit. Generally, setting K knots results in fitting $K+1$ polynomial models with a specified degree d (dependent on the spline type), which is associated with the number of parameters used in estimation, referred to as *degrees of freedom* (df). If the number of knots increases, the degree of freedom also increases accordingly, and vice versa. Therefore, differences in the number of knots can lead to different models. Typically, setting too large a number of knots may result in overfitting, while too few knots may not provide enough flexibility to capture the underlying structure of the data. Choosing the number of knots K involves considering two aspects: it should be large enough to provide sufficient degrees of freedom to represent the underlying data and small enough to maintain reasonable computational efficiency. In practice, this is usually done by increasing K until no considerable changes are observed in the plot. There are some recommended strategies for choosing the number of knots K. For instance, if the response variable is a continuous uncensored variable and the sample size is large enough (e.g. $n \geq 100$), $K = 5$ is a reasonable initial choice. Conversely, $K = 3$ is recommended for small sample sizes (e.g. $n < 30$) (Harrell, 2015, Chapter 2, p. 26).

Determining the placement or location of knots becomes straightforward when the relationship between the response (y) and predictor (x) is explicitly known (Harrell, 2015; Jamrozik et al., 2010). For instance, if there are critical points in the data where the curvature or relationship pattern changes, knots should be positioned at those points. However, in real-world problems, the underlying relationship between variables is often unknown or challenging to specify. Therefore, two popular strategies are often used: equidistant knots and quantile knots (Harrell, 2015, Chapter 2, p. 26). The former is the simplest strategy, which uses a set of equally spaced knots. However, this strategy may not effectively capture the underlying structure or variations in the data if the knots do not align well with regions of high data density or where the relationship between the predictor and response variables changes rapidly. The second strategy, quantile knots, offers a solution to this problem. This strategy places the knots according to the quantiles of the predictor x. This ensures the spline more flexible in regions with more data and less flexible in areas with less data.

3.8 Regression splines

Regression splines are another non-linear regression technique, widely used to estimate a smooth conditional function. The idea underlying this method is that a suitable set of p basis functions from Section 3.6 can be utilised to fit the effect of a smooth function, $g(x_i)$, of a covariate x_i, where $i = 1, \ldots, n$, on a response y_i (Hastie & Tibshirani, 1999; Wood, 2017a). Thus, the model can be expressed as

$$y_i = g(x_i) + \epsilon_i,$$

where

$$g(x_i) = \sum_{j=0}^{p} \gamma_j B_j(x) = \mathbf{B}\boldsymbol{\gamma}.$$

and $\epsilon \sim N(0, \sigma^2)$.

Given the p basis functions, $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_p)$ can be estimated using least squares estimation by minimising the least-squares criterion:

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 = ||\mathbf{y} - \mathbf{B}\boldsymbol{\gamma}||^2, \tag{3.39}$$

with respect to $\boldsymbol{\gamma}$. This calculation yields the basis coefficient estimators, $\hat{\boldsymbol{\gamma}} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y}$. Also, values of $\mathbf{y}$ can be fitted by

$$\hat{\mathbf{y}} = \mathbf{B}\hat{\boldsymbol{\beta}} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y} = \mathbf{S}\mathbf{y}, \tag{3.40}$$

where $\mathbf{S}$ is a smoothing matrix, widely known as the *hat* matrix.

The smoothness of the fitted function g depends on the number of bases, p. Increasing this number results in the increased "wiggliness" of g. Additionally, the number of knots (k) affects the number of bases, with the value of p increasing as k increases. In practice, the choice of both numbers is manually and subjectively determined, depending on the specific application. These conditions may render this approach less elegant (Wood, 2017a). However, some systematic methods for selecting these numbers are available in the literature, e.g. stepwise-based methods (Stone et al., 1997) and an artificial immune system (Ülker & Arslan, 2009).

3.9 Penalised regression splines

Penalised regression splines offer an alternative approach to avoid specifying the number of bases or locations and the number of knots. These approaches continue to define the function g by selecting the basis functions as outlined in Section 3.6. Rather that employing these conditions to modulate the smoothness of the fitted function, a penalty term is incorporated into the least-squares objective function.

Recall the least-squares objective function (3.39). This penalty term can be integrated into the function as follows:

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int_{x_{min}}^{x_{max}} (g^{(m)}(x))^2 dx = ||\mathbf{y} - \mathbf{B}\boldsymbol{\gamma}||^2 + \lambda\boldsymbol{\gamma}'\mathbf{D}\boldsymbol{\gamma}, \tag{3.41}$$

where an integral term represents the integrated square of the mth derivative, $\mathbf{D}$ is a symmetric $p \times p$ penalty matrix with an entry $D_{jk} = \int B_j^{(m)}(x)B_k^{(m)}(x)dx$, and λ is a penalty or smoothing parameter. This new objective function (3.41) is referred to as the *penalised least-squares* objective function (Green & Silverman, 1994). The term $\int(g^{(m)}(x))^2dx$ in (3.41) is known as a *roughness penalty*, which is utilised to measure the roughness of the function g, where the mth derivative usually serves as a control parameter for the roughness of function. When m is low (e.g. $m = 1$), the first derivative captures changes in the slope of the function, which can indicate the overall trend or smoothness of the function. As m increases, higher-order derivatives capture finer details of the function's behaviour, including sharp changes or fluctuations. In penalised regression splines, the second derivative ($m = 2$) is a common choice, resulting in the integral term being the integral of the square of the second derivative of the spline function g. This choice is made to control the curvature by measuring the rate of change of the first derivative, which corresponds to the curvature of the function. Penalising such a derivative controls changes in curvature, ensuring that the fitted spline function does not exhibit abrupt changes or oscillations. While higher derivatives (e.g. $m = 3$) can be used, they may introduce additional complexity and computational overhead without necessarily providing considerable improvement in model performance. Therefore, $m = 2$ is applied throughout this thesis due to this consideration. Minimising (3.41) with respect to $\boldsymbol{\gamma}$, given λ, yields the penalised least-squares estimator of $\boldsymbol{\gamma}$:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{B}'\mathbf{B} + \lambda\mathbf{D})^{-1}\mathbf{B}'\mathbf{y}.$$

The fitted values for $\mathbf{y}$ are then expressed by

$$\hat{\mathbf{y}} = \mathbf{B}\hat{\boldsymbol{\gamma}} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D})^{-1}\mathbf{B}'\mathbf{y} = \mathbf{S}\mathbf{y},$$

where $\mathbf{S}$ denotes a smoothing matrix or *hat* matrix.

In this model, the parameter λ assumes a significant role as it controls the smoothness of the fitted function. An increase in λ ($\lambda \rightarrow +\infty$) results in a smoother curve, whereas a decrease in λ ($\lambda \rightarrow 0$) may yield the opposite result. Figure 3.13 illustrates this phenomenon. If $\lambda = 0$, the least-squares criterion (3.41) equates to (3.39), returning the fitted values identical to (3.40).

Choice of smoothing parameter (λ)

The smoothing parameter, λ, can be chosen by the user to control the smoothness of the target function. Visual selection, such as that presented in Figure 3.13, is a simple way to handle this, especially in the case of simple univariate regression settings. In this example of visual selection, it seems that $\lambda = 10^{-4}$, represented by green line, may provide a suitable representation of the underlying signal in the data compared to the other two values of λ.

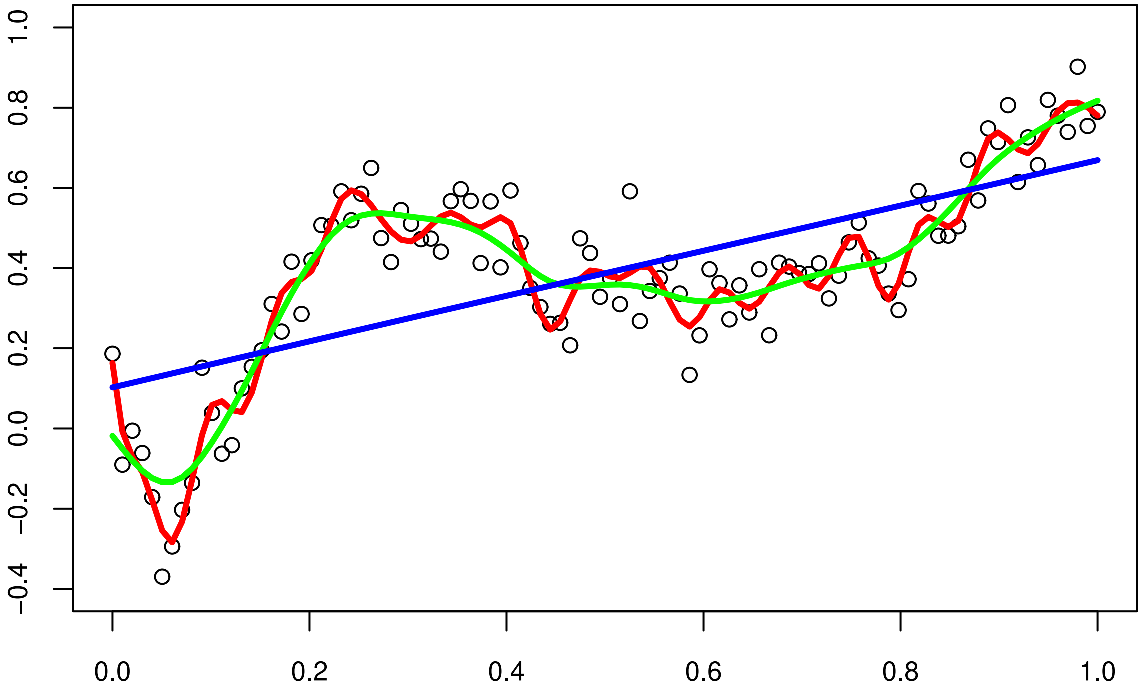


Figure 3.13: Three different smoothing parameter values applied to the data example. Red curve corresponds to $\lambda = 10^{-8}$, green curve to $\lambda = 10^{-4}$ and the blue curve to $\lambda = 10$.

However, such visualisation may not be adequate for selecting an optimal value of smoothing parameters in the context of complex models, such as those composed of smooth functions of multiple variables. The limitations arise in many aspects such as dimensionality and interactions. When the number of predictors increases, it can be challenging to visualise high-dimensional data accurately and effectively. Moreover, interactions between predictors may impact the relationship between those predictors and the target variable. Visual inspection may not capture these interactions comprehensively, leading to suboptimal selection of the smoothing parameter. Instead, a combination of visual exploration, statistical diagnostics, and objective evaluation criteria should be employed to ensure robust model selection and interpretation. There is intensive research in both statistical diagnostics and objective evaluation criteria in the context of selecting the optimal smoothing parameter, known as *cross-validation* and *information criterion*, respectively.

Cross-validation (CV) involves resampling and sample splitting techniques that consider separating data into two sets, usually by proportion, where one is used for testing and the other for training a model in different iterations. The goal of CV is to assess the performance and generalisation ability of a predictive model by ensuring that it performs well on new, unseen data. This is achieved by evaluating its predictive accuracy across different subsets of the dataset. Another method that utilises a different concept to emphasise the trade-off between goodness-of-fit and complexity is the information criterion (IC). Here, three methods are chosen to describe both CV and IC. The first is one of the classical forms of CV, leave-one-out cross-validation (LOOCV). The second is a specific method associated with IC, while the last goes into more detail on IC with three famous criteria such as the Akaike Information Criterion (AIC), the generalised AIC, and the Bayesian Information Criterion (BIC).

(1) Leave-One-Out Cross-Validation (LOOCV): The concept of this method lies in evaluating the smoothing parameter, λ, by minimising the function

$$\mathbf{LOOCV}(\lambda) = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{g}_{(-i)}(x_i;\lambda))^2, \quad i = 1, \ldots, N \tag{3.42}$$

where $\hat{g}_{(-i)}(x_i;\lambda)$ is the predicted value of the omitted observation $(-i)$ (Green & Silverman, 1994, p.30). The procedure can be expressed as follows:

1. Let y_i be data observations with a sample size N.
2. Fit the model $g_{(-i)}(x;\lambda)$ using the remaining $N-1$ observations to minimise the objective function (3.42).
3. Use the fitted model from Step 2 to predict $\hat{g}_{(-i)}(x_i;\lambda)$.
4. Calculate the differences between the observed y_i and its predicted values $\hat{g}_{(-i)}(x_i;\lambda)$.

5. Repeat Step 2 - 4 for each observation or N times.

6. Calculate the average sum of squares of the differences for each feasible value of λ .

7. Choose the λ which provides a minimised value of LOOCV to fit the model.

(2) Generalised cross-validation: The LOOCV method often faces the computational problem of having to fit the model N times. To overcome this problem, Craven and Wahba (1978) proposed a new cross-validation method called generalised cross-validation (GCV). The following is the procedure for this method:

1. Let y_i be data points with a sample size N.

2. Fit the model to minimise the objective function (3.42) with N sample for a range of values of the smoothing parameters, λ.

3. Calculate the GCV value for each λ as follows,

$$\mathbf{GCV}(\lambda) = \frac{n \times \text{RSS}}{(n - \text{tr}(\mathbf{S}))^2},$$

   where RSS $= \sum_{i=1}^{N} (y_i - \hat{g}(x_i; \lambda))^2$, is the residual sum of squares, $\hat{g}$ is the fitted values, and tr($\mathbf{S}$) is the trace of a smoothing matrix $\mathbf{S}$.

4. Choose λ which provides a minimised value of GCV to fit the model.

(3) Information criterion (IC): The IC serves as another tool for selecting the smoothing parameter, λ. This criterion balances the fit of the model to the data against the complexity of the model, aiming to prevent overfitting. The most commonly used information criterion is the Akaike Information Criterion (AIC) (Akaike, 1973): AIC $= 2p - 2\log(L)$, where p presents the number of parameters in the model, and L is the likelihood of the model. However, AIC is sensitive to the small sample sizes, leading to a tendency to overfit due to bias (Hurvich & Tsai, 1989). Therefore, an extension of AIC known as the generalised AIC (GAIC or AICc) was proposed to deal with this issue (Hurvich et al., 1998): GAIC $=$ AIC $+ 2p(p+1)/(N-p-1)$, where N is the sample size. An additional term of $2p(p+1)/(N-p-1)$ is included to adjust for the sample size relative to the number of parameters. Another widely used criterion is the Bayesian Information Criterion (BIC) (Schwarz, 1978): BIC $= \log(N)p - 2\log(L)$. The main difference between AIC and BIC lies in the penalty imposed for complexity: a term of $2p$ for AIC, compared to $\log(N)p$ for BIC, with BIC generally imposing a heavier penalty. Consequently, BIC is often considered more parsimonious than AIC (Burnham & Anderson, 2004). The following is a generalised step-by-step procedure to use information criterion for selecting the smoothing parameter in models:

1. Let the data have a sample size N.
2. Fit the model to the data for a range of candidate smoothing parameters, λ.
3. Calculate the IC value for each λ.
4. Choose λ which provides a minimised value of IC to fit the model.

3.10 Effective degrees of freedom

In a general linear model, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H}$ represents the idempotent projection matrix or hat matrix. This hat matrix is calculated as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, where $\mathbf{X}$ is the design matrix containing the predictors. The trace of the hat matrix, $\mathrm{tr}(\mathbf{H})$, equals the number of fitted parameters or coefficients (denoted as p) in the model. This number is commonly known as the *effective degrees of freedom* (*edf*) and is usually expressed as

$$\mathrm{df}_{\mathrm{model}} = \mathrm{tr}(\mathbf{H}) = p.$$

Essentially, it is used to measure the complexity of the model. A higher *edf* indicates a more complex model with more flexibility to fit the data, while a lower *edf* indicates a simpler model with fewer parameters.

In analogy to this idea in linear regression, the hat matrix in the context of regression splines can be defined as $\mathbf{S} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$. It appears to have the same form as the hat matrix $\mathbf{H}$ in the general linear model as mentioned earlier. In this case, $\mathrm{tr}(\mathbf{S})$ would represent the number of basis functions used in the spline model (Hastie et al., 2009). Therefore, in practice, the *edf* can be used to guide the selection of smoother in addition to or instead of specifying knots directly (Perperoglou et al., 2019). As the number of knots increases, the *edf* also tends to increase because the spline has more flexibility to fit the data. Conversely, reducing the number of knots decreases the flexibility of the spline and thus decreases its *edf*.

However, in penalised regression splines, the smoother matrix is a common term used instead of the hat matrix. It is defined as $\mathbf{S}_\lambda = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D})^{-1}\mathbf{B}'$ and may not directly correspond to the number of basis functions due to the presence of penalty terms. In this case, the *edf* of the model, denoted as

$$\mathrm{df}_{\mathrm{model}} = \mathrm{tr}(\mathbf{S}_\lambda) = \mathrm{tr}[\mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D})^{-1}\mathbf{B}'] = \mathrm{tr}[\mathbf{B}'\mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D})^{-1}],$$

still represent a measure of the model complexity after penalisation. It considers the trade-off between goodness of fit and model simplicity imposed by the penalty parameter λ and penalty structure $\mathbf{D}$ (Hastie et al., 2009).

3.11 P-splines

Eilers and Marx (1996) proposed another variant type of penalised regression splines, known as P-splines. The distinction between this method and the prior one lies in the specification of basis functions and a penalty term. P-splines form their basis functions from uniform B-splines on equidistant knots and penalise the sum-of-squared order-m difference of neighbourhoods of B-spline basis coefficients. The simplest form of the penalty is the sum-of-squared first-order $(m = 1)$ differences,

$$\boldsymbol{\gamma}'\mathbf{D}^{(1)}\boldsymbol{\gamma} = \sum_{j=1}^{p-1}(\gamma_{j+1} - \gamma_j)^2,$$

where $\mathbf{D}^{(1)}$ is a first-ordered difference matrix, which is defined as

$$\mathbf{D}^{(1)} = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

However, any order difference can be used. For example, the squared second-order $(m = 2)$ difference is given by

$$\boldsymbol{\gamma}'\mathbf{D}^{(2)}\boldsymbol{\gamma} = \sum_{j=1}^{p-2}(\gamma_{j+2} - 2\gamma_{j+1} + \gamma_j)^2,$$

where

$$\mathbf{D}^{(2)} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{pmatrix},$$

where $\mathbf{D}^{(2)}$ is a second-order difference matrix. Similarly to the penalised regression splines in the previous section, the roughness of the fitted function can be controlled by multiplying it by a scalar quantity λ. Therefore, the penalised least squares objective function for P-splines is

$$||\mathbf{y} - \mathbf{B}\boldsymbol{\gamma}||^2 + \lambda\boldsymbol{\gamma}'\mathbf{D}^{(m)\prime}\mathbf{D}^{(m)}\boldsymbol{\gamma}. \tag{3.43}$$

Minimising the objective function (3.43) with respect to $\boldsymbol{\gamma}$ for a fixed value of λ yields the

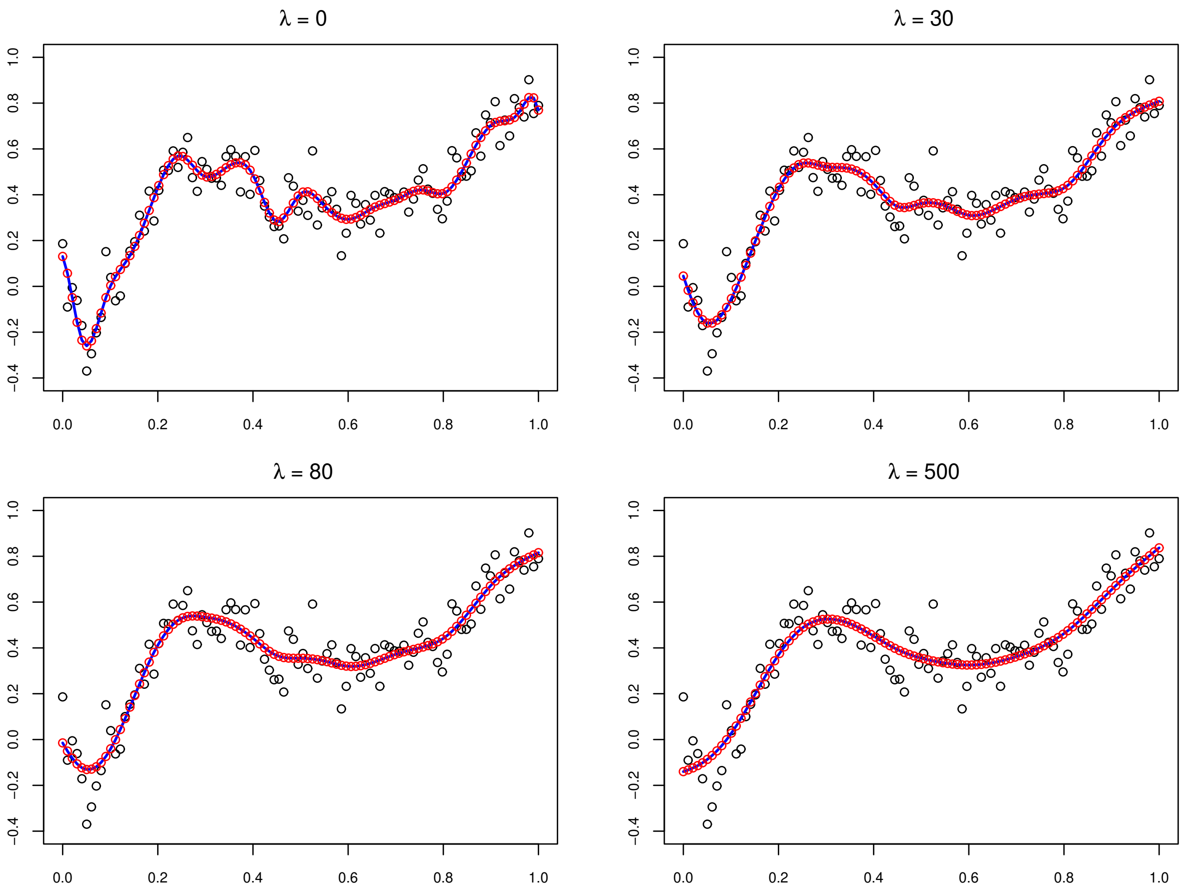


Figure 3.14: P-splines when varying the smoothing parameter, λ. Note that when λ is 0, the fitted curves is unpenalised which is equivalent to the regression splines.

estimate of $\boldsymbol{\gamma}$,

$$\hat{\gamma} = (\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}^{(m)\prime}\mathbf{D}^{(m)})^{-1}\mathbf{B}'\mathbf{y}.$$

The behaviour of the fitted function continues to rely on the smoothing parameter, λ (see Figure 3.14). Selection methods such as CV, GCV or IC can be employed to determine this parameter.

3.12 P-splines as mixed-effects models

In theory, any spline model can be rewritten as a mixed-effects model, and then the model can be fitted using the methodology developed in that context. P-splines are also capable of being transformed in this way (Currie & Durban, 2002; Eilers, 1999).

Let $\mathbf{B}$ be a set of B-spline basis functions, and $\boldsymbol{\gamma}$ be the corresponding coefficients. Thus, the model of response $\mathbf{y}$ based on B-splines can be expressed as:

$$\mathbf{y} = \mathbf{B}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{3.44}$$

where

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}).$$

Generally, the corresponding coefficients of B-spline basis functions can be estimated by minimising the penalised least squares objective function (3.43), as stated in Section 3.8. Alternatively, the model (3.44) can be rewritten, aligning the spline model with a mixed-effects model.

Let

$$\mathbf{I} - \mathbf{D}'(\mathbf{D}\mathbf{D}')^{-1}\mathbf{D} = \mathbf{L}\mathbf{L}',$$

where $\mathbf{D} = \mathbf{D}^{(m)}$ and the $p \times m$ matrix $\mathbf{L}$ has full column rank. According to this specification, the model (3.44) can be reformulated as mixed-effects models,

$$\begin{aligned} \mathbf{y} &= \mathbf{B}\mathbf{L}\mathbf{L}'\boldsymbol{\gamma} + \mathbf{B}\mathbf{D}'(\mathbf{D}\mathbf{D}')^{-1}\mathbf{D}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \end{aligned}$$

where $\mathbf{X} = \mathbf{B}\mathbf{L}\mathbf{A} = [\mathbf{1}, \mathbf{x}, \ldots, \mathbf{x}^{d-1}], \boldsymbol{\beta} = \mathbf{A}^{-1}\mathbf{L}'\boldsymbol{\gamma}, \mathbf{Z} = \mathbf{B}\mathbf{D}'(\mathbf{D}\mathbf{D}')^{-1}, \mathbf{u} = \mathbf{D}\boldsymbol{\gamma}$, and $\mathbf{A}$ is an existing full rank $m \times m$ matrix. Assuming that $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$ where $\mathbf{G} = \sigma_\gamma^2 \mathbf{I}$, the penalty term in the penalised least squares function for P-splines is equal to $\phi^{-1}\mathbf{u}'\mathbf{u}$. Thus, $\lambda = \phi^{-1} = \sigma_\epsilon^2 / \sigma_\gamma^2$. As a result, systematic techniques (e.g. CV, GCV, IC, etc.) are not required to select the smoothing parameter. In this case, all parameters in the model can be estimated using maximum likelihood estimation, such as restricted maximum likelihood estimation,

$$l_R(\mathbf{V}) = -\frac{1}{2}\Big\{\log|\mathbf{V}| + \log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}\Big\},$$

where $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \sigma_\epsilon^2\mathbf{I}$. The maximisation of log-likelihood function yields,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

and

$$\hat{\mathbf{u}} = \boldsymbol{\Sigma}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

3.13 Chapter summary

In this chapter, the statistical background and its associated methodologies are presented. The initial section elucidates why mixed-effects models have become a pivotal and increasingly popular approach for analysing longitudinal child growth data. The inclusion of random effects in the regression model stands as a principal methodology, offering the advantage of accounting for data dependency as well as between-individual variation.

In response to this advantage, a variety of mixed-effects models have been proposed to address additional complexities, such as modelling non-linear relationships, while maintaining adherence to the core principles of mixed-effects modelling. The second section outlines flexible mixed-effects models that extend beyond the previously mentioned random effects; these models also incorporate bases constructed from spline techniques as random effects to form smooth functions. These approaches enhance flexibility and facilitate implementation in standard mixed-models software, such as R. Consequently, one such approach, outlined in Chapter 4, has been applied using this methodology.

In the third section, another common approach used to model longitudinal child growth data for monitoring child growth, known as correlation models, is reviewed. This type of model is particularly helpful in informing a health professional's judgment about a child's growth based on predicting current growth using previous growth outcomes. To obtain this prediction, the correlation between growth outcomes at different time points is key, as it reflects the influences between those growth outcomes. However, these models may not be suitable for characterising all aspects of longitudinal child growth data, as they only condition on previous outcomes, lack consideration of other covariates, and fail to account for variability among children. Therefore, this kind of model is not considered in this thesis.

In the fourth section, a concise summary of quantile regression (QR) is presented. This method offers a significant advantage by providing a comprehensive description of the response variable conditional upon the covariates. Included in the section is a definition of QR, an explanation of quantiles for a random variable, a discussion on the concept of linear QR model, a brief review of the properties of the quantile function, and an overview of two types of QR: the distribution-free method and the likelihood-based method). The latter will be the core method employed in Chapter 4 and 5 of this thesis. Following this, the next section provides a brief summary of the Bayesian approach in QR. Two existing Bayesian linear QR model are discussed: one based on the asymmetric Laplace (AL) distribution and the other on the location-scale mixture of normals representation of the AL distribution. This thesis, especially in Chapter 5, will focus on the latter approach, which offers a more convenient methodology for the MCMC method, such as Gibbs sampling.

Subsequently, the existence of linear QR models within the context of longitudinal data is explored. This forms a foundational concept in the analysis of longitudinal child growth data, and is instrumental in extending some ideas towards more flexible QR approaches. The review initiates a discussion on two distinct linear QR models. The first is the classical QR model, which accounts for individual pure location shift effects, severing a role similar to individual random effects. The second model is an expanded version of the first, designed to facilitate the analysis of large datasets. Additionally, the last model discussed is the linear quantile mixed-effects models, which are based on the AL distribution. The last six sections present methodologies related to splines, which are frequently utilised to construct smooth functions that capture the non-linear relationships between responses and covariates. Some of these methodologies will be employed to model non-linear growth trajectories in Chapters 4 to 6.

Chapter 4

Flexible quantile regression models for longitudinal child growth data

4.1 Introduction

Longitudinal child growth data (LCGD) encompass data types that holistically reflect child growth developments since growth measurements are collected repeatedly from the same children over time (Cole, 1994). As a result, these data can account for dynamic effects. Such data can exhibit various characteristics, including diverse non-linear growth patterns and correlations in repeated (or clustered) data. Therefore, LCGD require appropriate statistical approaches to address these intricacies. Certainly, modelling LCGD via correlation models, as mentioned in Section 3.3, has been previously utilised for this purpose. Although correlation models are suitable for analysing the growth patterns and relationships within LCGD over time and estimate a child's growth to monitor or track their health, these models lack the ability to evaluate potential risk factors that can impact child growth, such as parental factors, nutritional deficiencies, health conditions, environmental factors, and social and economic factors. While many other statistical methods have been proposed for this purpose, most rely on conditional mean models, such as the generalised estimating equation (Hardin & Hilbe, 2013; M. Wang, 2014), the random-effects model or linear mixed model (Fitzmaurice et al., 2011; Laird & Ware, 1982), non-linear mixed models (Beath, 2007), piecewise models (Grajeda et al., 2016), semi-parametric models (Durbán et al., 2005), and additive mixed models (Wood, 2017a)). However, these models may fall short in providing a comprehensive view of the distribution of growth measurements. They may be unsuitable for analysing child growth data when the objective is to interpret changes in growth measurements by associating risk factors at specific locations (e.g. lower or upper) within the distribution of growth measurement. This type of interpretation is termed the "quantile treatment effect" (Koenker, 2005).

In this chapter, I review two flexible quantile regression (QR) approaches aimed at achieving the aforementioned goal. The first approach is the classical marginal quantile regression model proposed by Wei et al. (2006). In this approach, the authors utilised B-splines, to enhance the flexibility of QR, enabling it to describe non-linear growth patterns. They also incorporated the AR(1) model to characterise growth history. Furthermore, the incorporation of an AR(1) structure within the model addresses the within-subject variability. The primary aim of this model is to estimate a set of quantile curves, which can subsequently represent reference growth curves.

The second approach is the additive quantile mixed model (AQMM) introduced by Geraci (2019). AQMM emphasises the advantage of additive modelling in dealing with a variety of relationship patterns between response and covariates. It also provides the capability to distinguish between-subject and within-subject variations using the mixed model framework. This model considers the correlation between growth observations from the same child by integrating random effects (unobserved variables) into the additive quantile model, which already encompasses fixed effects (observed variables). Random effects possess two main abilities to address this aspect. Firstly, they can accommodate subject-specific variability, which is common in LCGD, for instance, the repeated growth measurements taken from the same children over time. These repeated measurements are likely to be more similar to each other than to measurements taken from different children. Random effects account for this by capturing the subject-specific patterns over time, such as a random intercepts and slopes, allowing each subject to have its own intercept and slope. Secondly, by including random effects, the model can explicitly address the within-subject correlation. Specifically, in LCGD, growth measurements taken closer in time might exhibit higher correlated compared to those taken further apart, akin to a first-order autocorrelation or AR(1). In this regard, random effects introduce dependencies among repeated measures within the same child, allowing the residuals (errors) from the same child to be treated as correlated rather than independent. This accounts for the fact that measurements taken closer together in time are more similar (more correlated) than those taken further apart, implicitly capturing the essence of the AR(1) structure. To clarify, consider a simple linear mixed model with random intercepts (u_{0i}) and slopes (u_{1i}):

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0i} + u_{1i} x_{ij} + \epsilon_{ij},$$

where $u_{0i} \sim N(0, \sigma^2_{u0})$, $u_{1i} \sim N(0, \sigma^2_{u1})$ and $\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon})$. Note that in this case, u_{0i} and u_{1i} are not correlated. This results in the covariance between two observations within the same subject i at times t and t' as:

$$\text{Cov}(y_{it}, y_{it'}) = \sigma^2_{u0} + \sigma^2_{u1} x_{it} x_{it'}.$$

The above covariance allows for a more complex correlation structure that can vary with time, resembling the AR(1) structure to some extent. The correlation structure is not limited solely to AR(1) but can take different forms, e.g. compound symmetry (CS), Toeplitz (TOEP), and unstructured forms (UN), making this type of model more flexible than others such as correlation models. Furthermore, AQMM provides insights into both population-level and subject-level effects, courtesy of the mixed model framework. As a result, AQMM appears to be a promising tool for analysing or modelling LCGD. However, to our updated knowledge, it is worth noting that AQMM has not yet been applied to this type of data.

This chapter comprises five sections, including the introduction. Sections 4.2 and 4.3 introduce two existing flexible QR models for LCGD. In Section 4.4, several simulation studies are presented: the first examines the performance of these models under typical LCGD characteristics, and the second looks at data that includes additional independent variables potentially influencing growth response. Another study explores the performance of the AQMM model in scenarios with between-individual differences. The fourth focuses on data highlighting between-individual differences in intra-individual variation. The final simulation study delves into a distinct LCGD feature: between-individual differences in autocorrelation. Section 4.5 concludes with a chapter summary.

4.2 Quantile autoregressive model

This section presents an expanded version of the classical quantile autoregression model for longitudinal child growth data as proposed by Wei et al. (2006), who termed this approach the "quantile specific autoregressive model". Henceforth, this model will be abbreviated as QSAM. QSAM relies on a combination of classical quantile regression, regression splines, and the first-order autoregressive model, AR(1). This model facilitates the use of various splines (e.g. B-splines, truncated polynomial splines) to capture the non-linear relationship between growth measurement and age at specific quantiles. In their original work, Wei et al. (2006) advocated for the use of B-splines due to their numerous benefits in the context of child growth modelling. These include high flexibility for modelling nonlinear growth trajectories, such as they can be easily adjusted to fit a diverse range of shapes and patterns. Local adjustments to the model can be performed, influencing only the pertinent section of the curve without altering the entire growth profile. Moreover, B-splines are numerically stable, ensuring that minor changes in the data do not result in large alterations to the growth curve. The autoregressive component of QSAM addresses the challenge of unequally spaced measurements in longitudinal child growth data. As previously noted, QSAM functions as a marginal model and thus primarily provides insights into population-level effects. Consequently, it is suited for the construction of reference growth charts.

Consider y_{ij} where $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$ to be the growth response at the jth time point for the ith child. The conditional quantile function of y_{ij} at τ, given $t_{ij}, y_i(t_{i,j-1})$, and $\mathbf{x}_i$, is given by:

$$\mathbf{Q}_{y_i(t_{ij})|t_{ij},y_i(t_{i,j-1}),\mathbf{x}_i}(\tau) = g_\tau(t_{ij}) + [\psi_{1,\tau} + \psi_{2,\tau}(t_{ij} - t_{i,j-1})]y_i(t_{i,j-1}) + \mathbf{x}_i'\boldsymbol{\beta}_\tau. \tag{4.1}$$

The model given by (4.1) comprises three distinct components: a non-parametric function, g_τ, associated with the time variable, t_{ij}; a first-order autoregressive model, AR(1); and a linear predictor function associated with the covariates vector, $\mathbf{x}_i$. The function g_τ can be represented using any spline method. If we assume that the unknown function g_τ is non-parametric and solely associated with t_{ij}, this function can be expressed in a general spline model as:

$$g_\tau(t_{ij}) \approx \sum_{k=0}^{\kappa} \gamma_{\tau,k} B_k(t_{ij}). \tag{4.2}$$

Here, $B_k(t_{ij})$ denotes a set of basis functions, and $\gamma_{\tau,k}$ represents the corresponding spline coefficients at a specific τ. As discussed in Section 3.7, the function (4.2) can be defined in multiple ways. For this study, B-splines were selected, in accordance with the recommendations of Wei et al. (2006). A pivotal question that arises is: “how many basis functions should we use to construct the smooth function g_τ?” The answer depends on the selection of the number and positions of knots, resulting in several models that should be considered.

All parameters in the model, as presented in (4.1), can be estimated by minimising the objective function

$$\rho_\tau\big(y_i(t_{ij}) - g(t_{ij}) - [\psi_1 + \psi_2(t_{ij} - t_{i,j-1})]y_i(t_{i,j-1}) - \mathbf{x}_i'\boldsymbol{\beta}\big), \tag{4.3}$$

where $\rho_\tau(\mathbf{e}) = \sum_{j=1}^{n} e_j(\tau - I(e_j < 0))$. The optimisation problem, given by (4.3), can be solved using a linear programming method, as outlined in Section 3.4.5.

4.3 Additive quantile mixed model

This section presents a flexible quantile regression model designed for longitudinal data. This model represents an extension of the additive quantile model, enabling the use of various non-parametric functions to accommodate diverse relationship patterns between growth measurements and covariates. Furthermore, it incorporates the mixed model framework to address the dependency between observations taken from the same subject at different times. Termed the “additive quantile mixed model” (Geraci, 2019) or AQMM for short, this approach parallels the additive mixed model in the context of the mean model. Notably, the AQMM yields insights into both population-level and subject-level effects, consistent with the capabilities of the mixed model framework.

Let x_{ij} denote the explanatory variables, $\mathbf{z}'_{ij}$ be the jth row of a random effects design matrix $\mathbf{Z}_i$, and $\mathbf{u}_{\tau,i}$ represent an individual-specific random effects vector. This vector comprises coefficients corresponding to each specific random effect and its respective individuals for a particular quantile level τ. The AQMM is expressed as

$$y_{ij}|\mathbf{u}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij} = \beta_{\tau,0} + \sum_{r=1}^{p} g_\tau^{(r)}(x_{ijr}) + \mathbf{z}'_{ij}\mathbf{u}_{\tau,i} + \epsilon_{ij}, \tag{4.4}$$

where $\mathrm{P}(\epsilon_{ij} \leq 0|\mathbf{u}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = \tau$ and $\epsilon_{ij} \sim \mathrm{AL}(0, \sigma_\tau, \tau)$. Alternatively, the model (4.4) can be rewritten as the τ-th quantile regression function of y as:

$$Q_{y_{ij}|\mathbf{u}_i,\mathbf{x}_{ij},\mathbf{z}_{ij}}(\tau) = \beta_{\tau,0} + \sum_{r=1}^{p} g_\tau^{(r)}(x_{ijr}) + \mathbf{z}'_{ij}\mathbf{u}_{\tau,i}. \tag{4.5}$$

Here $Q(\tau)$ is the τth conditional quantile function of the growth response based on the random effects $\mathbf{u}_i$ and the explanatory variables $\mathbf{x}_{ij}$. The functions $g_\tau^{(r)}$ on the right-hand side are unknown functions (typically smooth) of the explanatory variables x_r, and can be either linear or non-linear. For instance, partitioning the explanatory variables as $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)'$ with $\mathbf{x}_1 = (x_1, \ldots, x_s)$ for non-linear functions and $\mathbf{x}_2 = (x_{s+1}, \ldots, x_p)$ for linear ones, the summation term in either (4.4) or (4.5) can be elaborated as:

$$\sum_{r=1}^{s}\sum_{k=1}^{\kappa_r} \gamma_{\tau,kr} B_k^{(r)}(x_{ijr}) + \sum_{r=s+1}^{p} \beta_{\tau,r} x_{ijr},$$

where B_k are a set of basis functions and $\gamma_{\tau,k}$ are the corresponding coefficients. Commonly, these basis functions can be defined in several ways as detailed in Section 3.7. P-splines were selected for their utilisation of B-spline bases, often defined on uniformly spaced knots. They enable the use of a substantial number of B-splines (or knot count) and omit B-splines with no data support through a penalty term. Consequently, the specification of the number and positioning of knots is not a concern.

The model (4.5) can be expressed in matrix form as:

$$\mathbf{Q}_{\mathbf{y}_i|\mathbf{u}_i,\mathbf{X}_i,\mathbf{Z}_i}(\tau) = \mathbf{X}_i\boldsymbol{\beta}_\tau + \mathbf{B}_i\boldsymbol{\gamma}_\tau + \mathbf{Z}_i\mathbf{u}_{\tau,i}, \tag{4.6}$$

where $\mathbf{X}_i$ is a $n_i \times (p - s + 1)$ matrix, with the first column representing the population intercept and the subsequent columns representing the explanatory variables, often referred to as fixed effects. The vector $\boldsymbol{\beta}_\tau$ contains the corresponding regression coefficients associated these fixed effects. $\mathbf{B}_i$ is a $n_i \times \kappa$ matrix that encompasses the bases for the s covariates, and $\boldsymbol{\gamma}_\tau$ is a vector of the coefficients corresponding to those bases. $\mathbf{Z}_i$ is a $n_i \times q$ matrix, and $\mathbf{u}_{\tau,i}$ is a vector of the random-effect coefficients associated with $\mathbf{Z}_i$.

To apply the AQMM to LCGD, the model given by (4.5) or (4.6) can be reformulated into a simple growth model as follows:

$$Q_{y_{ij}|\mathbf{u}_i,\mathbf{t}_{ij},\mathbf{z}_{ij}}(\tau) = \beta_{\tau,0} + g_\tau(t_{ij}) + u_{\tau,0i} + u_{\tau,1i}t_{ij}. \tag{4.7}$$

For the estimation of the τ-th quantile regression function, both $\boldsymbol{u}_{\tau,i}$ and $\boldsymbol{\gamma}_\tau$ must adhere to the following assumptions:

1. $\mathbf{u}_{\tau,i} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{G}_\tau)$ and $\boldsymbol{\gamma}_\tau \sim \mathcal{N}_s(\mathbf{0}, \boldsymbol{\Phi}_\tau = \oplus_{r=1}^{s} \phi_{\tau,r}\mathbf{I}_{\kappa_r})$, respectively. Note that in the context of the simple growth model (4.7), these assumptions modify to $\mathbf{u}_{\tau,i} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{G}_\tau)$ and $\boldsymbol{\gamma}_\tau \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi}_\tau = \phi_\tau \mathbf{I}_\kappa)$,

2. $\mathbf{u}_{\tau,i}$ is assumed to be independent for each child i, and from $\boldsymbol{\gamma}_\tau$.

By incorporating the aforementioned assumptions and utilising L_2 penalised splines, the objective function can be derived as:

$$\sum_{i=1}^{N} \rho_\tau(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_\tau - \mathbf{B}_i\boldsymbol{\gamma}_\tau - \mathbf{Z}_i\mathbf{u}_{\tau,i}) + \sum_{i=1}^{N} ||\mathbf{u}_{\tau,i}||^2_{\mathbf{G}_\tau^{-1}} + \sum_{r=1}^{s} \phi_{\tau,r}^{-1} ||\boldsymbol{\gamma}_{\tau,r}||^2, \tag{4.8}$$

where $\rho_\tau(\mathbf{e}) = \sum_{j=1}^{n} e_j(\tau - I(e_j < 0))$ and $\phi_{\tau,r}$ are smoothing parameters. Consequently, all parameters in model (4.7) can be estimated using this objective function. Nonetheless, minimising the objective function (4.8) equates to maximising the likelihood function based on the asymmetric Laplace (AL) distribution, as posited by Geraci and Bottai (2007), Geraci (2019), Geraci and Bottai (2014), and Yu and Moyeed (2001). For this, let us assume that each $\mathbf{y}_i$, given $\boldsymbol{u}_i$, adheres to the AL distribution with location, scale, and skewness parameters denoted by $\boldsymbol{\mu}_{\tau,i} = \mathbf{X}_i\boldsymbol{\beta}_\tau - \mathbf{B}_i\boldsymbol{\gamma}_\tau - \mathbf{Z}_i\mathbf{u}_{\tau,i}$, σ_τ and τ, respectively. As such, the likelihood function can be formed as

$$L(\boldsymbol{\beta}_\tau, \sigma_\tau, \mathbf{G}_\tau, \boldsymbol{\Phi}_\tau | \mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i) = \mathcal{AL}(\mathbf{X}_i\boldsymbol{\beta}_\tau - \mathbf{B}_i\boldsymbol{\gamma}_\tau - \mathbf{Z}_i\mathbf{u}_{\tau,i}, \sigma_\tau, \tau).$$

To ascertain the standard errors associated with each parameter, the block bootstrap is utilised in this regard (Geraci, 2019).

4.4 Simulation studies

As discussed in Section 4.1, AQMM has not yet been applied to model LCGD. Therefore, it is essential to ascertain whether this model can effectively accommodate all the unique characteristics and features of such data. An inherent trait commonly found in LCGD is the presence of autocorrelation among repeated observations. Theoretically, the random effect component in the AQMM is posited to capture this trait, eliminating the need for an autoregressive residual model (Geraci, 2019). Furthermore, it is not limited to just this trait; for instance, children might show variations in their initial growth at birth and in growth rates. This suggests that the data possesses between-individual differences. Consequently, a pressing question emerges: Is the AQMM adequetely suited for addressing LCGD?

In addition, the choice of spline model to capture non-linear growth patterns requires meticulous consideration. Among the available options, P-splines appear to be a sensible choice to achieve this objective. These splines are constructed from B-splines combined with a discrete roughness penalty on their coefficients. This method offers several advantages, particularly in eliminating the need to specify the number and placement of knots (Eilers & Marx, 2021). However, the literature also documents an alternative spline method similar to P-splines. This alternative employs B-splines with a derivative-based penalty on the basis coefficients (Wood, 2017b), providing another option for exploration.

Therefore, when dealing with LCGD data, it is necessary to consider both the effectiveness of AQMM and the appropriateness of various spline modelling strategies.

4.4.1 Study 4.1

In AQMM, autocorrelation among repeated measurements for the same individual is addressed through the inclusion of random effects, denoted as $\mathbf{u}_{\tau,i}$. These effects are assumed to follow zero-centred multivariate normal distributions, with their variance-covariance matrices represented by $\mathbf{G}_{\tau,i}$. Consequently, these matrices play a crucial role in accounting for autocorrelation. However, an alternative approach to address this correlation involves incorporating an autoregressive model into the regression framework. The AR(1) model is often employed in this context, given that within-subject correlations typically decrease over time, reflecting the growth mechanism of a child. This methodology is indeed adopted by the QSAM approach.

Aim

A simulation study was conducted to evaluate the accuracy of AQMM and QSAM in estimating conditional quantile functions for child growth measurements using longitudinal data. Additionally, the study investigated the behaviour of splines in modelling nonlinear growth patterns.

Data generation

Longitudinal child growth data were generated by determining a growth outcome as the standard score of weight (WAZ). Each child had repeated growth outcomes over a specified interval of time. To ensure that the simulated data reflected the natural progression of child growth development, repeated growth outcomes were generated to correlate with one another.

The data were generated using the model described below. For simplification, only one covariate, the time variate, is included:

$$y_{ij} = \mu_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, N, j = 1, \ldots, n_i, \tag{4.9}$$

where

$$\begin{aligned} \mu_{ij} = \; & 0.06 - 13.65t_{ij} + 209.83t_{ij}^2 - 1067.35t_{ij}^3 + \\ & 2634.10t_{ij}^4 - 3446.05t_{ij}^5 + 2301.40t_{ij}^6 - 617.65t_{ij}^7. \end{aligned} \tag{4.10}$$

Here, μ_{ij} represents the mean function, designed to follow the mean weight-for-age Z score for 4563 children from the “Growing up in Scotland” study, as per the UK-WHO child growth charts (refer to Figure 4.1 and Chapter 2 for further details about this dataset). The time variable, t_{ij}, was generated within the range [0,1] and adhered to two data designs: balanced and unbalanced data. Note that the mean function (4.10) encompasses two periods of growth: childhood (0-9 years) and adolescence (10-14 years). Typically, these two periods exhibit distinct growth patterns, which also vary between males and females. However, in this simulation study and subsequent ones, this aspect has been ignored. This is due to the limitation of the example dataset, the GUS data, which includes only eight repeated physical growth measurements. Furthermore, the primary goal of the simulation study is to investigate the performance of target models (AQMM and QSAM) in modelling general longitudinal child growth data without specifying a particular gender. Therefore, the function (4.10) is utilised as an example of growth pattern to generate longitudinal child growth data.

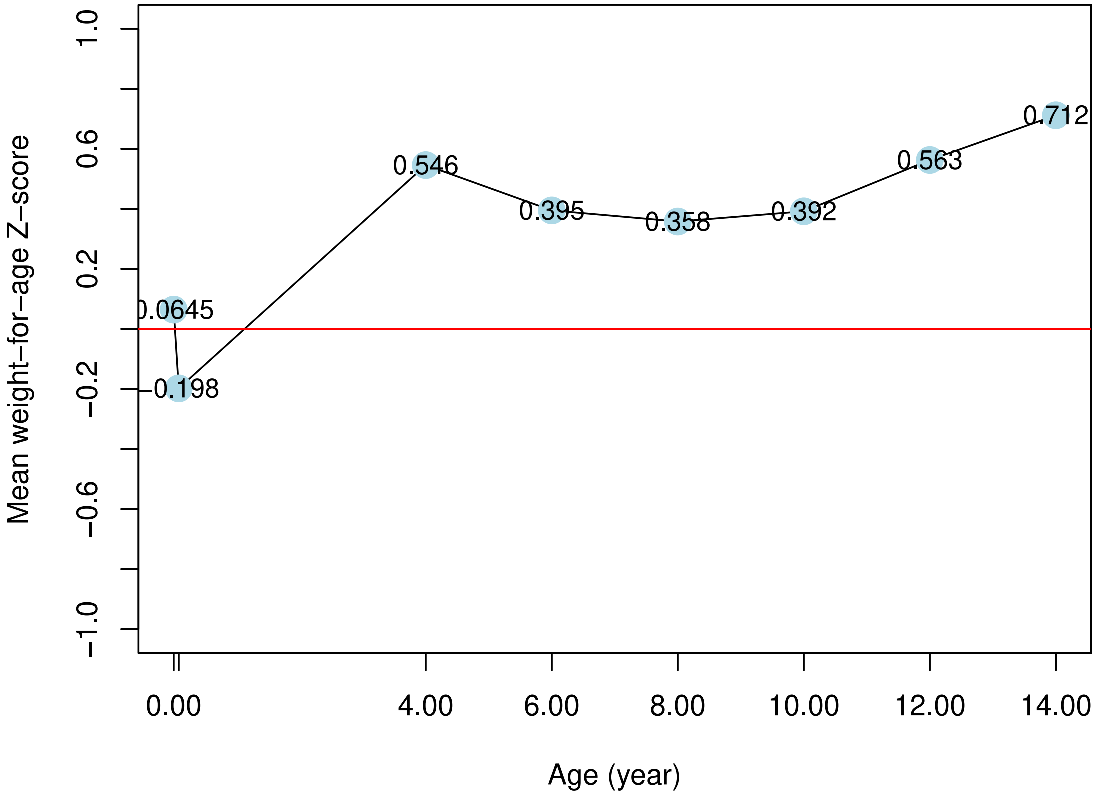


Figure 4.1: Mean weight-for-age Z-scores for 5210 Growing up in Scotland (GUS) children according to the UK-WHO child growth charts.

Balanced data

In this scenario, each child had eleven repeated measurements denoted as n_i. To ensure the simulated data reflected the real LCGD data, time observations for each child were assigned at the following scheduled time points: 0, 0.06, 0.14, 0.21, 0.29, 0.36, 0.43, 0.57, 0.71, 0.86, 1. These settings represent a transformation of a child's age in the "Growing up in Scotland" study to fit a range between 0 to 1. Notably, the intervals between the first seven time points were approximately 0.07, while the remaining intervals were 0.14. As a result, these represent unequally spaced time observations.

Unbalanced data

The number of repeated measurements (n_i) was set to range between 2 and 11, determined by rounding uniformly random numbers from the uniform distribution $\mathcal{U}(2, 11)$, generated using the function `runif(N,2,11)` in R, where `N` represents the number of children. Every child had a growth outcome recorded at baseline, which corresponds to birth or $t_{i1} = 0$. To generate distinct sets of observation times deviating from scheduled time points for each child, starting from the second scheduled time point, the observation times were drawn from uniform distributions $\mathcal{U}(a, b)$ with the following lower and upper limits: [0.057,0.063], [0.136,0.144], [0.205,0.215], [0.284,0.215], [0.284,0.296], [0.353,0.367], [0.42,0.44], [0.55,0.59], [0.67,0.75], [0.81,0.91] and [0.88,1], respectively. Using this method ensured that the simulated observation times slightly deviated from the scheduled time points of the balanced data, with the exception of the last time point.

To ensure that the simulated data exhibited (auto)correlated growth outcomes, the residual errors (ϵ_{ij}) for each child were generated according to two different scenarios of variance-covariance structure.

Homogeneous exponential covariance

In this scenario, the residual (within-subject) errors were assumed to be both homoscedastic (i.e. possessing equal variances) and autocorrelated. The residual errors for each child were generated as:

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i), \quad \mathbf{R}_i = \sigma^2 \mathbf{C}_i, \tag{4.11}$$

where

$$\mathbf{C}_i = \begin{pmatrix} 1 & e^{-s_{12}/\phi} & e^{-s_{13}/\phi} & \cdots & e^{-s_{1n_i}/\phi} \\ & 1 & e^{-s_{12}/\phi} & \cdots & e^{-s_{1(n_i-1)}/\phi} \\ & & 1 & \cdots & e^{-s_{1(n_i-2)}/\phi} \\ & & & \ddots & \vdots \\ & & & & 1 \end{pmatrix}. \tag{4.12}$$

Here, $\mathbf{C}_i$ represents a correlation matrix derived from an exponential correlation structure. σ^2 denotes the residual variance, while $s_{jj'}$ is a real number representing the distance between two time points t_j and $t_{j'}$, defined as $s_{jj'} = |t_j - t_{j'}|$. The range parameter $\phi > 0$, which is always positive, defines the extent to which the covariance decays to zero. Note that the exponential correlation structure corresponds to the continuous AR(1) or CAR(1) correlation structure, as defined by Pinheiro and Bates (2000, p. 232):

$$\rho(s) = e^{-s/\phi} = (e^{-1/\phi})^s = \Psi^s.$$

The final term represents the form of the CAR(1) correlation structure. For instance, when $\phi = 1.45$, this gives rise to the CAR(1) with $\Psi \approx 0.5017$.

Heterogeneous exponential covariance

In this scenario, the residual (within-subject) errors were assumed to be heteroscedastic (i.e. they exhibit unequal variances) and autocorrelated. The heterogeneous variance-covariance was generated as an exponential function of the variance covariate, as described by Pinheiro and Bates (2000, p.211 - 212),

$$g(t_{ij}, \alpha) = \exp(\alpha t_{ij}).$$

Following this function, the variance model for residual errors can be represented as:

$$\mathrm{Var}(\epsilon_{ij}) = \sigma^2 g^2(t_{ij}, \alpha) = \sigma^2\{\exp(\alpha t_{ij})\}^2 = \sigma^2[\mathbf{\Phi}_i]^2_{jj},$$

where

$$\mathbf{\Phi}_i = \begin{pmatrix} \Phi_{i1} & 0 & 0 & \cdots & 0 \\ 0 & \Phi_{i2} & 0 & \cdots & 0 \\ 0 & 0 & \Phi_{i3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \Phi_{i,n_i} \end{pmatrix}. \tag{4.13}$$

Here, the parameter α dictates the variance trend. Specifically, if $\alpha > 0$, the variance increases over time; conversely, if $\alpha < 0$, the variance decreases over time. Consequently, a diagonal matrix, $\mathbf{\Phi}_i$, describes the variance pattern. The decomposition of the within-subject variance-covariance structure can be represented as:

$$\mathbf{\Lambda}_i = \mathbf{\Phi}_i \mathbf{C}_i \mathbf{\Phi}_i, \tag{4.14}$$

where $\mathbf{C}_i$ is a correlation matrix that follows the structure defined in 4.12. Given these parameters, the residual errors for each child were generated as:

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R_i}), \quad \mathbf{R_i} = \sigma^2 \mathbf{\Lambda}_i. \tag{4.15}$$

In this context, the primary objective is to evaluate the performance of target models featuring two different covariance structures (homogeneous and heterogeneous covariance), while ensuring an adequate representation of the CAR(1) correlation structure to reflect autocorrelated growth outcomes. This choice is not tailored specifically to mimic the GUS data but aims to be applicable to a broader range of datasets and to assess the ability of the random effects part of AQMM in accounting for autocorrelation in repeated observations. However, the heterogeneous exponential covariance most closely reflects the GUS data. This conclusion is supported by fitting the GUS data with simple linear mixed-effects models under four different covariance structures and comparing their goodness-of-fit metrics, including AIC, BIC, and log-likelihood (see Table 4.1). The model is specified as follows:

$$\begin{aligned} Weight_{ij} = \beta_0 + \beta_1(Age)_{ij} + \beta_2(Age)^2_{ij} + \beta_3(Age)^3_{ij} + \beta_4(Age)^4_{ij} \\ + \beta_5(Age)^5_{ij} + \beta_6(Age)^6_{ij} + \beta_7(Age)^7_{ij} + u_{0i}(Age)_{ij} + u_{1i}(Age)_{ij}. \end{aligned}$$

Table 4.1 shows that the heteroscedastic model with an exponential covariance structure yields the lowest AIC and BIC values, as well as the highest log-likelihood. These results suggest that this covariance structure is likely the best fit for the GUS data.

Table 4.1: Summary of goodness-of-fit metrics for linear mixed-effects models with four different covariance structures.

Model	**AIC**	**BIC**	**log-likelihood**
Homoscedastic model	81592.12	81637.34	-40790.06
Heteroscedastic model with exponential variance structure[a]	**72600.42**	**72653.17**	**-36293.21**
Heteroscedastic model with power variance structure[b]	73452.69	73505.45	-36719.34
Heteroscedastic model with fixed variance structure[c]	76935.52	76970.74	-38456.76

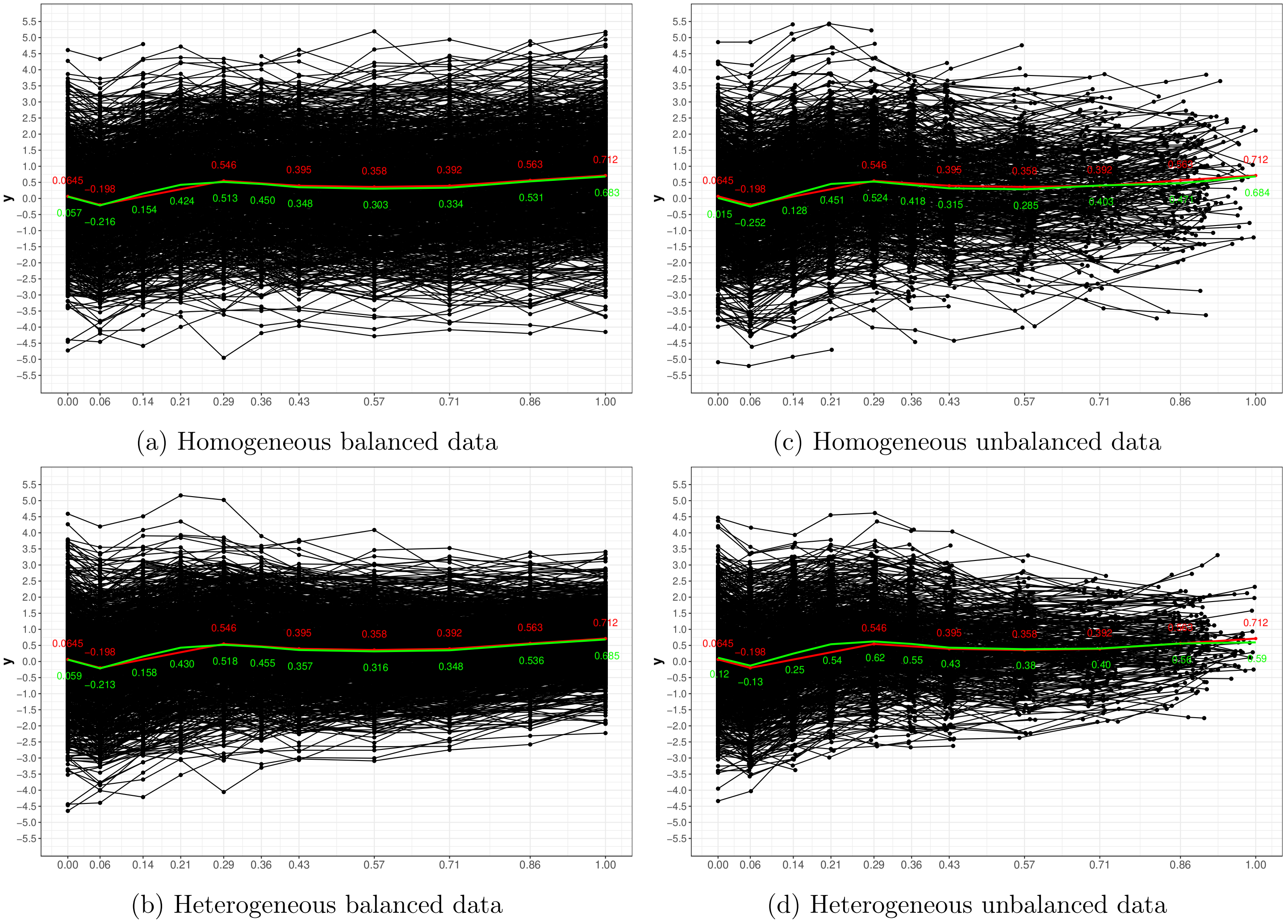
[a] Variance model: $Var(\epsilon_{ij}) = \sigma^2 \exp(2\delta Age_{ij})$ and Variance function: $g(Age_{ij}, \delta) = \exp(\delta Age_{ij})$
[b] Variance model: $Var(\epsilon_{ij}) = \sigma^2 |Age.y_{ij}|^{2\delta}$ and Variance function: $g(Age.y_{ij}, \delta) = |Age.y_{ij}|^{\delta}$
[c] Variance model: $Var(\epsilon_{ij}) = \sigma^2 Age.y_{ij}$ and Variance function: $g(Age.y_{ij}) = \sqrt{Age.y_{ij}}$

To reflect Z-score growth outcomes varying -5 and 5, the residual variance (σ^2) is set to 2. However, it is acknowledged that a few values may fall outside of this range, provided that these values are not overly restrictive. Additionally, the range parameter ϕ is fixed at 1.45, equivalent to CAR(1) with $\Psi = 0.50$, representing moderate correlation between two observations one unit of time apart.

In the scenario of heterogeneous covariance, the variance trend parameter α is set at -0.50 to reflect a decrease in variance over time, a pattern that is explicitly observed in WAZ measurement in GUS data and other longitudinal child growth data (e.g. a South African Birth Cohort (Biljon et al., 2023)).

To evaluate the performance of the target models across different data sample sizes, two distinct sample sizes were selected: 100 (defined as small), and 1000 (defined as large). The later was chosen to reflect the GUS dataset adequately without impacting computational time during simulation.

Figure 4.2 shows plots derived from some simulated data, while Table 4.2 lists all the scenarios considered in this study.



(a) Homogeneous balanced data

(c) Homogeneous unbalanced data

(b) Heterogeneous balanced data

(d) Heterogeneous unbalanced data

Figure 4.2: Example datasets of Study 4.1, generated from the model (4.9) using 1000 children ($N = 1000$) with unequally spaced time observations. Figures (a) and (b) contain the dataset for the balanced data scenario with the *homogeneous* and *heterogeneous* scenario, respectively. Figures (c) and (d) contain the dataset for the unbalanced data scenario with the *homogeneous* and *heterogeneous* scenario, respectively. Red solid lines show the true WAZ means, and green solid lines show the mean WAZ at each age of the simulated

Table 4.2: The scenarios used in Study 4.1

Scenario	Data design	Variance-covariance types of errors ($\mathbf{R}_i$)	Sample sizes (N)
1	Balanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	100
2			1000
3		Heterogeneous ($\sigma^2 = 2, \phi = 1.45,\ \alpha = -0.50$)	100
4			1000
5	Unbalanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	100
6			1000
7		Heterogeneous ($\sigma^2 = 2, \phi = 1.45,\ \alpha = -0.50$)	100
8			1000

Fitting the simulated data

The subsequent step involves fitting the simulated data using QSAM (4.1) and AQMM (4.4), with a simplification that includes only time variable (t_{ij}). Two versions of the model (4.4) were fitted: AQMM1 and AQMM2. Each version utilises distinct penalised methods employing uniform B-splines with 10 equidistant knots spanning the entire range of time variable values. (Eilers & Marx, 1996, 2010; Eilers et al., 2015; Eilers & Marx, 2021; Wood, 2017b). The rationale for setting 10 knots for both AQMM1 and AQMM2 is to adequately capture the time points used in the simulated data. Regarding the placement of knots, equidistant knots were chosen for AQMM2 because they are mandatory for P-splines. To ensure fairness, the same knot placement was also applied to AQMM1. Furthermore, the variance-covariance matrix of random effects was specified as a general positive-definite matrix (also known as "pdSymm" in `aqmm` package in R) for both AQMM1 and AQMM2, in order to account for autocorrelated repeated growth observations.

For the QSAM approach, the non-parametric function was modelled using cubic B-splines with variations in the degrees of freedom (both small and large) to fit the QSAM model with varying degrees of smoothness. In regression splines, the positioning of knots stands as another crucial fixed parameter, shaping the smoothness of the curve. Quantile knots (i.e. unevenly spaced knots), determined by the predictor variable's distribution quantiles, naturally conform to the data's distribution characteristics. This choice proves advantageous when dealing with predictor variable that exhibit non-uniform or skewed, especially in cases of longitudinal child data with age differences between subjects. Furthermore, by employing quantile knots, the spacing of interior knots of cubic B-splines maintains flexibility, adept at accommodating densely clustered data points, as seen in simulated data scenarios. Therefore, flexibility in knot placement was permitted in both QSAM1 and QSAM2, employing quantile knots. It should be noted that QSAM is dependent on regression splines, so P-splines cannot be used to fit the non-parametric function of QSAM.

Furthermore, the number of knots is also typically fixed and cannot be altered during (regression splines) model fitting. While certain algorithms can determine optimal knots, this process often requires computational resources, potentially resulting in extensive time consumption when implementing this approach in simulation studies. In this thesis, the suggestion of Harrell (2015) is adopted to determine the number of knots. Specifically, if the response variable is a continuous (uncensored) variable and the sample size is sufficiently large (e.g. greater than 100), it is recommended to use 5 knots. As mentioned in Section 3.10, the number of effective degrees of freedom (*edf*) represents the number of parameters used in the model and can guide the selection of smothers in addition to or instead of specifying knots directly. Therefore, an *edf* of 5 is chosen for cubic B-splines with quantile knots, which is equivalent to specifying 5 knots (3 interior knots and 2 boundary knots). This number should be adequate to avoid overfitting. However, it is important to note that these recommendations cannot fully guarantee the prevention of overfitting. Note that, in cubic B-splines, the *edf* is defined as *edf* $=$ (number of interior knots) $+$ 3 - 1. Furthermore, the number of *edf* is augmented as much as possible without exceeding the number of time points (i.e. $n_i = 11$), to explore the behaviour of predictive performance in terms of potential overfitting relative to the *edf* of 5. In this instance, the *edf* was set to 10 ($edf = 10$) for balanced data and 8 for unbalanced data ($edf = 8$). This adjustment is necessary as some B-splines might lack data support in situations with extensive degrees or dimensions, particularly with unbalanced data. This can lead to a singularity issue while fitting the quantile regression.

In this chapter, all models were fitted using fixed values of τ at 0.10, 0.50, and 0.90. These three quantiles represent the lower, middle, and upper locations of the distribution of WAZ growth measurements (response distribution). In essence, the 0.10 quantile signifies the lower 10% of children who may be experiencing slower or less favorable growth compared to the rest of the population. The 0.50 quantile indicates the median, with 50% of children falling below or above this point, representing the typical or average growth level within the population. Lastly, the 0.90 quantile indicates the top 10% of children with higher growth compared to the rest of the population. In summary, the four models are:

- AQMM1: AQMM utilising penalised cubic B-splines for a non-linear term associated with t_{ij}, based on a conventional cubic spline penalty. Specifically, it utilises a B-spline basis with a second derivative or quadratic penalty on the basis coefficients. This penalty is sensible for controlling changes in curvature without requiring excessive computational time compared to other higher derivatives, as discussed in Section 3.9.

- AQMM2: AQMM using P-splines for a non-linear term linked to t_{ij}, constructed on a cubic B-spline basis with a discrete quadratic penalty on the basis coefficients.

- QSAM1: QSAM employing cubic B-splines for a non-linear term related to t_{ij}, using quantile knots with a small effective degree of freedom ($edf = 5$).
- QSAM2: QSAM using cubic B-splines for a non-linear term associated with t_{ij}, based on quantile knots with a larger effective degree of freedom ($edf = 10$ for balanced data and $edf = 8$ for unbalanced data).

Additionally, the true model (4.9) was fitted using linear mixed models via the `lmer` function in the `lme4` package in R to compare with the 0.50th quantile models serving as an oracle baseline. In this modelling, polynomial degrees up to 7 were specified for the time variable, t_{ij}, and random effects (intercepts and slopes) were assumed to be correlated, each to both AQMM1 and AQMM2. This model is referred to as MTRUE.

Summarising the results

To assess the accuracy with which each model estimates the conditional quantiles, three metrics were utilised: the coefficient of determination (R-squared) for quantile regression, mean weighted absolute errors (MAE), and the proportion of negative residuals (PNR).

1) Coefficient of determination or R-squared for QR (RS)

RS_τ is a goodness-of-fit statistic used to measure how well the QR model fits a set of observations. This indicator is analogous to the R^2 statistic in classical ordinary least square regression (OLS), as noted by Koenker and Machado (1999b). However, its interpretation is specific to a particular quantile, hence it can be seen as a local measure of goodness of fit (Koenker & Machado, 1999b; Uribe & Guillen, 2020). The formula for R^2_τ is given by:

$$\text{RS}_\tau = 1 - \frac{\hat{V}_\tau}{\tilde{V}_\tau},$$

where $\hat{V}_\tau = \min \sum_{i=1}^{N} \sum_{j=1}^{n_i} \rho_\tau(y_{ij} - \hat{g})$ and $\tilde{V}_\tau = \min \sum_{i=1}^{N} \sum_{j=1}^{n_i} \rho_\tau(y_{ij} - \tilde{g})$. Here, $\hat{g}$ and $\tilde{g}$ are the unrestricted (fully parameterised models) and restricted (non-conditional) quantile estimated models, respectively. Mathematically, since $\hat{V}_\tau$ is always less than or equal to $\tilde{V}_\tau$, it follows that $0 \leq R^2_\tau \leq 1$. If R^2_τ is close to 1, it indicates that the QR model fits the data well at that specific quantile. Note that in the context of quantile regression, there is no modified version of the R-squared, such as an adjusted R-squared, that adjusts the standard R-squared metric to account for the addition of new predictors to a model.

2) Mean weighted absolute errors (MAE)

The MAE measures the variation in the predicted values around the observations. Smaller MAE values indicate more accurate predictions. The MAE of the QR model, at specific quantiles, is given by:

$$\mathrm{MAE}_\tau = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i} \epsilon^*_{ij,\tau}\Big(\tau - I(\epsilon^*_{ij,\tau} < 0)\Big)}{\sum_{i}^{N} n_i}, \tag{4.16}$$

where $\epsilon^*_{ij,\tau} = y_{ij} - \hat{Q}_{y_{ij}|t_{ij}}(\tau)$, $(\tau - I(\epsilon^*_{ij,\tau} < 0))$ is a weighted term, and $I(\cdot)$ is an indicator function.

3) Proportion of negative residuals (PNR)

Another metric to consider is the proportion of negative residuals (PNR). It is given by

$$\mathrm{PNR}_\tau = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i} \Big(I(\epsilon^*_{ij,\tau} < 0)\Big)}{\sum_{i}^{N} n_i},$$

where $\epsilon^*_{ij,\tau}$ and $I(\cdot)$ are defined as in (4.16). Ideally, this metric should be approximately equal to τ.

Results

Five hundred datasets were simulated for each scenario, and all four models with the true model were applied to each dataset.

For brevity, this section presents only the results from the heterogeneous variance scenario; the outcomes for the homogeneous variance scenario are detailed in Appendix B. Overall, the trend of these outcomes was similar to that of the former scenario. Nevertheless, it is noteworthy that the results from the homogeneous variance scenario exhibited a higher MAE and a lower R-squared compared to those observed in the heterogeneous variance scenario, but these did not impact on the PNR. This can be attributed to the AQMM method being based on mixed models, which are designed to accommodate heterogeneous data. Therefore, applying this method to homogeneous data might introduce unnecessary complexity, leading to increased prediction errors.

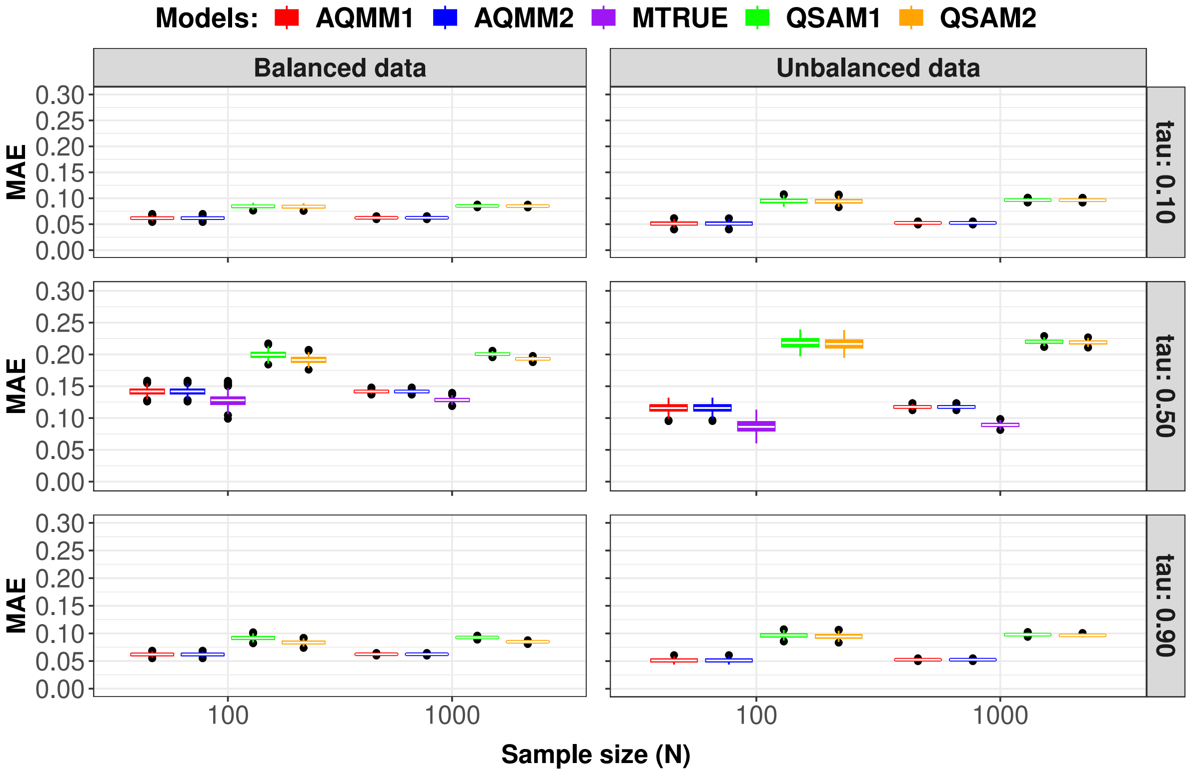


Figure 4.3: The MAE of four models, including the MSE for the true model (MTRUE), in the *heterogeneous* scenario of Study 4.1. The left column presents the results for the balanced data scenario, while the right column shows the results for the unbalanced data scenario. The three rows display the results for quantile levels at 0.10, 0.50 and 0.90, respectively.

Figure 4.3 shows that the MAE values of two extreme quantile models ($\tau = 0.10$ and $\tau = 0.90$) were relatively similar and smaller than those at the median quantile ($\tau = 0.50$), consistent across various data design scenarios and sample sizes. This occurs because the weights applied to the residuals differ between quantile models, affecting the total error differently based on the distribution and magnitude of the residuals. Additionally, this indicates that the two extreme quantile models provided accurate predictions not only for the lower 10% of the data but also for the upper 10%, implying a robust performance across various segments of the data distribution. When comparing different models, AQMM1 and AQMM2 consistently showed lower MAE values than QSAM1 and QSAM2 across all quantiles, regardless of sample sizes and data design scenarios. However, the MAE values at the median quantile for both AQMM1 and AQMM2 were slightly higher than the MSE values from the true model (MTRUE). Additionally, both QSAM models showed higher MAEs with unbalanced data compared to balanced data. These findings suggest that AQMM approaches are generally more accurate in predicting outcomes in across various scenarios. Note that, in the homogeneous variance scenarios with balanced data, the MSE value from MTRUE was higher than the AQMM models for both sample sizes (see Figure B.1). This demonstrates an evident that when modelling the homogeneous data using the models based on mixed model framework may introduce unnecessary complexity with resulting in high prediction errors.

Regarding the two data designs, the MAE values for AQMM1 and AQMM2 were slightly smaller in the unbalanced data compared to the balanced data across three quantiles. Conversely, the MAE values for QSAM1 and QSAM2 were smaller in the balanced data than in the unbalanced data, particularly at the median quantile. This suggests that AQMM models are suitable for both data designs, while QSAM models may require certain conditions to optimize their performance, especially in unbalanced data. Considering sample sizes, the MAE values for all models showed slight variation in the large sample size ($N = 1,000$) compared to the small sample size ($N = 100$). This indicates that an increase in sample size leads to smaller variations in predictions for all models. Hence, the differences in MAE values between AQMM models and QSAM models are considerable, indicating that the former provides more reliable predictions. This superiority is evident across all scenarios tested, highlighting the robustness of AQMM models.

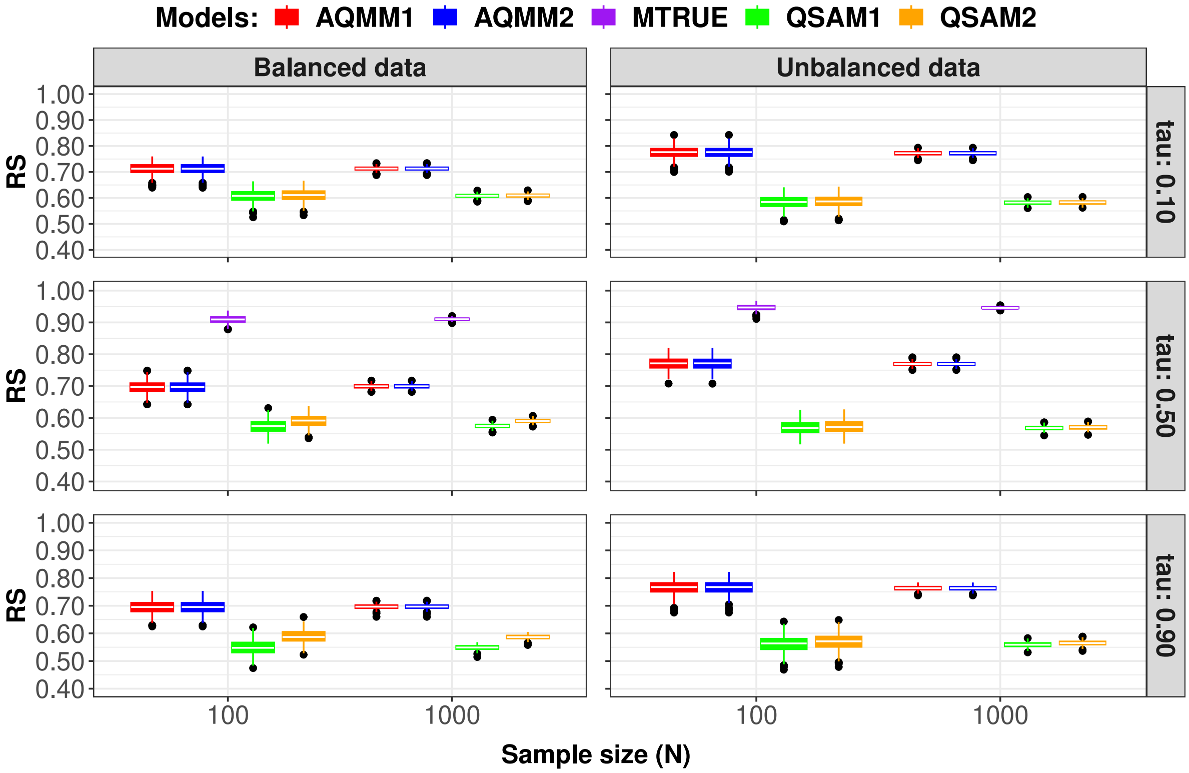


Figure 4.4: The RS of four models, including the true model (MTRUE), in the *heterogeneous* scenario of Study 4.1. The left column presents the results for the balanced data scenario, while the right column shows the results for the unbalanced data scenario. The three rows display the results for quantile levels at 0.10, 0.50 and 0.90, respectively.

Figure 4.4 presents the RS values for the three quantile models in the heterogeneous variance scenario across two distinct data designs and sample sizes. The RS values of all models were relatively similar across three quantiles. Notably, the AQMM1 and AQMM2 models provided a superior fit to the set of observations compared to the models fitted by QSAM1 and QSAM2 across quantiles. However, at the median quantile, the RS values from these models were smaller than of the MTRUE model, suggesting that the MTRUE model provides a better overall fit to the data as it represents the true model. Considering different data designs, the AQMM approaches demonstrated slightly better performance in the unbalanced data scenario than in the balanced data scenario, whereas the QSAM approaches seemed to be similar. In terms of sample sizes, increasing sample sizes to be large ($N = 1,000$) tended to yield slight variation in RS values compared to the small sample size ($N = 100$).

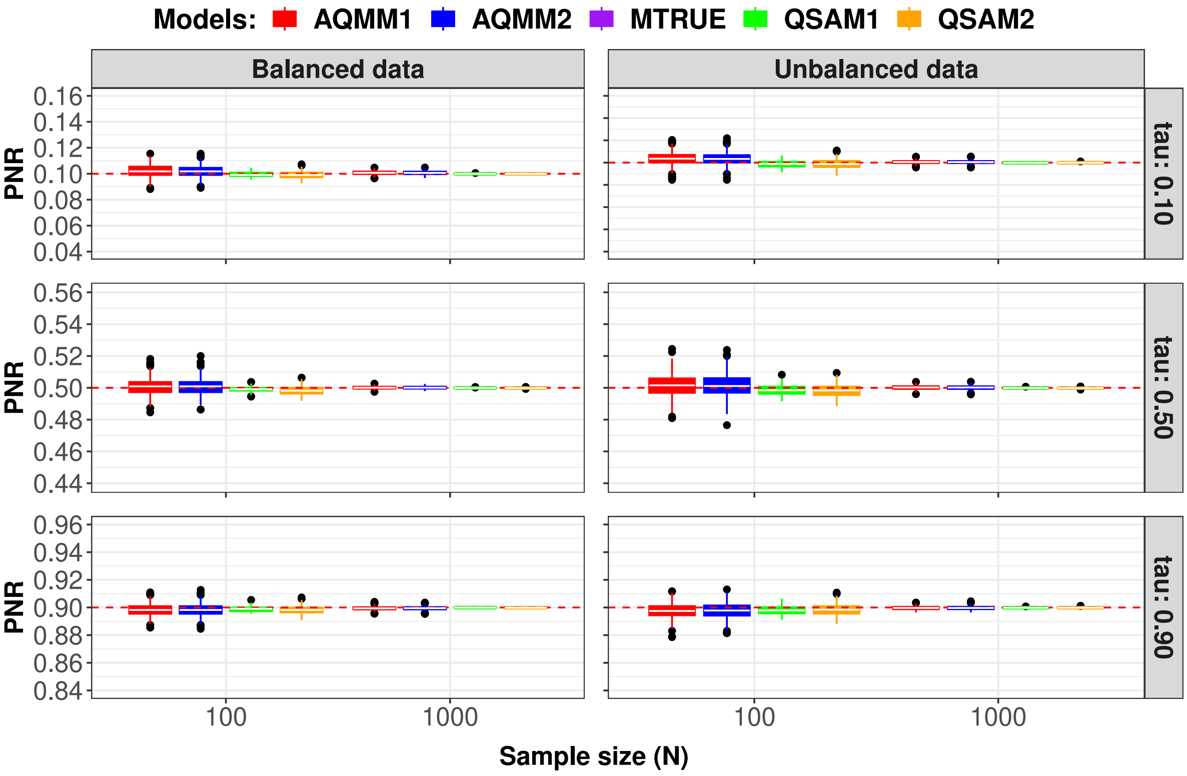


Figure 4.5: The PNR of four models in the *heterogeneous* scenario of Study 4.1. The left column presents the results for the balanced data scenario, while the right column shows the results for the unbalanced data scenario. The three rows display the results for quantile levels at 0.10, 0.50 and 0.90, respectively. The red dashed lines represent the expected quantile levels, $\tau = 0.10, 0.50$, and 0.90, respectively.

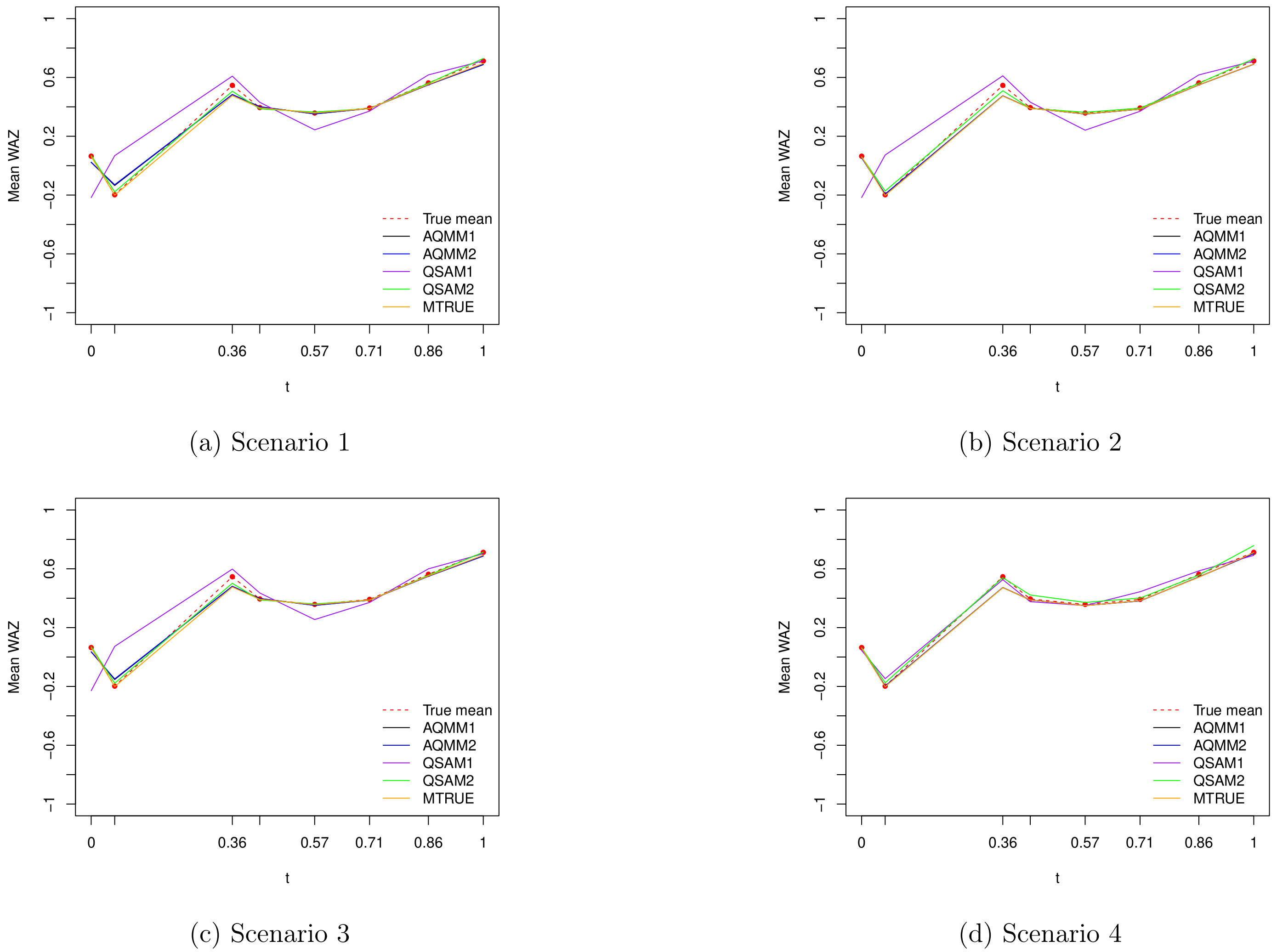
Figure 4.5 depicts the PNR values for the three quantile models in the heterogeneous variance scenario across two distinct data designs and sample sizes. It is evident that each quantile model derived from the four approaches yielded PNR values closely aligned with the expected quantile levels ($\tau = 0.10, 0.50, 0.90$) across both data designs and sample sizes. When considering the two sample sizes, the 0.50th quantile model provided average PNR values close to the expected quantile level at 0.50 for both sample sizes. Specifically, when the sample size was small, the PNR values of the two extreme quantile models slightly deviated from their expected quantile levels at 0.10 and 0.90, respectively. This suggests that the AQMM model is particularly sensitive to sample sizes, especially for extreme quantiles. Moreover, the AQMM1 and AQMM2 exhibited greater variability in PNR values compared to those of QSAM1 and QSAM2, especially when the sample size was small. Notably, the AQMM approaches demonstrated increased variation in PNR in the unbalanced data scenario compared to the balanced data scenario.

Furthermore, to explore how well all models fit the simulated data, Figures 4.6 and 4.7 present plots of the true mean versus the predicted mean at specified time points for the 0.50th quantile model of AQMM1, AQMM2, QSAM1, QSAM2, and MTRUE across the eight scenarios presented in Table 4.2. The results indicate that MTRUE, AQMM1 and AQMM2 provided predicted means that closely matched the true means at specified time points in all scenarios. These models also excelled in capturing growth patterns, while the QSAM model with small *edf* underperformed in some scenarios, particularly in the case of balanced data with a small sample size.

A repeated measures ANOVA was conducted to examine the influence of the "Model" factor and the three LCGD design factors ("Data design", "Variance-covariance types of errors", and "Sample size") on each metric. This approach was taken because each metric was measured by the four models on each of the 500 simulated datasets, resulting in repeated outcomes. The model for this approach can be expressed as follows:

$$\begin{aligned} Y_{ijklm} =& \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\delta)_{il} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + s_m + \epsilon_{ijklm}, \\ & m = 1, \dots, 4000, i = 1, 2, j = 1, 2, k = 1, 2, l = 1, \dots, 4, \end{aligned}$$

where y_{ijklm} is the metric measurement (i.e. MAE, RS, and PNR) for the m-th simulated dataset at the i-th level of the variance-covariance of errors ($\mathbf{R}$), the j-th level of the data design, and the k-th level of the sample sizes, and the l-th level of the models, α_i is the effect of the i-th level of the variance-covariance of errors, β_j is the effect of the j-th level of the data design, γ_k is the effect of the k-th level of the sample size, δ_l is the effects of the l-th level of the models, $(\alpha\delta)_{il}, (\beta\delta)_{jk}, (\gamma\delta)_{kl}$ are the two-way interactions, s_m is the random effect of the m-th simulated dataset, and ϵ_{ijklm} is the residual error. Note that only two-way interactions between the models and other factors were observed in this analysis. Additionally, the Eta-squared (η^2), a measure of effect size used in ANOVA, was employed to quantify the proportion of variance in the dependent variable explained by each of the model's terms. It ranges from 0 to 1, with larger values indicating a greater proportion of variance explained by the model's term(s).



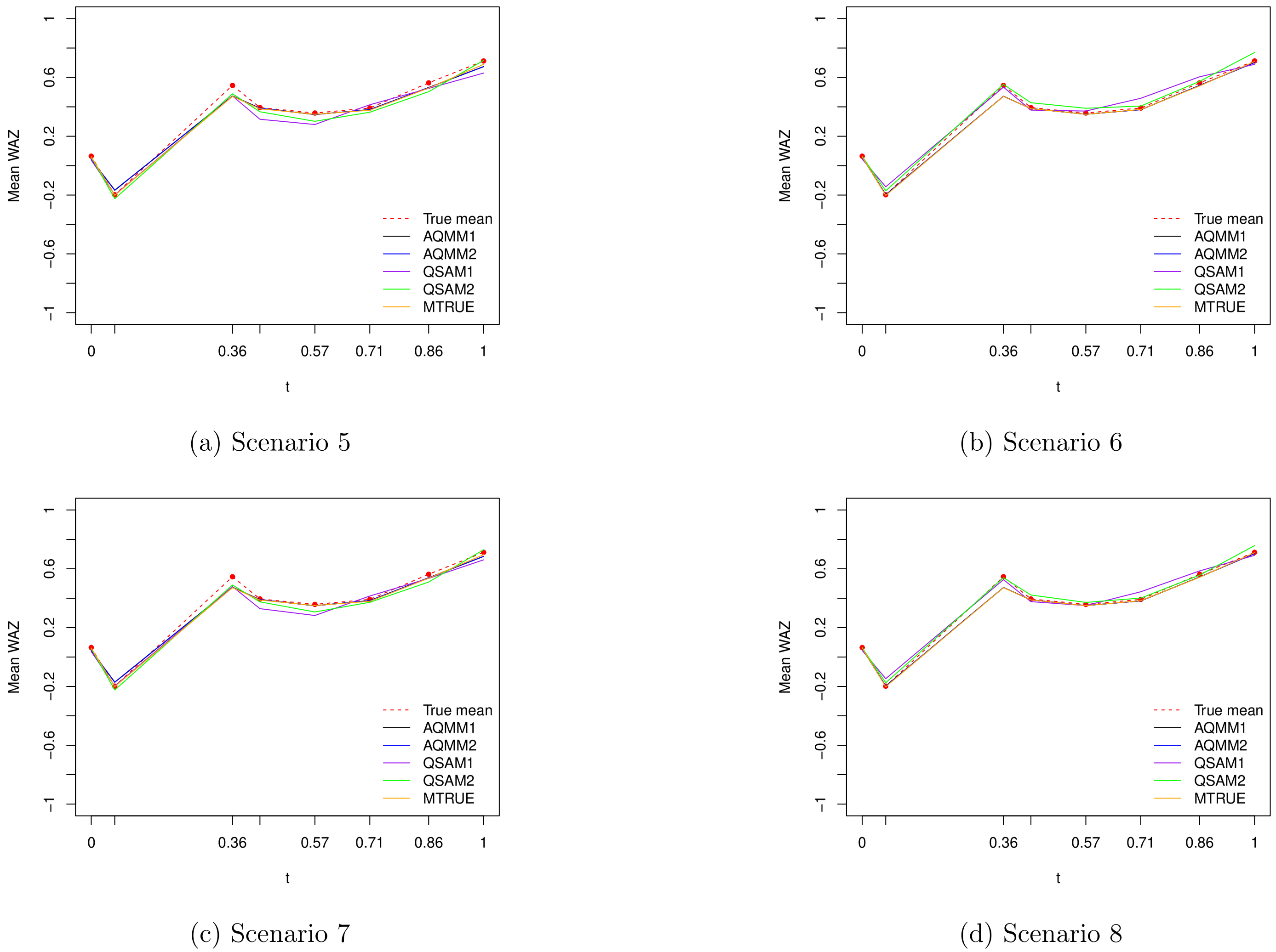
(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 4.6: The true mean versus predicted mean at the 0.50th quantile model for Scenarios 1 through 4.



(a) Scenario 5

(b) Scenario 6

(c) Scenario 7

(d) Scenario 8

Figure 4.7: The true mean versus predicted mean at the 0.50th quantile model for Scenarios 5 through 8.

As shown in Tables 4.3 to 4.4, the "Model" factor emerged as the most significant influence on predictive performance across three quantile models, particularly on MAE and R-squared. The interaction between the "Model" factor and the "Data design" was identified as the second most influential factor. This indicates that the effect of the "Model" factor depends on the level of the "Data design" factor. For instance, both AQMM1 and AQMM2 outperformed in situations with unbalanced data compared to balanced data. Moreover, in terms of the "Variance-covariance error" factor, they performed better under heterogeneous variance than under homogeneous variance, suggesting that AQMM can effectively account for this type of variance. The "Sample size" factor had a relatively minor influence in this context. These trends are especially apparent in metrics such as MAE and R-squared across the three quantile models. Importantly, while the influence of these factors was evident in terms of MAE and R-squared, the "Variance-covariance error" factor did not substantially affect predictive performance as measured by PNR, as indicated in Table 4.5.

Table 4.3: Repeated measures ANOVA for three quantile models on MAE of Study 4.1

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	**Between subjects**				
	Variance-covariane of errors (**R**) (Homogeneous and Heterogeneous)	1	0.56	19678.70***	0.83
	Data design (Balanced and Unbalanced data)	1	0.04	1346.30***	0.25
	Sample size ($N = 100$ and $N = 1000$)	1	0.01	208.60***	0.05
	Residuals	3996	0.11		
	Within subjects				
	Model (AQMM1, AQMM2, QSAM1, and QSAM2)	3	4.76	475750.00***	0.99
	Model * R	3	0.01	840.92***	0.17
	Model * Data design	3	0.43	42530.00***	0.91
	Model * N	3	0.00	38.94***	0.01
	Residuals	11988	0.04		
0.50	**Between subjects**				
	Variance-covariane of errors (**R**)	1	2.85	22220.64***	0.85
	Data design	1	0.35	2740.21***	0.41
	Sample size (N)	1	0.01	45.81***	0.01
	Residuals	3996	0.51		
	Within subjects				
	Model	3	26.53	806226.82***	1.00
	Model * R	3	0.05	1355.72***	0.25
	Model * Data design	3	2.19	66460.62***	0.94
	Model * N	3	0.00	42.08***	0.01
	Residuals	11988	0.13		
0.90	**Between subjects**				
	Variance-covariane of errors (**R**)	1	0.56	19740.30***	0.83
	Data design	1	0.08	2877.30***	0.42
	Sample size (N)	1	0.01	179.40***	0.04
	Residuals	3996	0.11		
	Within subjects				
	Model	3	5.44	564885.12***	0.99
	Model * R	3	0.01	809.92***	0.17
	Model * Data design	3	0.34	35488.55***	0.90
	Model * N	3	0.00	49.42***	0.01
	Residuals	11988	0.04		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Table 4.4: Repeated measure for three quantile models on R-squared of Study 4.1

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	**Between subjects**				
	Variance-covariane of errors (**R**) (Homogeneous and Heterogeneous)	1	0.47	445.54***	0.10
	Data design (Balanced and Unbalanced data)	1	1.98	1866.65***	0.32
	Sample size ($N = 100$ and $N = 1000$)	1	0.00	3.49	0.00
	Residuals	3996	4.23		
	Within subjects				
	Model (AQMM1, AQMM2, QSAM1, and QSAM2)	3	84.38	631166.43***	0.99
	Model * R	3	0.01	81.43***	0.02
	Model * Data design	3	6.91	51648.93***	0.93
	Model * N	3	0.00	21.28***	0.01
	Residuals	11988	0.53		
0.50	**Between subjects**				
	Variance-covariane of errors (**R**)	1	0.03	32.39***	0.01
	Data design	1	3.72	4474.07***	0.54
	Sample size (N)	1	0.00	42.25	0.00
	Residuals	3996	3.11		
	Within subjects				
	Model	3	94.12	1031000.00***	1.00
	Model * R	3	0.09	951.00***	0.19
	Model * Data design	3	6.83	74890.00***	0.95
	Model * N	3	0.00	21.19***	0.01
	Residuals	11988	0.36		
0.90	**Between subjects**				
	Variance-covariane of errors (**R**)	1	0.14	111.75***	0.03
	Data design	1	4.16	3373.48***	0.46
	Sample size (N)	1	0.00	2.02***	0.00
	Residuals	3996	4.93		
	Within subjects				
	Model	3	101.21	714903.61***	0.99
	Model * R	3	0.15	1091.24***	0.21
	Model * Data design	3	5.53	39087.03***	0.91
	Model * N	3	0.00	30.56***	001
	Residuals	11988	0.57		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Table 4.5: Repeated measure ANOVA for three quantile models on PNR of Study 4.1

τ	Source of Variation	DF	SS	F	η^2
0.10	**Between subjects**				
	Variance-covariane of errors (**R**) (Homogeneous and Heterogeneous)	1	0.00	1.80	0.00
	Data design (Balanced and Unbalanced data)	1	0.00	2.50	0.00
	Sample size ($N = 100$ and $N = 1000$)	1	0.00	93.12***	0.02
	Residuals	3996	0.06		
	Within subjects				
	Model (AQMM1, AQMM2, QSAM1, and QSAM2)	3	0.02	910.48***	0.19
	Model * R	3	0.00	2.04	0.00
	Model * Data design	3	0.00	33.63***	0.01
	Model * N	3	0.01	418.66***	0.09
	Residuals	11988	0.08		
0.50	**Between subjects**				
	Variance-covariane of errors (**R**)	1	0.00	0.27	0.00
	Data design	1	0.00	3.66	0.00
	Sample size (N)	1	0.00	14.61***	0.00
	Residuals	3996	0.07		
	Within subjects				
	Model	3	0.01	296.49***	0.07
	Model * R	3	0.00	0.03	0.00
	Model * Data design	3	0.00	21.37***	0.01
	Model * N	3	0.01	193.31***	0.05
	Residuals	11988	0.10		
0.90	**Between subjects**				
	Variance-covariane of errors (**R**)	1	0.00	6.49*	0.00
	Data design	1	0.00	6.88**	0.00
	Sample size (N)	1	0.01	747.70***	0.16
	Residuals	3996	0.05		
	Within subjects				
	Model	3	0.00	49.89***	0.01
	Model * R	3	0.00	1.74	0.00
	Model * Data design	3	0.00	4.73**	0.00
	Model * N	3	0.00	34.49**	0.00
	Residuals	11988	0.08		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Additionally, to compare the different approaches in terms of computational efficiency, Table 4.6 presents the average computational times (in seconds) over all 500 simulation runs. The numbers indicate that the AQMM models noticeably consumed more computational time than the QSAM models and MTRUE across the eight scenarios and three quantiles. When the sample size increased from 100 to 1000, the average computational times of the AQMM models increased by approximately 10 times. In this case, they consumed times around 4 minutes to fit two extreme quantile models and around 3 minutes for fitting the 0.50th quantile model.

Table 4.6: Average computational times (in seconds) for five models across eight scenarios.

τ	**Scenario**	**AQMM1**	**AQMM2**	**QSAM1**	**QSAM2**	**MTRUE**
0.10	1 ($N = 100$)	21.26	22.32	0.01	0.04	-
	2 ($N = 1000$)	251.82	247.87	0.31	0.47	-
	3 ($N = 100$)	21.68	22.25	0.01	0.04	-
	4 ($N = 1000$)	253.57	252.78	0.32	0.46	-
	5 ($N = 100$)	18.97	19.62	0.01	0.01	-
	6 ($N = 1000$)	232.82	246.52	0.10	0.12	-
	7 ($N = 100$)	18.99	19.81	0.01	0.01	-
	8 ($N = 1000$)	230.31	244.40	0.10	0.11	-
0.50	1 ($N = 100$)	15.74	16.46	0.01	0.03	0.11
	2 ($N = 1000$)	192.23	189.67	0.32	0.44	0.88
	3 ($N = 100$)	16.47	16.70	0.01	0.03	0.12
	4 ($N = 1000$)	188.09	193.48	0.33	0.43	0.85
	5 ($N = 100$)	14.41	15.21	0.01	0.01	0.08
	6 ($N = 1000$)	198.53	208.08	0.11	0.12	0.54
	7 ($N = 100$)	14.62	15.68	0.01	0.01	0.08
	8 ($N = 1000$)	199.99	208.76	0.11	0.11	0.51
0.90	1 ($N = 100$)	20.46	20.97	0.01	0.03	-
	2 ($N = 1000$)	244.07	247.91	0.35	0.45	-
	3 ($N = 100$)	21.26	21.92	0.01	0.03	-
	4 ($N = 1000$)	251.89	249.11	0.35	0.45	-
	5 ($N = 100$)	18.11	19.37	0.01	0.01	-
	6 ($N = 1000$)	228.20	242.09	0.13	0.13	-
	7 ($N = 100$)	18.30	19.00	0.01	0.01	-
	8 ($N = 1000$)	230.43	244.26	0.12	0.12	-

Summary

In conclusion, it can be inferred that AQMM has the capability to address both homoscedastic and heteroscedastic features, as well as autocorrelation among repeated measurements in LCGD, regardless of whether the data is balanced. However, AQMM appears to perform exceptionally well with unbalanced data, as its mixed model framework serves this aspect effectively. The observation that AQMM's performance improves with an increase in sample size aligns with expectations. Additionally, employing two distinct penalised methods (a derivative penalty and a discrete penalty) based on the cubic B-spline basis does not markedly influence the overall performance of the model, especially in the case of AQMM. This suggests that P-splines with a discrete (quadratic) penalty are flexible and well-suited for modelling child growth patterns in this context. Although the AQMM models require extensive computational time compared to the QSAM models, their computational efficiency is still acceptable.

4.4.2 Study 4.2

In the context of real-world data, various additional risk factors come into play, including biological sex, parental influences (such as genetics and lifestyle), socioeconomic status, and various other variables that can significantly impact child growth development. Fortuitously, both AQMM and QSAM methods offer a convenient way for addressing these risk factors by integrating them as linear predictors. Consequently, a comprehensive assessment of the performance of both approaches becomes imperative when applying them to model the longitudinal child growth.

Moreover, it is important to highlight that random errors within empirical data might deviate from a normal distribution. This deviation can manifest as a profusion of outliers featuring remarkably high values situated at the extreme tails of the distribution. Such distributions are termed heavier-tailed distributions.

Aim

This simulation study was conducted to evaluate the accuracy with which AQMM and QSAM can estimate conditional quantile functions of child growth measurement when incorporated alongside other independent variables. Additionally, it assesses the performance of parameter estimation for the linear predictor term.

Data generation

The data were generated from the model below:

$$y_{ij} = \mu_{ij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \epsilon_{ij}, \tag{4.17}$$

where μ_{ij} is defined to correspond to the (intercept) function (4.10), $\beta_1 = 1$, $\beta_2 = 0.5$, $x_{1ij} \sim \text{Bin}(1, 0.50)$ (representing the factor variable, for example sex), $x_{2ij} \sim N(0,1)$ (representing the covariate, for example birth weight), and ϵ_{ij} is a random error. Following model (4.17), the conditional quantile functions of y can be given by

$$Q_{y_{ij}|t_{ij},x_{1ij},x_{2ij}}(\tau) = \mu_{ij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + F_\epsilon^{-1}(\tau) = \alpha(\tau) + \beta_1 x_{1ij} + \beta_2 x_{2ij},$$

where $\alpha(\tau) = \mu_{ij} + F_\epsilon^{-1}(\tau)$, with $F_\epsilon^{-1}(\tau)$ being the quantile function of ϵ. In this case, $\alpha(\tau)$ are vertical translations of one another, relying on the concept of the location-scale shift model. In contrast, β_1 and β_2 are identical across all quantile levels.

In this simulation study, the time variable was generated using the same technique as in Study 4.1. Two designs were considered for the data: balanced and unbalanced. To generate the autocorrelated growth outcomes, the residual errors (ϵ_{ij}) were still assumed to follow a zero-centred multivariate normal distribution with two covariance structure scenarios, namely homogeneous exponential covariance and heterogeneous exponential covariance, as in Study 4.1. In another case, a heavier-tailed distribution, a zero-centered multivariate t-distribution with four degrees of freedom ($\nu = 5$), was also assumed for random errors:

$$\boldsymbol{\epsilon}_i \sim \mathcal{T}_{n_i,5}(\mathbf{0}, \boldsymbol{\Sigma}_i),$$

where $\boldsymbol{\Sigma}_i$ is a scale matrix (not the covariance matrix). Theoretically, for $\nu > 2$, the covariance matrix of $\boldsymbol{\epsilon}_i$ can be defined as $\nu/(\nu-2)\boldsymbol{\Sigma}_i$; otherwise it is undefined (Kotz & Nadarajah, 2004). In order for the simulated random errors to exhibit an exponential covariance structure, $\boldsymbol{\Sigma}_i$ can be defined as $\boldsymbol{\Sigma}_i = ((\nu-2)/\nu)\sigma^2\mathbf{C}_i$, where $\mathbf{C}_i$ is defined in 4.12 for homogeneous covariance, and $\boldsymbol{\Sigma}_i = ((\nu-2)/\nu)\sigma^2\boldsymbol{\Lambda}_i$, where $\boldsymbol{\Lambda}_i$ is defined in 4.14 for heterogeneous covariance. In this study, $\nu = 5$ and $\sigma^2 = 2$ yield simulated growth outcomes consistent with actual Z-score growth outcomes, which range between -5 and 5. However, it is acknowledged that a few values may fall outside of this range, given that these instances are not too restrictive.

To fit the simulated data, the same models used in Study 4.1 were applied. These models include AQMM1 (AQMM with penalised cubic B-splines), AQMM2 (AQMM with cubic P-splines), QSAM1 (with the small edf), QSAM2 (with the large edf), and MTRUE (the true model). Each model was fitted following the processes outlined in Section 4.4.1.

Furthermore, to assess the performance of each model, the same metrics were employed (i.e. R-squared for QR, MAE, and PNR) as outlined in Section 4.4.1. Similar to Study 4.1, a repeated measured ANOVA was conducted to investigate the factor influencing on each metric, based on the model:

$$Y_{pijklm} = \mu + \lambda_p + \alpha_i + \beta_j + \gamma_k + \delta_l + (\lambda\delta)_{pl} + (\alpha\delta)_{il} + (\beta\delta)_{jl} + (\gamma\delta)_{kl} + s_m + \epsilon_{pijklm},$$

where $m = 1, \ldots, 4000, p = 1, 2, i = 1, 2, j = 1, 2, k = 1, 2, l = 1, \ldots, 4$.

Here, Y_{pijklm} is the metric measurement for the m-th simulated dataset at the p-th level of the error distribution ($f(\epsilon)$), the i-th level of the variance-covariance of errors ($\mathbf{R}$), the j-th level of the data design, the k-th level of the sample sizes, and the l-th level of the models. λ_p is the effect of the p-th level of the error distribution. λ_p α_i, β_j γ_k, δ_l are effects representing the error distribution, the variance-covariance of errors, the data design, the sample sizes, and the models, respectively. $(\lambda\delta)_{pl}$, $(\alpha\delta)_{il}$, $(\beta\delta)_{jl}$, $(\gamma\delta)_{kl}$ are the two-way interactions. s_m is the random effect of the m-th simiulated dataset, and ϵ_{pijklm} is the residual error.

Additionally, the bias and the root-mean-squared error (RMSE) of the coefficients for the linear predictors, β_1 and β_2, were calculated to evaluate the estimation methods in both models.

Bias

The *bias* values provide insight into the performance of the parameter estimation methods employed. This metric aims to gauge the average discrepancies between the estimated parameters and their true values. The bias of the parameters associated with the linear predictors within each model can be summarised as the Mean Bias Error (MBE):

$$\text{MBE}(\beta_h) = \frac{1}{500} \sum_{r=1}^{500} \left(\hat{\beta}_{hr} - \beta_h \right), \quad h = 1, 2.$$

A MBE approaching zero suggests that the chosen parameter estimation method effectively approximates the parameter of interest. However, it is important to note that the bias is suitable for assessing precision, as it solely represents the average discrepancy between estimates and true values, without accounting for their direction or magnitude.

Root-mean-squared error (RMSE)

Variability is an important quantity for assessing the precision of estimation. In this regard, the Root Mean Square Error (RMSE) assumes a crucial role. RMSE is pivotal in quantifying the dispersion of estimates around the true value. A smaller RMSE value signifies a higher degree of precise estimation. Mathematically, the RMSE of the parameters within each model is expressed as:

$$\text{RMSE}(\beta_h) = \sqrt{\frac{1}{500} \sum_{r=1}^{500} \left(\hat{\beta}_{hr} - \beta_h \right)^2}, \quad h = 1, 2.$$

By utilising RMSE, the extent of dispersion in the estimates from the actual values can be effectively measured, ultimately indicating the level of precision attained.

Results

The results of this study are only presented for a specific heterogeneous variance scenario, while all findings related to homogeneous variance scenario are provided in Appendix B. Overall, the outcomes obtained from the homogeneous variance scenario exhibited a considerable degree of similarity of those observed in the heterogeneous scenario.

In terms of predictive performance, Figures 4.8, 4.9, and 4.10 demonstrate that the values for three metrics (i.e., MAE, R-squared, PNR) in each quantile for the normal error scenario and the Student's t-distribution showed a similar trend across scenarios involving two data designs and two distinct sample sizes. However, in the case of the Student's t-distribution error, there was greater variability or dispersion in the MAE and R-squared values compared to the normal error scenario. When specifically considering the three quantiles, it is noted that the MAE values for the 0.50th quantile were higher than those for the two extreme quantiles, which appeared to have similar values. Additionally, at the 0.50th quantile model, the MSE values of the true model (MTRUE) were smaller than those of other models, while the R-squared values were vice versa. However, two AQMM models tended to provide values close to MTRUE compared to two QSAM models, especially the MAE values. This suggests that both AQMM and MTRUE models are making predictions with similar levels of accuracy at the central location of the response distribution.

It is important to highlight that the R-squared values of the three quantile models obtained from both AQMM1 and AQMM2 displayed a noticeable increase compared to those observed in Study 4.1. This observation is consistent with the property of this metric, wherein an increase in the number of independent variables in the model leads to an increase in R-squared value. In the quantile regression context, there is no adjusted R-squared value to account for this aspect. Furthermore, the PNR values were closely aligned with their corresponding quantile levels. Similar to Study 1, both AQMM1 and AQMM2 provided PNR values that slightly deviated from the corresponding quantile levels at the two extreme quantiles.

The results concerning parameter estimation performance are presented in Tables 4.7 and 4.8. Across the two error distributions, both AQMM models exhibited smaller MBE and RMSE values than those of the QSAM models for all three quantiles. Notably, both QSAM models exhibited suboptimal performance for β_1, resulting in high bias and RMSE values. Evidently, as sample sizes increased, all models demonstrated a tendency to yield reduced values of both bias and RMSE.

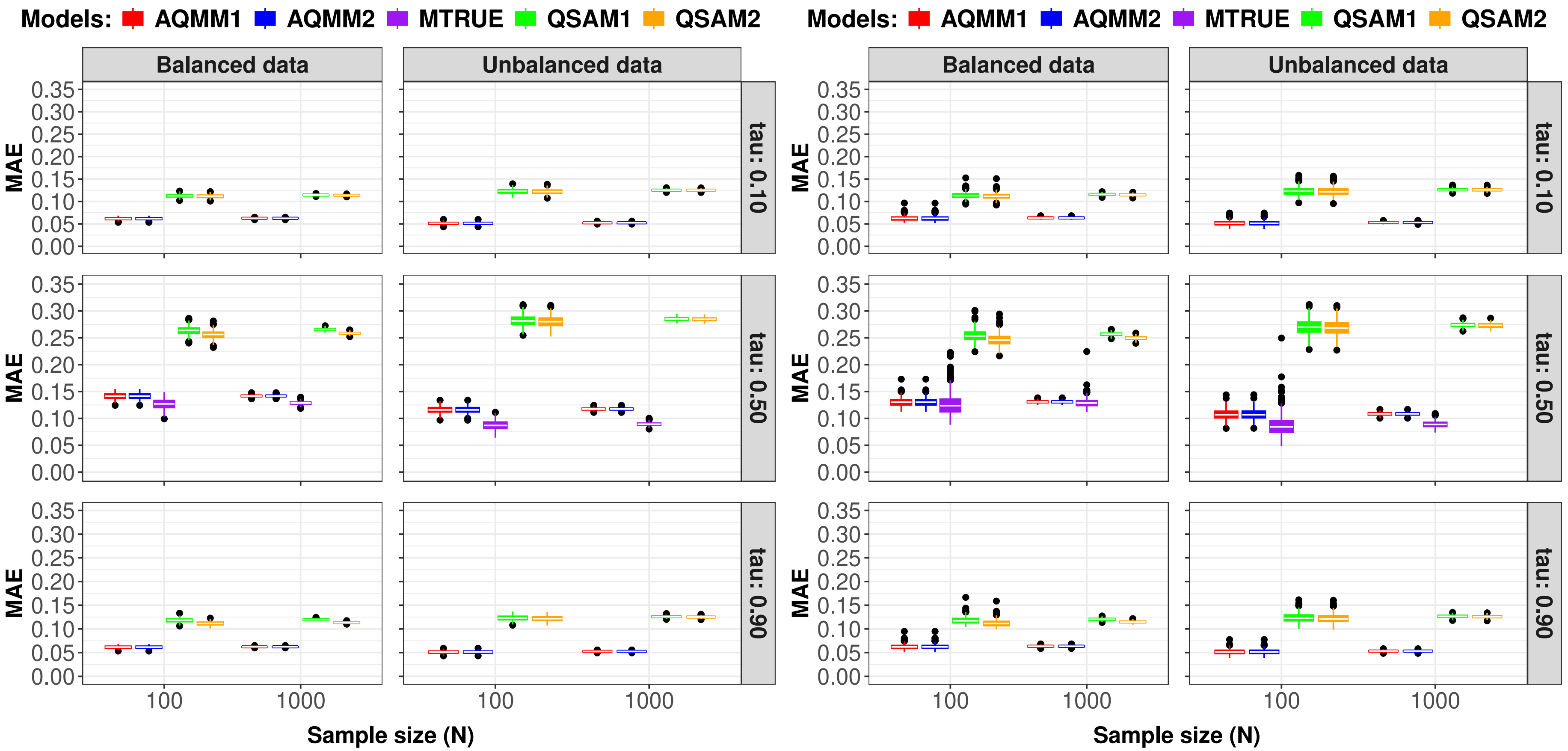


Figure 4.8: The MAE of the four models in the *heterogeneous* scenario of Study 4.2. The left figure represents the normal error case, while the right figure represents the t ($\mathcal{T}_5$) error case. The three rows in each figure contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively.

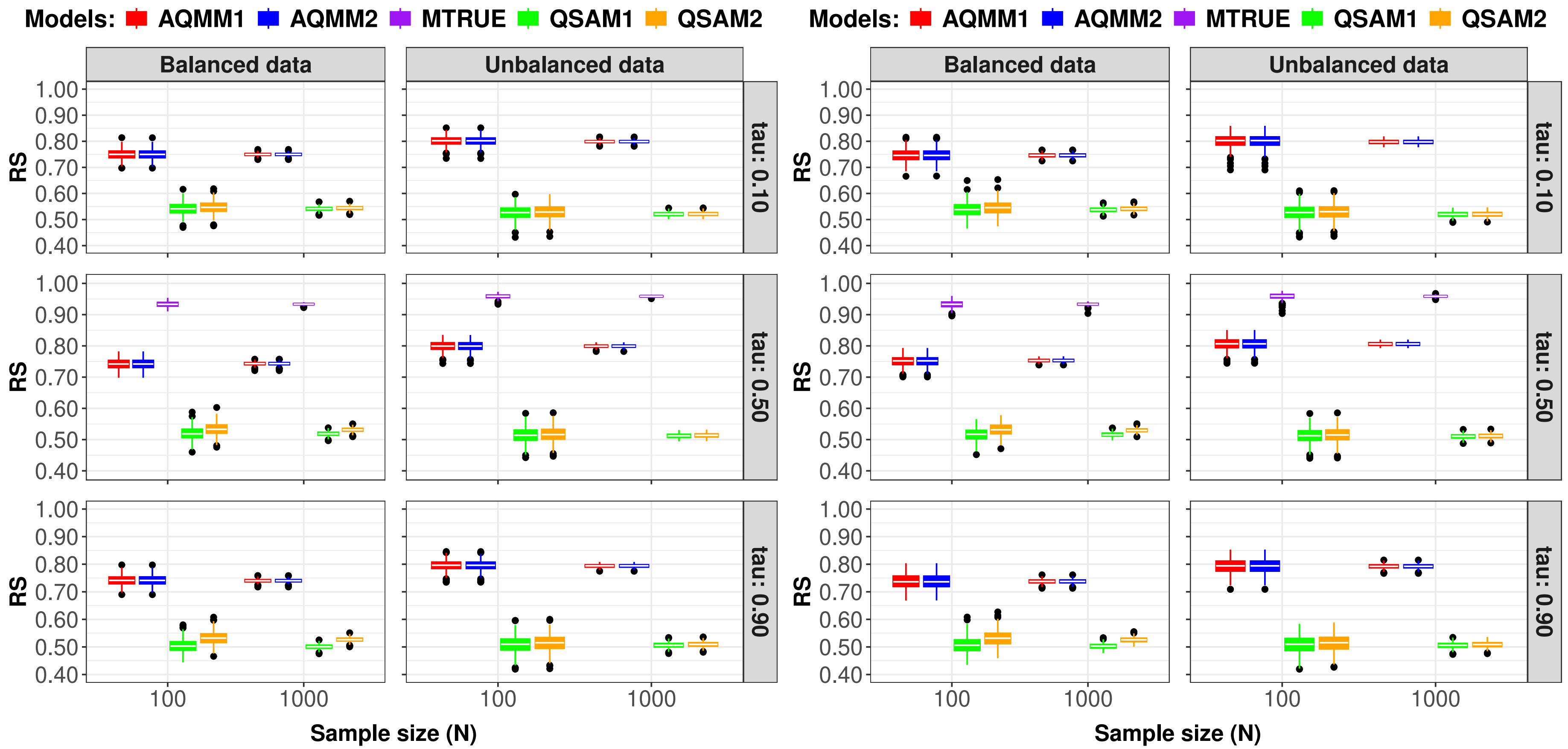


Figure 4.9: The RS of the four models in the *heterogeneous* scenario of Study 4.2. The left figure represents the normal error case, while the right figure represents the t ($\mathcal{T}_5$) error case. The three rows in each figure contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively.

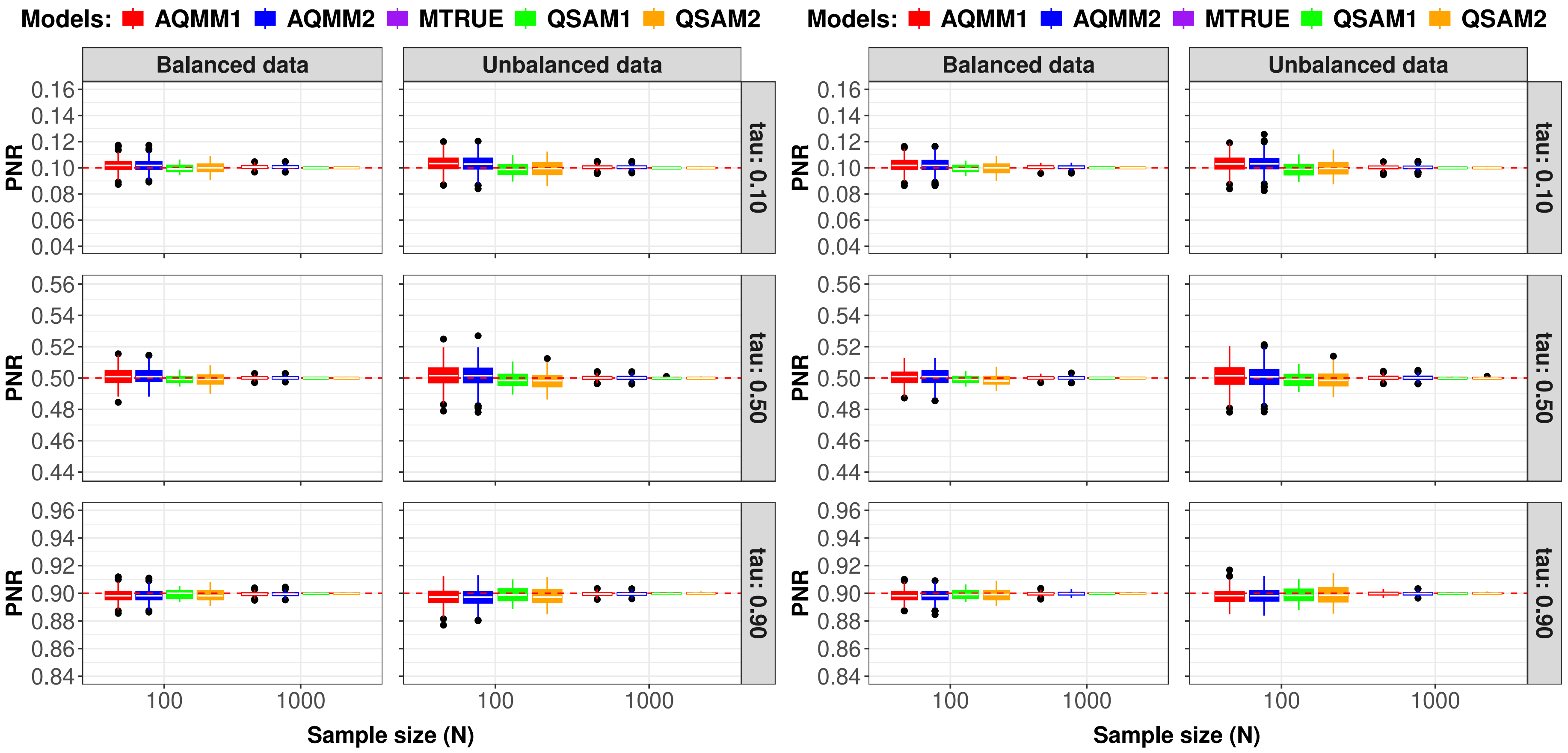


Figure 4.10: The PNR of the four models in the *heterogeneous* scenario of Study 4.2. The left figure represents the normal error case, while the right figure represents the t ($\mathcal{T}_5$) error case. The three rows in each figure contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively.

Table 4.7: MBE and RMSE concerning the simulated data under Study 4.2, specifically focusing on the **unbalanced** data design, **heterogeneous** variance-covariance of errors with two distinct error distributions, and a sample size of **100**.

Error	**Model**	τ	**MBE**	**RMSE**	**MBE**	**RMSE**
			β_1	β_1	β_2	β_2
$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i)$	AQMM1	0.10	0.0111	0.2338	0.0015	0.0229
	AQMM2	0.10	0.0113	0.2338	0.0015	0.0229
	QSAM1	0.10	-0.7200	0.7293	0.0042	0.0450
	QSAM2	0.10	-0.7212	0.7306	0.0035	0.0451
	MTRUE	Mean	0.0151	0.2313	0.0000	0.0153
	AQMM1	0.50	0.0159	0.2320	0.0001	0.0179
	AQMM2	0.50	0.0160	0.2318	0.0000	0.0180
	QSAM1	0.50	-0.7114	0.7162	0.0014	0.0314
	QSAM2	0.50	-0.7150	0.7198	0.0012	0.0308
	AQMM1	0.90	0.0148	0.2327	0.0004	0.0223
	AQMM2	0.90	0.0147	0.2328	0.0005	0.0223
	QSAM1	0.90	-0.7050	0.7135	0.0005	0.0439
	QSAM2	0.90	-0.7137	0.7217	0.0004	0.0442
$\boldsymbol{\epsilon}_i \sim \mathcal{T}_{n_i,4}(\mathbf{0}, \boldsymbol{\Sigma}_i)$	AQMM1	0.10	-0.0102	0.2527	0.0013	0.0252
	AQMM2	0.10	-0.0102	0.2529	0.0013	0.0251
	QSAM1	0.10	-0.7218	0.7308	0.0028	0.0441
	QSAM2	0.10	-0.7214	0.7303	0.0023	0.0447
	MTRUE	Mean	-0.0121	0.2512	-0.0005	0.0163
	AQMM1	0.50	-0.0117	0.2506	-0.0008	0.0152
	AQMM2	0.50	-0.0116	0.2505	-0.0008	0.0152
	QSAM1	0.50	-0.6910	0.6966	0.0008	0.0295
	QSAM2	0.50	-0.6954	0.7012	0.0004	0.0295
	AQMM1	0.90	-0.0133	0.2566	-0.0017	0.0232
	AQMM2	0.90	-0.0132	0.2566	-0.0016	0.0232
	QSAM1	0.90	-0.7229	0.7326	-0.0011	0.0460
	QSAM2	0.90	-0.7273	0.7370	-0.0011	0.0457

Table 4.8: MBE and RMSE concerning the simulated data under Study 4.2, specifically focusing on the **unbalanced** data design, **heterogeneous** variance-covariance of errors with two distinct error distributions, and a sample size of **1000**.

Error	**Model**	τ	**MBE**	**RMSE**	**MBE**	**RMSE**
			β_1	β_1	β_2	β_2
$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i)$	AQMM1	0.10	0.0013	0.0732	0.0000	0.0070
	AQMM2	0.10	0.0013	0.0733	0.0001	0.0070
	QSAM1	0.10	-0.7268	0.7276	0.0009	0.0139
	QSAM2	0.10	-0.7276	0.7284	0.0009	0.0140
	MTRUE	Mean	0.0019	0.0718	-0.0002	0.0049
	AQMM1	0.50	0.0022	0.0726	-0.0002	0.0053
	AQMM2	0.50	0.0022	0.0726	-0.0002	0.0053
	QSAM1	0.50	-0.7221	0.7225	0.0003	0.0097
	QSAM2	0.50	-0.7245	0.7250	0.0004	0.0096
	AQMM1	0.90	0.0011	0.0723	0.0004	0.0068
	AQMM2	0.90	0.0011	0.0722	0.0004	0.0068
	QSAM1	0.90	-0.7219	0.7277	0.0006	0.0134
	QSAM2	0.90	-0.7266	0.7273	0.0006	0.0130
$\boldsymbol{\epsilon}_i \sim \mathcal{T}_{n_i,4}(\mathbf{0}, \boldsymbol{\Sigma}_i)$	AQMM1	0.10	0.0049	0.0747	0.0007	0.0074
	AQMM2	0.10	0.0049	0.0747	0.0007	0.0074
	QSAM1	0.10	-0.7326	0.7336	0.0013	0.0134
	QSAM2	0.10	-0.7335	0.7345	0.0012	0.0134
	MTRUE	Mean	0.0039	0.0732	0.0004	0.0048
	AQMM1	0.50	0.0037	0.0733	0.0002	0.0044
	AQMM2	0.50	0.0037	0.0733	0.0001	0.0044
	QSAM1	0.50	-0.6946	0.6952	0.0004	0.0093
	QSAM2	0.50	-0.6971	0.6976	0.0004	0.0093
	AQMM1	0.90	0.0031	0.0753	-0.0001	0.0077
	AQMM2	0.90	0.0030	0.0753	-0.0001	0.0077
	QSAM1	0.90	-0.7264	0.7274	0.0003	0.0135
	QSAM2	0.90	-0.7314	0.7324	0.0003	0.0134

From the findings presented in Tables 4.9 to 4.11, it is evident that the performance of the "Model" factor emerged as the most significant influence on predictive performance. Particularly, both AQMM1 and AQMM2 performed better in situations involving unbalanced data compared to balanced data. Each model performed well when errors were assumed to follow a normal distribution, suggesting that all models were sensitive to the distribution of error.

Table 4.9: ANOVA for three quantile models from AQMM with P-splines on MAE of Study 4.2

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	**Between subjects**				
	Error distribution (f(ϵ)) (Standard normal and Student's t)	1	0.00	102.95***	0.01
	Variance-covariance of errors (**R**) (Homogeneous and Heterogeneous)	1	0.12	17808.00***	0.69
	Data design (Balanced and Unbalanced data)	1	0.01	1473.50***	0.16
	Sample size (N = 100 and N = 1000)	1	0.00	515.30***	0.06
	Residuals	7995	0.48		
	Within subjects				
	Model (AQMM1, AQMM2, QSAM1, and QSAM2)	3	31.34	1512700.00***	0.99
	Model * f(ϵ)	3	0.00	0.70	0.00
	Model * R	3	0.01	497.07***	0.06
	Model * Data design	3	0.86	41588.00***	0.84
	Model * N	3	0.01	268.82***	0.03
	Residuals	23985	0.17		
0.50	**Between subjects**				
	Error distribution (f(ϵ)	1	0.10	4863.90***	0.38
	Variance-covariance of errors (**R**)	1	0.54	26510.00***	0.77
	Data design	1	0.07	3230.70***	0.29
	Sample size	1	0.00	186.72***	0.02
	Residuals	7995	1.54		
	Within subjects				
	Model	3	168.87	2764500.00***	1.00
	Model * f(ϵ)	3	4.19	0.01**	0.00
	Model * R	3	0.06	1022.50***	0.11
	Model * Data design	3	4.05	66351.00***	0.89
	Model * N	3	0.02	259.44***	0.03
	Residuals	23985	0.49		
0.90	**Between subjects**				
	Error distribution (f(ϵ))	1	0.00	77.26***	0.01
	Variance-covariance of errors (**R**)	1	0.12	17016.00***	0.68
	Data design	1	0.02	2440.00***	0.23
	Sample size	1	0.00	509.40***	0.06
	Residuals	7995	0.50		
	Within subjects				
	Model	3	32.76	1535400.00***	0.99
	Model * f(ϵ)	3	0.00	11.73***	0.00
	Model * R	3	0.01	494.17***	0.06
	Model * Data design	3	0.72	33545.00***	0.81
	Model * N	3	0.01	291.61***	0.04
	Residuals	23985	0.17		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Table 4.10: ANOVA for three quantile models from AQMM with P-splines on R-squared of Study 4.2

τ	Source of Variation	DF	SS	F	η^2
0.10	**Between subjects**				
	Error distribution (f(ϵ)) (Standard normal and Student's t)	1	0.00	12.54***	0.00
	Variance-covariance of errors ($\mathbf{R}$) (Homogeneous and Heterogeneous)	1	0.04	523.00***	0.06
	Data design (Balanced and Unbalanced data)	1	0.21	2849.40***	0.26
	Sample size ($N = 100$ and $N = 1000$)	1	0.00	38.12***	0.00
	Residuals	7995	9.52		
	Within subjects				
	Model (AQMM1, AQMM2, QSAM1, and QSAM2)	3	437.99	2024200.00***	1.00
	Model * f(ϵ)	3	0.00	16.21***	0.00
	Model * R	3	0.37	1691.60***	0.17
	Model * Data design	3	10.28	47507.00***	0.86
	Model * N	3	0.03	121.25***	0.01
	Residuals	23985	1.73		
0.50	**Between subjects**				
	Error distribution (f(ϵ)	1	0.00	45.14***	0.01
	Variance-covariance of errors ($\mathbf{R}$)	1	0.00	82.42***	0.01
	Data design	1	0.32	6471.40***	0.45
	Sample size	1	0.00	3.16	0.00
	Residuals	7995	5.70		
	Within subjects				
	Model	3	489.78	3307700.00***	1.00
	Model * f(ϵ)	3	0.24	1623.20***	0.17
	Model * R	3	0.62	4161.90***	0.34
	Model * Data design	3	9.75	65866.00***	0.89
	Model * N	3	0.02	113.67***	0.01
	Residuals	23985	1.18		
0.90	**Between subjects**				
	Error distribution (f(ϵ))	1	0.00	6.57*	0.00
	Variance-covariance of errors ($\mathbf{R}$)	1	0.00	10.95***	0.00
	Data design	1	0.31	3973.90***	0.33
	Sample size	1	0.00	31.81***	0.00
	Residuals	7995	9.99		
	Within subjects				
	Model	3	474.36	2028600.00***	1.00
	Model * f(ϵ)	3	0.01	59.22***	0.01
	Model * R	3	0.87	3733.90***	0.32
	Model * Data design	3	8.72	37278.00***	0.82
	Model * N	3	0.03	108.01***	0.01
	Residuals	23985	1.87		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Table 4.11: ANOVA for three quantile models on PNR of Study 4.2

τ	Source of Variation	DF	SS	F	η^2
0.10	**Between subjects**				
	Error distribution (f(ϵ)) (Standard normal and Student's t)	1	0.00	0.01	0.00
	Variance-covariance of errors (**R**) (Homogeneous and Heterogeneous)	1	0.00	1.24	0.00
	Data design (Balanced and Unbalanced data)	1	0.00	6.15*	0.00
	Sample size ($N = 100$ and $N = 1000$)	1	0.00	269.14***	0.03
	Residuals	7995	0.15		
	Within subjects				
	Model (AQMM1, AQMM2, QSAM1, and QSAM2)	3	0.03	903.27***	0.10
	Model * f(ϵ)	3	0.00	0.03	0.00
	Model * R	3	0.00	1.09	0.00
	Model * Data design	3	0.00	29.52***	0.00
	Model * N	3	0.00	439.46***	0.05
	Residuals	23985	0.23		
0.50	**Between subjects**				
	Error distribution (f(ϵ)	1	0.00	0.85	0.00
	Variance-covariance of errors (**R**)	1	0.00	2.81	0.00
	Data design	1	0.00	1.77	0.00
	Sample size	1	0.00	7.68**	0.00
	Residuals	7995	0.17		
	Within subjects				
	Model	3	0.01	334.14***	0.04
	Model * f(ϵ)	3	0.00	1.86	0.00
	Model * R	3	0.00	1.71	0.00
	Model * Data design	3	0.00	22.77***	0.00
	Model * N	3	0.01	217.27***	0.03
	Residuals	23985	0.26		
0.90	**Between subjects**				
	Error distribution (f(ϵ))	1	0.00	8.07**	0.00
	Variance-covariance of errors (**R**)	1	0.00	6.83**	0.00
	Data design	1	0.00	30.88***	0.00
	Sample size	1	0.01	906.52***	0.10
	Residuals	7995	0.14		
	Within subjects				
	Model	3	0.00	119.14***	0.01
	Model * f(ϵ)	3	0.00	0.92	0.00
	Model * R	3	0.00	0.23	0.00
	Model * Data design	3	0.00	1.05	0.00
	Model * N	3	0.00	31.05***	0.00
	Residuals	23985	0.24		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Summary

It is evident that when AQMM was integrated with linear predictors, its predictive performance remained efficient across all simulation scenarios, especially in the normal error case. This indicates that AQMM maintained efficient performance when dealing with linear predictors. It is important to note that AQMM is sensitive to the assumptions of the error distribution. Furthermore, while AQMM excelled in parameter estimation, QSAM struggled to estimate certain coefficients, particularly β_1. This discrepancy might be attributed to the influence of the first-order autoregressive (AR(1)) component.

4.4.3 Study 4.3

The simulation studies outlined in Study 4.1 and 4.2 were primarily centred on the broader context of LCGD. This framework entails that the growth of child, such as weight, depends only upon their preceding growth trajectory (autocorrelated process) and aligns with the population average (the mean function). Nevertheless, it is worth noting that real-world LCGD characteristics often encompass complexities beyond this simplified representation. For example, individual children may diverge in their initial measurements, like birth weight or height (intercept), and their varying growth rate over time (slope). This phenomenon, referred to as "between-individual differences", is a common facet. Furthermore, many longitudinal child development studies encounter deviations from the initially planned data collection schedule. Instances may arise where data collection is missed during certain age intervals, leading to deviations from the scheduled points. Consequently, these additional features/characteristics should be taken into account for a comprehensive understanding.

Aim

This simulation study was conducted to evaluate the precision of AQMM in estimating conditional quantile functions for child growth measurement. It focused on situations where longitudinal data incorporated supplementary attributes such as between-individual variations in baseline and growth trend, along with non-uniformly scheduled time points.

Data generation

The data were generated in accordance with the non-linear model provided by the equation below:

$$y_{ij} = \mu_{ij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{0i} + u_{1i} t_{ij} + \epsilon_{ij}. \tag{4.18}$$

Here, μ_{ij} is defined to align with the non-linear function (4.10), while β_1, β_2, x_{1ij}, and x_{2ij} retain the same specifications as presented in Study 4.2. In this context, the random individual effects are represented by $\mathbf{u} = (u_0, u_1)'$, denoting random intercept and random slope, respectively.

To generate the simulated data with autocorrelated and heteroscedastic observations, the random errors (ϵ_{ij}) were presumed to adhere to a heterogeneous exponential covariance structure. This structure, defined by parameters $\sigma^2 = 2$ and $\alpha = -0.50$, mirrors that of Study 4.1. Additionally, two distinct values of ϕ were taken into account: a medium value ($\phi = 1.45$) and a large value ($\phi = 4.48$). This choice was made to examine the influence of autocorrelation on the observations.

To exhibit the simulated data with between-individual differences, the random effects $\mathbf{u}$ were assumed the follow the distribution:

$$\mathbf{u} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{G}),$$

where

$$\mathbf{G} = \begin{pmatrix} 1 & -0.2 \\ -0.2 & 0.5 \end{pmatrix}.$$

In this context, the matrix $\mathbf{G}$ was constructed based on the understanding that the variance of random intercepts (σ_0^2) generally exceeds that of random slopes (σ_1^2) in the realm of child physical growth and development. Regarding the covariance value between random intercepts and random slopes (σ_{01}), it is common for children with substantial slopes to have small intercepts. As a result, this covariance is typically negative.

This simulation study was centered on two cases involving unbalanced data featuring unequally spaced time observations, a common occurrence in longitudinal datasets concerning child development:

1. Case 1: This case involved a specific arrangement of scheduled time points: 0, 0.06, 0.14, 0.21, 0.29, 0.36, 0.43, 0.57, 0.71, 0.86, 1. Consequently, the time intervals between observations were set at 0.07 for the initial six points and 0.14 for the subsequent ones. To simulate instances where observed time points deviate from the scheduled ones, the same methodology employed in Study 4.1 was utilised (as elaborated in the "unbalanced data" subsection). In this instance, the number of observations (n_i) per child varied between 2 and 11. Thus, this particular case effectively represented scenarios in which the collection of growth measurements and other associated variables encountered difficulties as age advanced.

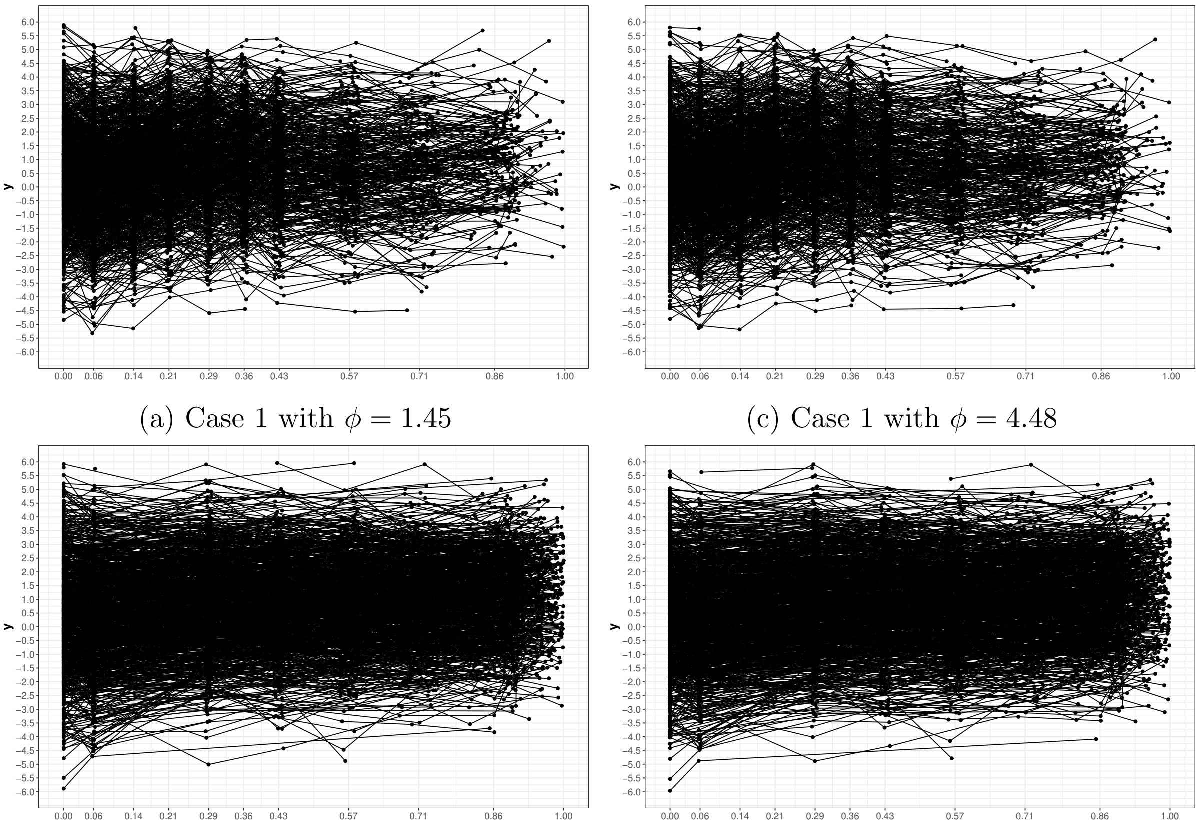
2. Case 2: The scheduled time points were defined similarly to Case 1, with the exception of excluding 0.14, 0.21, and 0.36. Consequently, for this case, the set of scheduled time points was 0, 0.06, 0.29, 0.43, 0.57, 0.71, 0.86, 1. This scenario reflects instances where growth measurements were not obtained at certain scheduled time points. Within this context, the number of observations (n_i) per child ranged from 2 to 8. Thus, this particular case was illustrative of situations in which the collection of growth measurements and associated variables encountered challenges at specific time points.

Figure 4.11 shows plots derived from some simulated data, while and Table 4.12 lists all the scenarios considered in this study.

Fitting the simulated data

For each of the eight simulated datasets, a comprehensive analysis was conducted involving the fitting of three distinct quantile models (with quantile levels of $\tau = 0.10, 0.50, 0.90$). The model employed for this purpose was represented as follows:

$$\mathbf{Q}_{y_{ij}|\mathbf{u}_i,\mathbf{t}_{ij},\mathbf{x}_i}(\tau) = \beta_{\tau,0} + g_\tau(t_{ij}) + \beta_{\tau,1}x_{1ij} + \beta_{\tau,2}x_{2ij} + u_{\tau,0i} + u_{\tau,1i}t_{ij}.$$



(a) Case 1 with $\phi = 1.45$

(c) Case 1 with $\phi = 4.48$

(b) Case 2 with $\phi = 1.45$

(d) Case 2 with $\phi = 4.48$

Figure 4.11: Example datasets of Study 4.3, generated from the model (4.18) using 1,000 children ($N = 1000$) with unequally spaced time observations.

Table 4.12: The scenarios used in Study 4.3

Scenario	**Data design**	**Autocorrelation coefficient (ϕ)**	**Sample size (N)**
1	Case 1	Medium, $\phi = 1.45$ (or $\Psi \approx 0.50$)	100
2		Medium, $\phi = 1.45$ (or $\Psi \approx 0.50$)	1000
3		Large, $\phi = 4.48$ (or $\Psi \approx 0.80$)	100
4		Large, $\phi = 4.48$ (or $\Psi \approx 0.80$)	1000
5	Case 2	Medium, $\phi = 1.45$ (or $\Psi \approx 0.50$)	100
6		Medium, $\phi = 1.45$ (or $\Psi \approx 0.50$)	1000
7		Large, $\phi = 4.48$ (or $\Psi \approx 0.80$)	100
8		Large, $\phi = 4.48$ (or $\Psi \approx 0.80$)	1000

In order to model the non-linear function (g_τ), the second order P-spline basis (cubic spline) with a discrete quadratic penalty on the basis coefficients was used. To save computational time when fitting the model, the number of knots was determined using the rule of thumb, $K = \max(5, \min(\tilde{N}/4, 35))$, where $\tilde{N}$ is the number of unique observed times and $\tilde{N}/4$ is the largest integer not exceeding $[\tilde{N}/4]$ (Ngo & Wand, 2004). To specify the variance-covariance matrix of $\mathbf{u}$, I employed a general positive-definite matrix (also known as "pdSymm" in `aqmm` package in R).

To assess the performance of each model, the same metrics (i.e. R-squared for QR, MAE, and PNR) were utilised as outlined in Section 4.4.1. Additionally, an analysis of variance (ANOVA) was conducted to examine the influence of the degree of the autocorrelation coefficient (ϕ) factor, data design, and sample size on each metric. Let Y_{ijkl} represent observations of each metric. The ANOVA model can be expressed as follows:

$$\begin{aligned} Y_{ijkl} =& \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{ik} + \epsilon_{ijkl}, \\ & l = 1, \ldots, 4000, i = 1, 2, j = 1, 2, k = 1, 2, \end{aligned}$$

where μ is the overall mean, α_i is the main effect of the i-th level of the degree of the autocorrelation coefficient factor ($\phi = 1.45$ and $\phi = 4.48$), β_i is the main effect of the j-th level of the data design factor (Case1 and Case2), γ is the main effect of the k-th level of the sample size factor ($N = 100$ and $N = 1,000$). Here, $\alpha\beta)_{ij}, (\alpha\gamma)_{ik}, (\beta\gamma)_{ik}$, are the two-way interaction effects between these factors, and ϵ_{ijkl} is the random error term.

Results

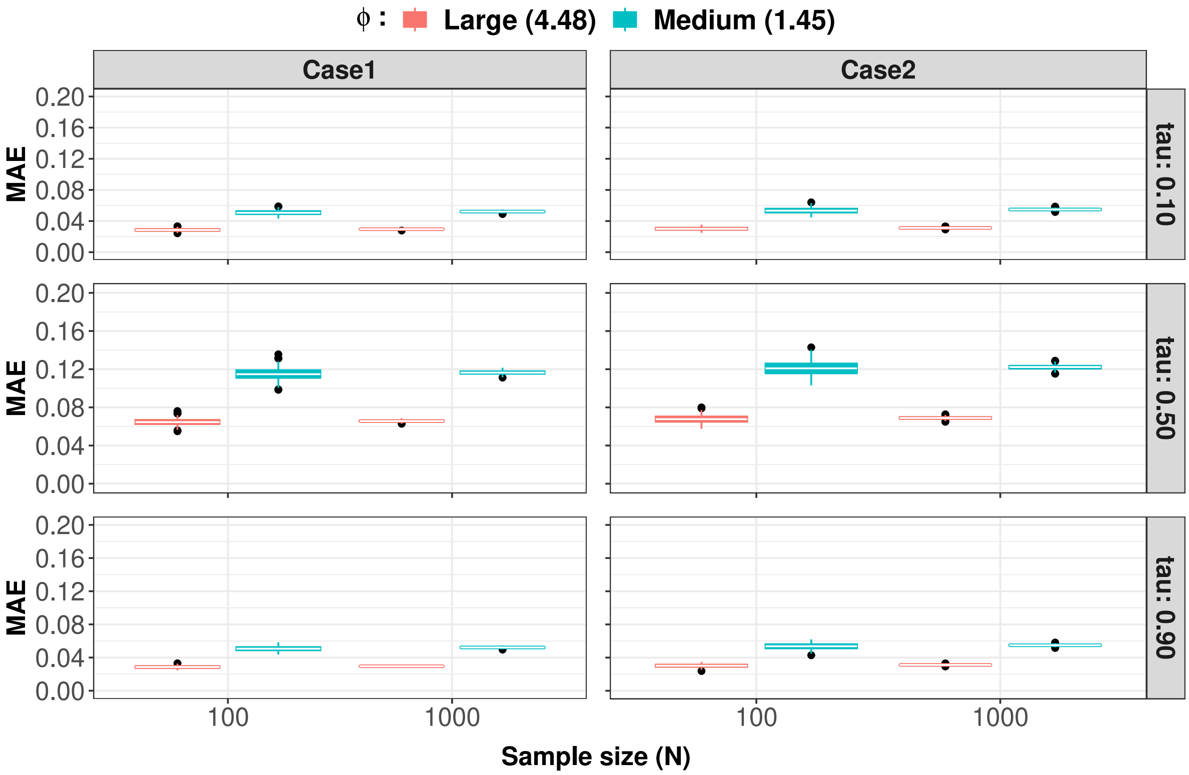


Figure 4.12: The MAE of the AQMM approach with P-splines in Study 4.3. The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

Figure 4.12 presents the MAE values of the AQMM approach with P-splines across three quantiles, two degrees of autocorrelation, two data designs, and two sample sizes. The results indicate that both extreme quantile models (the 0.10th and 0.90th quantile models) provided similar MAE values and trends, whereas the 0.50th quantile models yielded higher MAE values than the two extreme quantile models. When comparing scenarios with different degrees of autocorrelation, the AQMM model provided smaller MAE values for the large degree of autocorrelation compared to the medium degree across all three quantile models, data designs, and sample sizes. This suggests that when the data exhibit highly correlated measurements, the random effects part of the AQMM model can effectively address such correlations.

Regarding the two distinct data designs (Case 1 and 2), this model exhibited consistent performance, displaying similar values and trends in terms of MAE across three quantile models. This indicates that the AQMM model with P-splines can be effective in capturing the characteristics of these two unbalanced datasets featuring unequally spaced time observations.

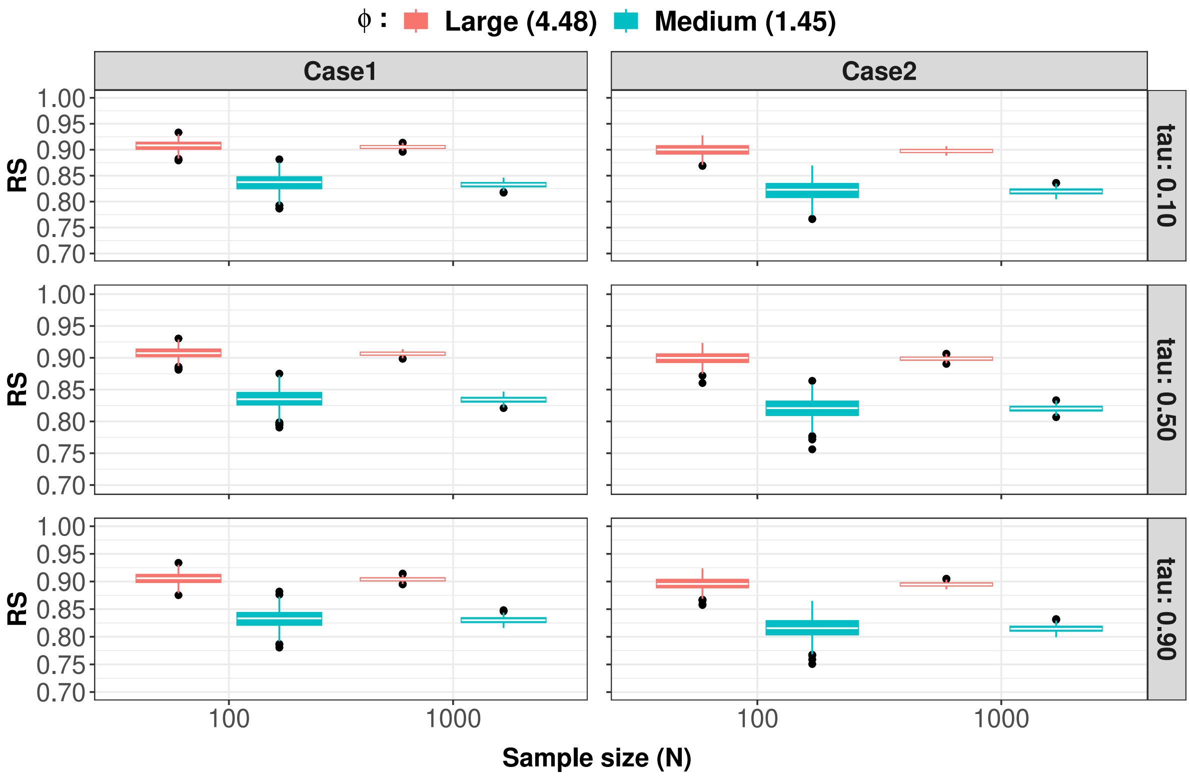


Figure 4.13: The RS of the AQMM approach with P-splines in Study 4.3. The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

Figure 4.13 displays the RS values of the AQMM approach with P-splines across three quantiles, two degrees of autocorrelation, two data designs, and two sample sizes. It is evident that the AQMM approach exhibited similar RS values and trends across all three quantile models. Specifically, within the context of two distinct degrees of autocorrelation, the AQMM model yielded higher RS values in the large degree compared to the medium degree. This suggests a similarity to the MAE values mentioned previously. When considering the two data designs, the RS values from the AQMM model with P-splines were relatively similar in both data designs.

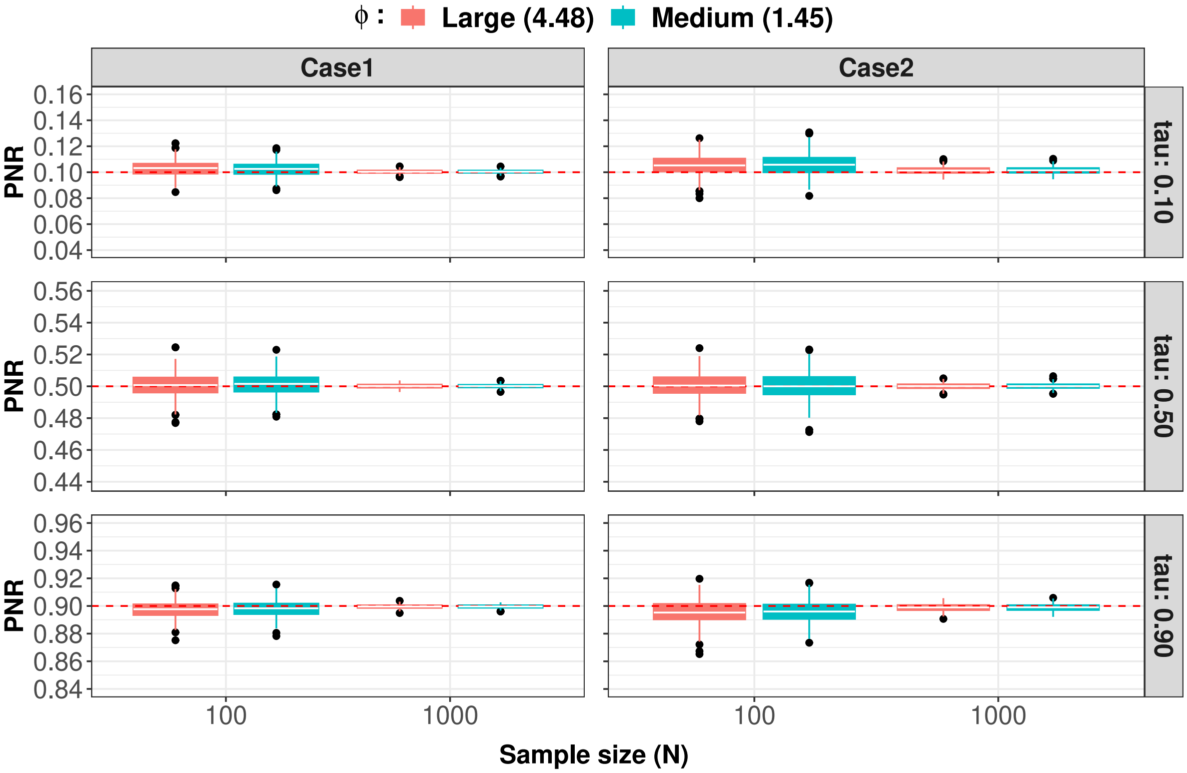


Figure 4.14: The PNR of the AQMM approach with P-splines in Study 4.3. The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively. The red dashed lines represent the expected quantile levels, $\tau = 0.10, 0.50$, and 0.90, respectively.

Figure 4.14 shows the PNR values of the AQMM approach with P-splines across three quantiles, two degrees of autocorrelation, two data designs, and two sample sizes. It appears that the three quantile models from the AQMM approach tended to provide PNR values close to their respective quantile levels across both degrees of autocorrelation, data designs, and sample sizes. Specifically, when considering the two sample sizes, the 0.50th quantile model provided average PNR values close to the expected quantile level at 0.50 for both sample sizes. However, when the sample size was small, the PNR values of the two extreme quantile models deviated from their expected quantile levels at 0.10 and 0.90, respectively. This suggests that the AQMM model is particularly sensitive to sample sizes, especially for extreme quantiles.

Furthermore, an ANOVA on MAE, RS and PNR, as presented in Tables 4.13 to 4.15, indicates that both the degrees of autocorrelation and the data design predominantly influenced AQMM's performance on MAE and RS, whereas the sample size factor was the main influence on PNR.

Table 4.13: ANOVA for three quantile models from AQMM with P-splines on MAE of Study 4.3

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	Degree of the autocorrelation coefficient ($\phi = 1.45$ and $\phi = 4.48$)	1	0.53	141800.00***	0.97
	Data design (Case1 and Case2)	1	0.00	1219.00***	0.23
	Sample size ($N = 100$ and $N = 1000$)	1	0.00	416.40***	0.09
	ϕ * Data design	1	0.00	87.69***	0.02
	ϕ * N	1	0.00	11.12***	0.00
	Data design * N	1	0.00	0.34	0.00
	Residual	3993	0.01		
0.50	Degree of the autocorrelation coefficient (ϕ)	1	2.70	176100.00***	0.98
	Data design	1	0.02	1310.00***	0.25
	Sample size (N)	1	0.00	105.20***	0.03
	ϕ * Data design	1	0.00	123.00***	0.03
	ϕ * N	1	0.00	0.97	0.00
	Data design * N	1	0.00	0.15	0.00
	Residual	3993	0.06		
0.90	Degree of the autocorrelation coefficient (ϕ)	1	0.53	148200.00***	0.97
	Data design	1	0.00	1363.00***	0.25
	Sample size (N)	1	0.00	390.20***	0.09
	ϕ * Data design	1	0.00	98.94***	0.02
	ϕ * N	1	0.00	9.59**	0.00
	Data design * N	1	0.00	0.65	0.00
	Residual	3993	0.01		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Table 4.14: ANOVA for three quantile models from AQMM with P-splines on R-squared of Study 4.3

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	Degree of the autocorrelation coefficient (ϕ) ($\phi = 1.45$ and $\phi = 4.48$)	1	5.63	51023.78***	0.93
	Data design (Case1 and Case2)	1	0.12	1041.11***	0.21
	Sample size ($N = 100$ and $N = 1000$)	1	0.01	55.90***	0.01
	ϕ * Data design	1	0.01	74.46***	0.02
	ϕ * N	1	0.00	0.87	0.00
	Data design * N	1	0.00	1.40	0.00
	Residual	3993	0.44		
0.50	Degree of the autocorrelation coefficient (ϕ)	1	5.70	69516.49***	0.95
	Data design	1	0.13	1522.32***	0.28
	Sample size (N)	1	0.00	3.88*	0.00
	ϕ * Data design	1	0.01	131.23***	0.03
	ϕ * N	1	0.00	0.16	0.00
	Data design * N	1	0.00	1.47	0.00
	Residual	3993	0.33		
0.90	Degree of the autocorrelation coefficient (ϕ)	1	5.91	52331.34***	0.93
	Data design	1	0.16	1433.57***	0.26
	Sample size (N)	1	0.00	14.92***	0.00
	ϕ * Data design	1	0.00	103.78***	0.03
	ϕ * N	1	0.00	0.02	0.00
	Data design * N	1	0.00	2.16	0.00
	Residual	3993	0.45		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Table 4.15: ANOVA for the three quantile models from AQMM with P-splines on PNR of Study 4.3

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	Degree of the autocorrelation coefficient (ϕ) ($\phi = 1.45$ and $\phi = 4.48$)	1	0.00	0.01	0.00
	Data design (Case1 and Case2)	1	0.00	137.04***	0.03
	Sample size ($N = 100$ and $N = 1000$)	1	0.01	424.48***	0.10
	ϕ * Data design	1	0.00	2.60	0.00
	ϕ * N	1	0.00	0.21	0.00
	Data design * N	1	0.00	34.11***	0.01
	Residual	3993	0.10		
0.50	Degree of the autocorrelation coefficient (ϕ)	1	0.00	0.99	0.00
	Data design	1	0.00	2.11	0.00
	Sample size (N)	1	0.00	10.41**	0.00
	ϕ * Data design	1	0.00	0.52	0.00
	ϕ * N	1	0.00	0.62	0.00
	Data design * N	1	0.00	1.13	0.00
	Residual	3993	0.11		
0.90	Degree of the autocorrelation coefficient (ϕ)	1	0.00	2.41	0.00
	Data design	1	0.00	65.46***	0.02
	Sample size (N)	1	0.00	246.39***	0.06
	ϕ * Data design	1	0.00	0.11	0.00
	ϕ * N	1	0.00	2.06	0.00
	Data design * N	1	0.00	14.17***	0.00
	Residual	3993	0.10		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Summary

Within the context of LCGD, characterised by heteroscedasticity, correlated within-child errors, discrepancies in initial measurements, and temporal variations in individual growth rates, AQMM consistently showcased resilient performance. The random effects component in this model effectively addresses correlated within-child errors or the correlation structure in repeated growth measurements, such as the CAR(1) structure. Furthermore, modelling AQMM with P-splines exhibited notable proficiency in accommodating LCGD characterised by non-uniformly scheduled temporal points.

4.4.4 Study 4.4

In theoretical terms, the mixed model framework employs the incorporation of random errors to capture within-individual variability. These random errors conform to a prior distribution, often characterised by a zero-centred normal distribution, complete with a predetermined variance-covariance structure. Typically, it is assumed that each individual shares an identical residual variance (σ^2). However, in practice, additional systematic differences contribute to variations in within-individual attributes across individuals. This phenomenon is commonly referred to as "between-individual differences in intra-individual variation". In the context of LCGD, such variations may emerge, resulting in data that have this variation with common features such as between-individual differences in intercept and trend, and autocorrelation. Hence, it is reasonable to investigate this aspect when applying the AQMM approach to the fitting of LCGD.

Aim

This simulation study was conducted with the purpose of assessing the precision with which AQMM can estimate conditional quantile functions concerning child growth measurement withing the context of longitudinal data that incorporates the manifestation of between-individual differences in intra-individual variation.

Data generation

The data were generated from the model (4.18), but each child was allowed to have his/her own residual variance by assuming the residual errors to follow

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma_i^2 \mathbf{R_i}), \tag{4.19}$$

where

$$\sigma_i^2 = \exp(\delta_0 + \omega_i). \tag{4.20}$$

Here, $\delta_0 = 0.69$ signifies the random intercept model for σ_i^2. This setting implies an average residual variance $(\bar{\sigma}^2)$ of 2. The term $\omega_i \sim N(0, \sigma_\omega^2)$ represents an individual-specific random effect for the residual variance, where σ_ω^2 is its variance (Hedeker et al., 2008).

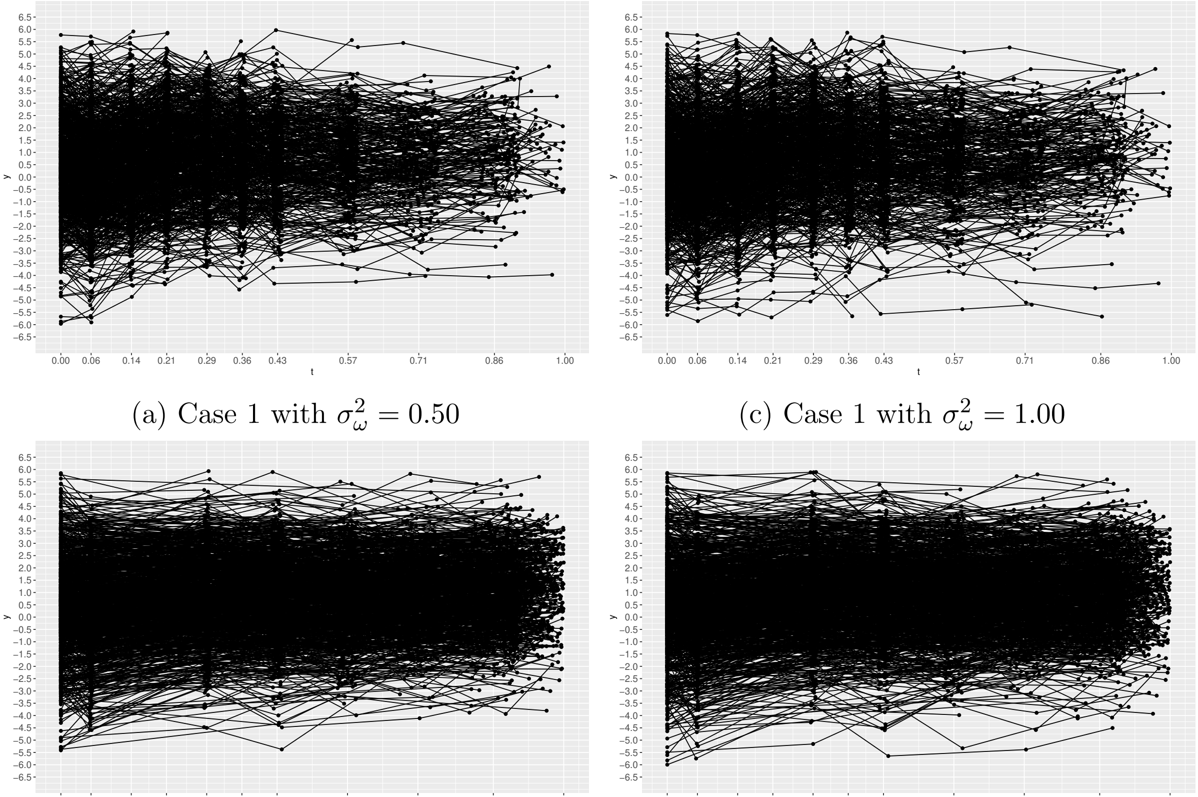
In order to investigate the model's performance under different data settings, the variance of ω_i was set to two different values: 0.5 and 1.0. The former indicates an initial dispersion of individual-specific random effects around the mean, while the latter represents a doubling of this dispersion to assess its impact on the model's performance. Both $\mathbf{C}_i$ and $\mathbf{\Phi}_i$ were defined to correspond to (4.12) and (4.13), respectively, to hold the heterogeneous exponential covariance structure. In this case, each child was assumed to have the identical structure of autocorrelation, where $\phi = 1.45$. Figure 4.15 shows plots derived from some simulated data, while and Table 4.16 lists all the scenarios considered in this study.

To fit the simulated data, the same model of Study 4.3 was applied, i.e. AQMM with a cubic P-spline. Each simulated dataset was fitted to follow the same processes that mentioned in Section 4.4.3.

To assess the performance of the model, the same metrics (i.e. R-squared for QR, MAE, and PNR) were used. Furthermore, the analysis of variance (ANOVA) was conducted to examine the influence of three factors (i.e. "Variance of ω_i (σ_ω^2)", "Data design", "Sample size (N)" on each metric. Let each metric is denoted by Y_{ijkl}, where Y_{ijkl} is the l-th observation at the i-th level of the variance of ω_i (σ^2_ω) factor, the j-th level of the data design factor, and the k-th level of the sample size factor. The ANOVA model can be written as:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{ik} + \epsilon_{ijkl},$$

where μ is the overall mean, α_i is the main effect of the i-th level of the variance of ω_i factor ($\sigma^2_\omega = 0.50$ and $\sigma^2_\omega = 1.00$), β_i is the main effect of the j-th level of the data design factor (Case1 and Case2), γ is the main effect of the k-th level of the sample size factor ($N = 100$ and $N = 1,000$), $\alpha\beta)_{ij}, (\alpha\gamma)_{ik}, (\beta\gamma)_{ik}$, are the interaction effects, and ϵ_{ijkl} is the random error term, assumed to be normally distributed with mean 0 and variance σ^2.



(a) Case 1 with $\sigma^2_\omega = 0.50$

(c) Case 1 with $\sigma^2_\omega = 1.00$

(b) Case 2 with $\sigma^2_\omega = 0.50$

(d) Case 2 with $\sigma^2_\omega = 1.00$

Figure 4.15: Example datasets of Study 4.4, generated from models (4.18) and (4.20) using 1,000 children ($N = 1000$) with unequally spaced time observations.

Table 4.16: The scenarios used in Study 4.4

Scenario	Data design	Variance of ω_i (σ^2_ω)	Sample size (N)
1	Case 1	$\sigma^2_\omega = 0.50$	100
2		$\sigma^2_\omega = 0.50$	1000
3		$\sigma^2_\omega = 1.00$	100
4		$\sigma^2_\omega = 1.00$	1000
5	Case 2	$\sigma^2_\omega = 0.50$	100
6		$\sigma^2_\omega = 0.50$	1000
7		$\sigma^2_\omega = 1.00$	100
8		$\sigma^2_\omega = 1.00$	1000

Results

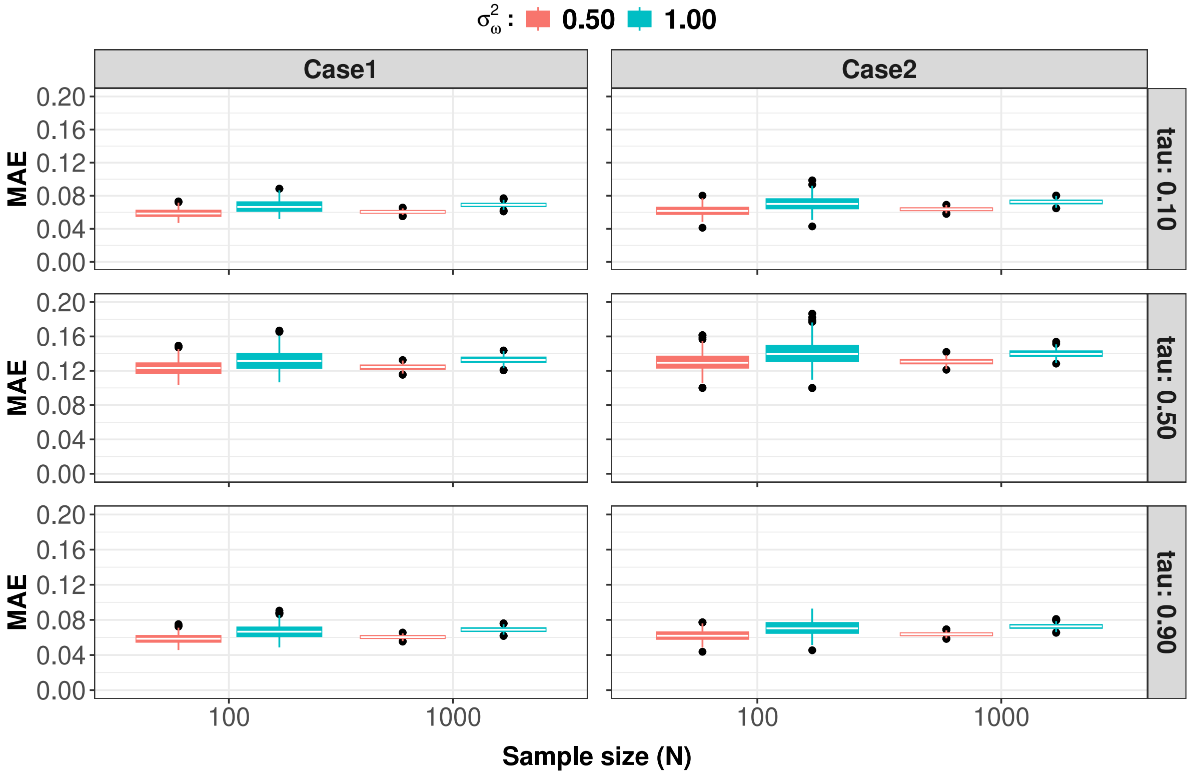


Figure 4.16: The MAE of the AQMM approach with P-splines in Study 4.4. The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

Figure 4.16 illustrates that, even with between-individual differences in intra-individual variation within LCGD, the two extreme quantile models (the 0.10 and 0.90th quantiles) from the AQMM with P-splines consistently yielded similar MAE values. This similarity in MAE values indicates that the AQMM approach provides relatively consistent predictions across different segments of the data distribution. In contrast, the MAE values of the

0.50the quantile model were explicitly higher than those of the two extreme quantile models. When considering the scenario of two distinct values of the variance of individual-specific random effect for the residual variance (σ^2_ω), the MAE values in the scenario of $\sigma^2_\omega = 0.50$ tended to be smaller than in the case of $\sigma^2_\omega = 1.0$ across two data designs and two sample sizes. Furthermore, compared to the scenarios of LCGD with no between-individual differences in intra-individual variation in Study 4.3, the MAE values were higher across all scenarios. This indicates that the AQMM model is sensitive to other between-individual differences within LCGD, as expected. Regarding the two distinct data designs, the AQMM seemed to yield slightly smaller MAE values in Case 1 compared to Case 2. When the sample size increased to be large ($N = 1,000$), the AQMM model provided higher precision in MAE values compared to a small sample size across all other scenarios.

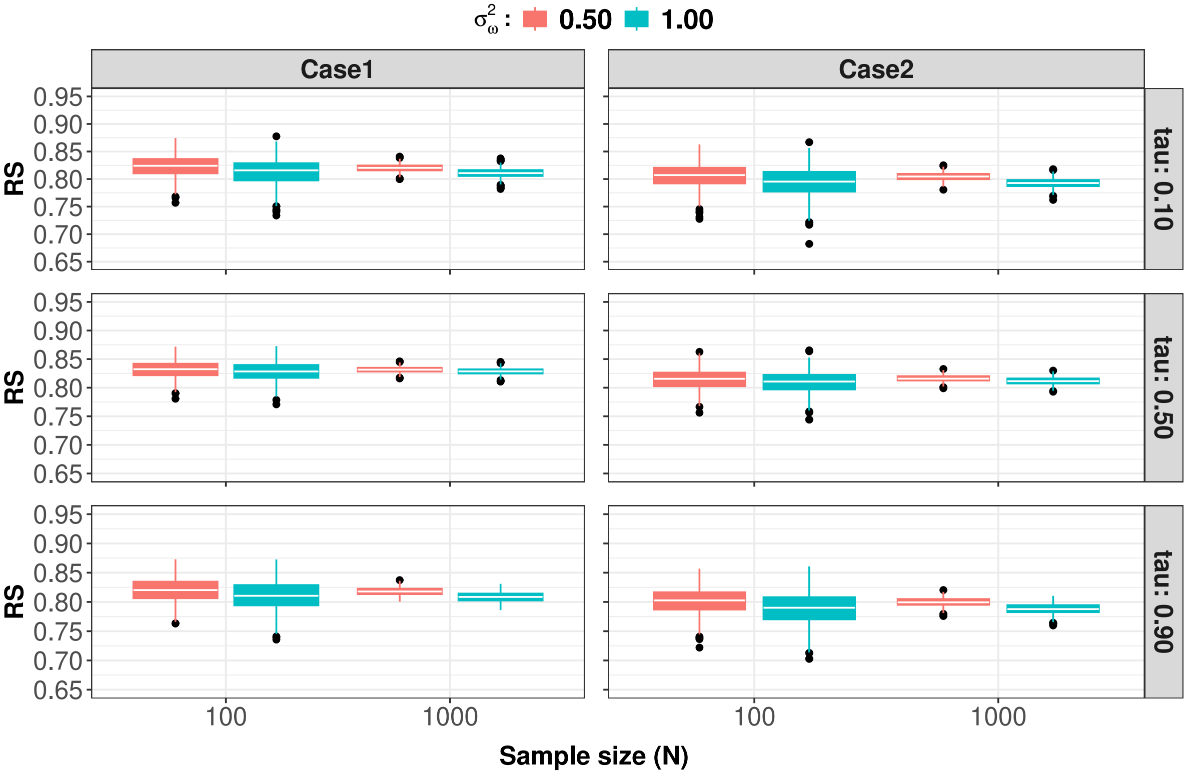


Figure 4.17: The RS of the AQMM approach with P-splines in Study 4.4. The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

Figure 4.17 shows that the 0.10 and 0.90th quantile models from the AQMM with P-splines relatively yielded similar RS values. However, the 0.50the quantile model provided slightly higher RS values compared to those of the two extreme quantile models. The RS values in the scenario of $\sigma^2_\omega = 0.50$ were slightly higher than in the case of $\sigma^2_\omega = 1.0$ across two data designs and two sample sizes. When considering two distinct data design, the RS values from three quantile models in Case1 appeared to be slightly smaller than in Case2. Increasing the sample sizes to a large provide more precision in RS values than the small sample size for all other scenarios.

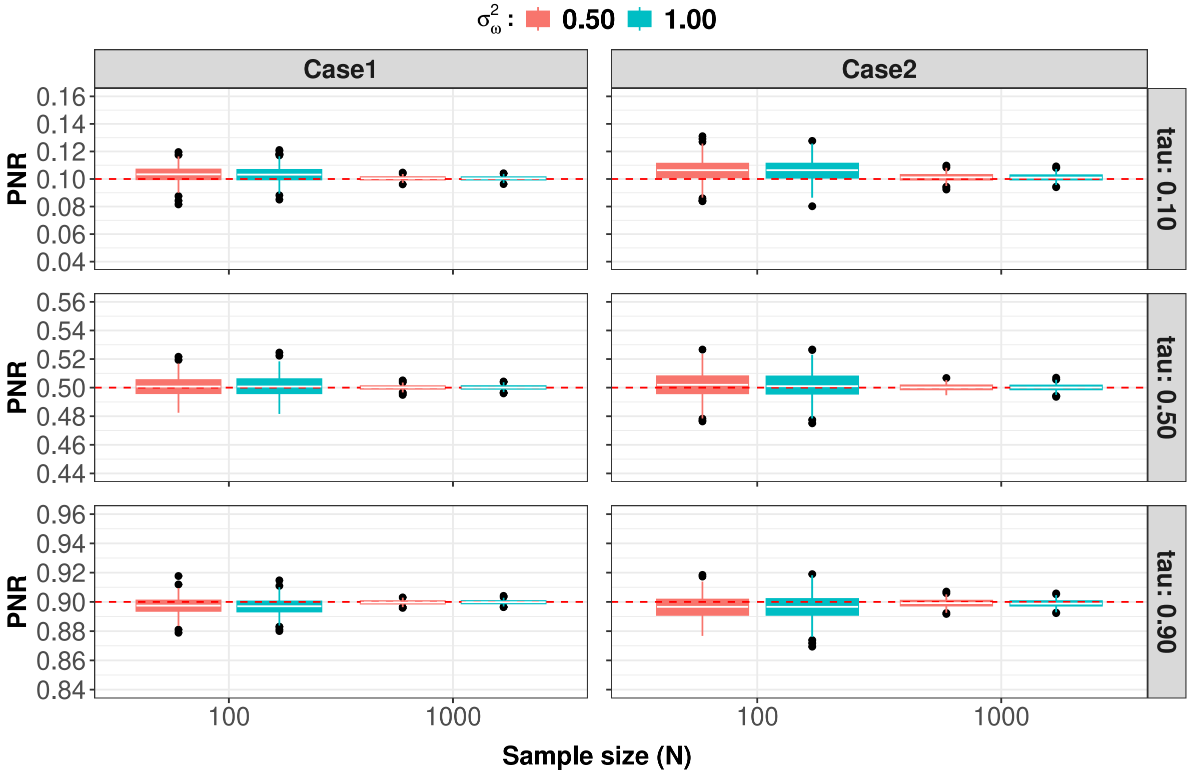


Figure 4.18: The PNR of the AQMM approach with P-splines in Study 4.4. The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

Figure 4.18 shows that the three quantile models from the AQMM approach with P-splines yielded the PNR values close to their expected quantile level at 0.10, 0.50, and 0.90 across all scenarios. However, in the case of a small sample size, both extreme quantile models (the 0.10 and 0.50 quantiles) provided average PNR values that deviated relatively from their expected quantile levels. This trend held true for all scenarios.

Clearly, Tables 4.17 to 4.19 show that the variance of the individual-specific random effect for the residual (σ^2_ω) was the main factor impacting to the predictive performance of AQMM with P-splines, especially in terms of MAE. Meanwhile, the data design was the main factor affecting the R-squared value. However, this scenario did not affect the PNR (Table 4.18).

Table 4.17: ANOVA for three quantile models from AQMM with P-splines on MAE of Study 4.4

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	Variance of ω_i ($\sigma^2_\omega = 0.50$ and $\sigma^2_\omega = 1.00$)	1	0.07	3187.65***	0.44
	Data design (Case1 and Case2)	1	0.01	473.74***	0.11
	Sample size ($N = 100$ and $N = 1000$)	1	0.00	135.43***	0.03
	σ^2_ω * Data design	1	0.00	2.250	0.00
	σ^2_ω * N	1	0.00	0.55	0.00
	Data design * N	1	0.00	0.05	0.00
	Residual	3993	0.09		
0.50	Variance of ω_i (σ^2_ω)	1	0.08	1283.02***	0.24
	Data design	1	0.05	825.58***	0.17
	Sample size (N)	1	0.00	2.54	0.00
	σ^2_ω * Data design	1	0.00	4.03*	0.00
	σ^2_ω * N	1	0.00	2.23	0.00
	Data design * N	1	0.00	4.11*	0.00
	Residual	3993	0.26		
0.90	Variance of ω_i (σ^2_ω)	1	0.07	3160.18***	0.44
	Data design	1	0.01	540.47***	0.12
	Sample size (N)	1	0.00	114.03***	0.03
	σ^2_ω * Data design	1	0.00	2.74	0.00
	σ^2_ω * N	1	0.00	0.10	0.00
	Data design * N	1	0.00	0.83	0.00
	Residual	3993	0.09		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Table 4.18: ANOVA for three quantile models from AQMM with P-splines on R-squared of Study 4.4

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	Variance of ω_i ($\sigma^2_\omega = 0.50$ and $\sigma^2_\omega = 1.00$)	1	0.12	398.53***	0.09
	Data design (Case1 and Case2)	1	0.29	977.30***	0.20
	Sample size ($N = 100$ and $N = 1000$)	1	0.00	9.60**	0.00
	σ^2_ω * Data design	1	0.00	3.76	0.00
	σ^2_ω * N	1	0.00	0.08	0.00
	Data design * N	1	0.00	0.44	0.00
	Residual	3993	1.19		
0.50	Variance of ω_i (σ^2_ω)	1	0.02	107.37***	0.03
	Data design	1	0.29	1755.14***	0.31
	Sample size (N)	1	0.00	9.17**	0.00
	σ^2_ω * Data design	1	0.00	4.02*	0.00
	σ^2_ω * N	1	0.00	1.73	0.00
	Data design * N	1	0.00	4.74*	0.00
	Residual	3993	0.66		
0.90	Variance of ω_i (σ^2_ω)	1	0.12	375.39***	0.09
	Data design	1	0.39	1219.94***	0.23
	Sample size (N)	1	0.00	3.26	0.00
	σ^2_ω * Data design	1	0.00	4.84*	0.00
	σ^2_ω * N	1	0.00	0.46	0.00
	Data design * N	1	0.00	0.30	0.00
	Residual	3993	1.29		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Table 4.19: ANOVA for three quantile models from AQMM with P-splines on PNR of Study 4.4

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	Variance of ω_i ($\sigma^2_\omega = 0.50$ and $\sigma^2_\omega = 1.00$)	1	0.00	0.02	0.00
	Data design (Case1 and Case2)	1	0.00	141.37***	0.03
	Sample size ($N = 100$ and $N = 1000$)	1	0.02	624.22***	0.14
	σ^2_ω * Data design	1	0.00	0.16	0.00
	σ^2_ω * N	1	0.00	0.30	0.00
	Data design * N	1	0.00	49.55***	0.01
	Residual	3993	0.10		
0.50	Variance of ω_i (σ^2_ω)	1	0.00	0.12	0.00
	Data design	1	0.00	3.32	0.00
	Sample size (N)	1	0.00	33.71***	0.01
	σ^2_ω * Data design	1	0.00	1.68	0.00
	σ^2_ω * N	1	0.00	0.12	0.00
	Data design * N	1	0.00	2.60	0.00
	Residual	3993	0.14		
0.90	Variance of ω_i (σ^2_ω)	1	0.00	0.10	0.00
	Data design	1	0.00	22.59***	0.01
	Sample size (N)	1	0.01	292.12***	0.07
	σ^2_ω * Data design	1	0.00	0.35	0.00
	σ^2_ω * N	1	0.00	0.09	0.00
	Data design * N	1	0.00	0.17	0.00
	Residual	3993	0.09		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Summary

In scenarios where LGCD exhibited between-individual differences in intra-individual variation, AQMM's performance remained satisfactory, with MAE approaching to zero, R-squared values close to one, and PNR closely aligned with the quantile levels, particularly when the intra-individual variation was at a medium level. However, its performance appeared to decline when LGCD demonstrated a high level of this variation.

4.4.5 Study 4.5

In real-world datasets, children exhibit variation not only in residual errors but also in terms of autocorrelation patterns. This heterogeneity implies that some children might demonstrate a higher or lower degree of the AR(1) process compared to others, influenced by individual circumstances. Consequently, this characteristic presents another essential consideration when analysing or modelling LCGD.

Aim

The objective of this simulation study was to evaluate the accuracy of AQMM in estimating conditional quantile functions for child growth measurements in the presence of both between-individual differences in intra-individual variation and autocorrelation within longitudinal data.

Data generation

The dataset was generated according to the model (4.18), with the additional consideration that each child had unique residual variance and autocorrelation patterns. This was achieved by assuming the residual errors followed distribution (4.19) and (4.20), and by specifying an individual exponential correlation structure as follows:

$$C_{i,j} = \exp(-s_{ij}/\phi_i), \quad \phi_i > 0, \tag{4.21}$$

where

$$\phi_i = \exp(\zeta_0 + \pi_i). \tag{4.22}$$

Here, $\zeta_0 = 0.37$ denotes a random intercept of ϕ_i. This setting implies an average autocorrelation coefficient ($\bar{\phi}$) of 1.45. In other words, this is equivalent to the continuous first-order autocorrelation coefficient of 0.50. The term $\pi_i \sim N(0, \sigma^2_\pi)$ represents an individual-specific random effect for the autocorrelation, where σ^2_π is its variance.

In order to investigate how well the model performed with different data settings, the variance of ω_i and π_i was adjusted to two different situations: 0.50 and 0.10 (as shown in Table 4.20). The adjustment was made to examine how the model behaves when the dispersion of both σ^2_i and ϕ_i is altered by half.

Table 4.20: The scenarios used in Study 4.5

Scenario	Data design	Variance of ω_i (σ^2_ω)	Variance of π_i (σ^2_π)	Sample size (N)
1	Case 1	$\sigma^2_\omega = 0.50$	$\sigma^2_\pi = 0.50$	100
2			$\sigma^2_\pi = 1.00$	100
3			$\sigma^2_\pi = 0.50$	1000
4			$\sigma^2_\pi = 1.00$	1000
5		$\sigma^2_\omega = 1.00$	$\sigma^2_\pi = 0.50$	100
6			$\sigma^2_\pi = 1.00$	100
7			$\sigma^2_\pi = 0.50$	1000
8			$\sigma^2_\pi = 1.00$	1000
9	Case 2	$\sigma^2_\omega = 0.50$	$\sigma^2_\pi = 0.50$	100
10			$\sigma^2_\pi = 1.00$	100
11			$\sigma^2_\pi = 0.50$	1000
12			$\sigma^2_\pi = 1.00$	1000
13		$\sigma^2_\omega = 1.00$	$\sigma^2_\pi = 0.50$	100
14			$\sigma^2_\pi = 1.00$	100
15			$\sigma^2_\pi = 0.50$	1000
16			$\sigma^2_\pi = 1.00$	1000

To fit the simulated data, the same model of Study 4.3 were applied, i.e. AQMM with a cubic P-spline. Each simulated dataset was fitted to follow the same processes that mentioned in Section 4.4.3.

The performance of each model was evaluated using the same metrics (i.e. R-squared for

QR, MAE, and PNR) as described in Section 4.4.1.

Results

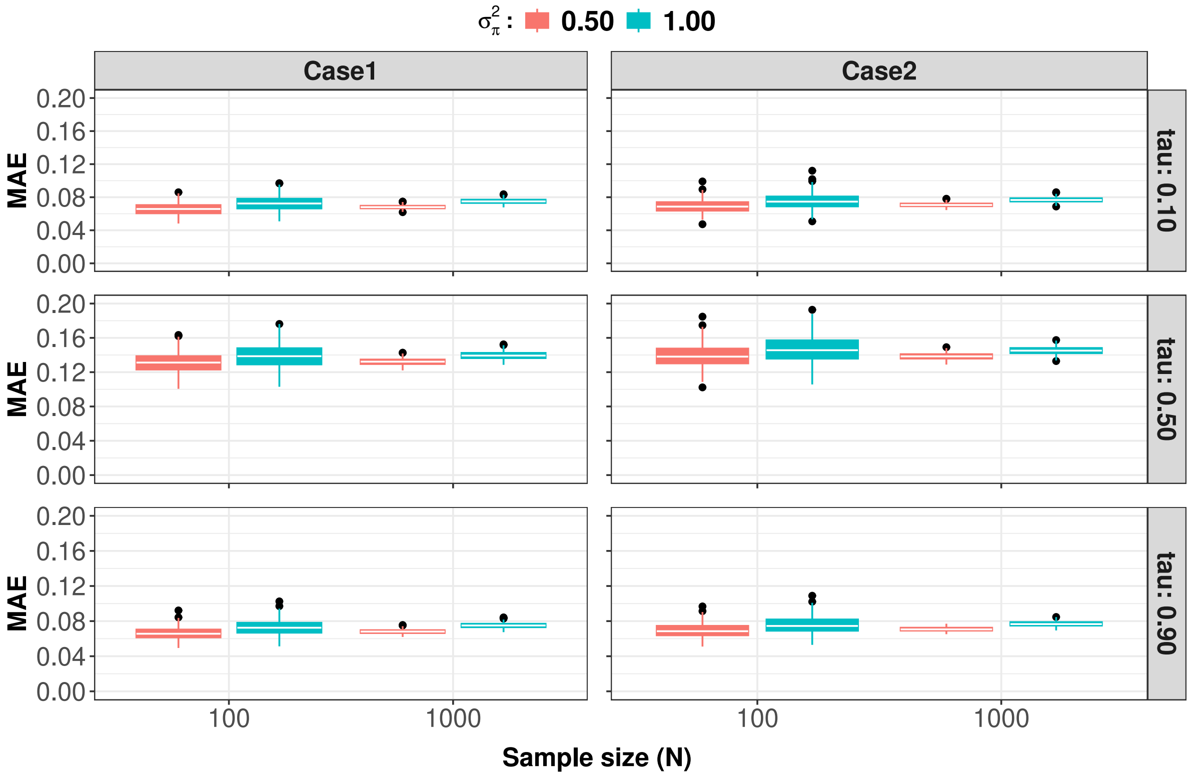


Figure 4.19: The MAE of the AQMM approach with P-splines in Study 4.5 when the simulated data had a medium variance of ω_i ($\sigma_\omega^2 = 0.50$). The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

In the scenario where the variance of ω_i (σ_ω^2) was fixed at 0.50 and the variance of π_i (σ_π^2) was varied to be 0.50 and 1.00, Figure 4.19 shows that the MAE values of the two extreme quantile models (the 0.10th and 0.90th quantiles) were relatively similar, indicating that the AQMM approach provides consistent prediction across different segments of the data distribution. However, the 0.50th quantile model provided higher MAE values than those of the two extreme quantile. The MAE values in the scenario of $\sigma_\pi^2 = 0.50$ tended to be smaller than $\sigma_\pi^2 = 1.00$. This trend was consistent across all scenarios and three quantile levels, suggesting that the AQMM approach is sensitive to this feature in LCGD.

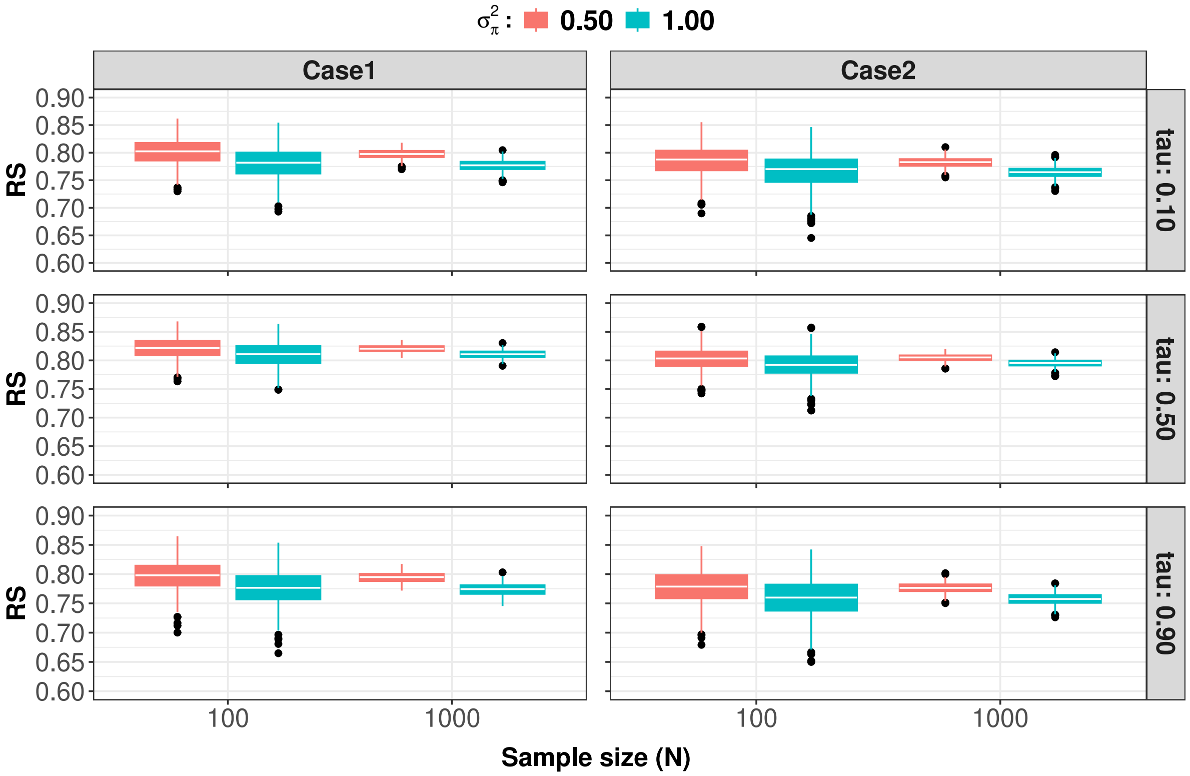


Figure 4.20: The RS of the AQMM approach with P-splines in Study 4.5 when the simulated data had a medium variance of ω_i ($\sigma^2_\omega = 0.50$). The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

In term of R-squared, Figure 4.20 demonstrates that the two extreme quantile models yielded the RS values that were relatively similar, whereas the 0.50th quantile model provided slightly higher RS values compared to the two extreme quantile models. The RS values declined when the variance of π_i (σ^2_π) increased to 1.00, indicating that this feature in LCGD impacts to the performance of AQMM.

When considering the PNR metric, Figure 4.21 shows that the AQMM approach with P-splines still provided the PNR values as expected. The PNR values of the three quantile models were close to their expected quantile levels in all scenarios. However, in the case of the two extreme quantile models with a small sample size, the average PNR values appeared to deviate from the expected quantile levels.

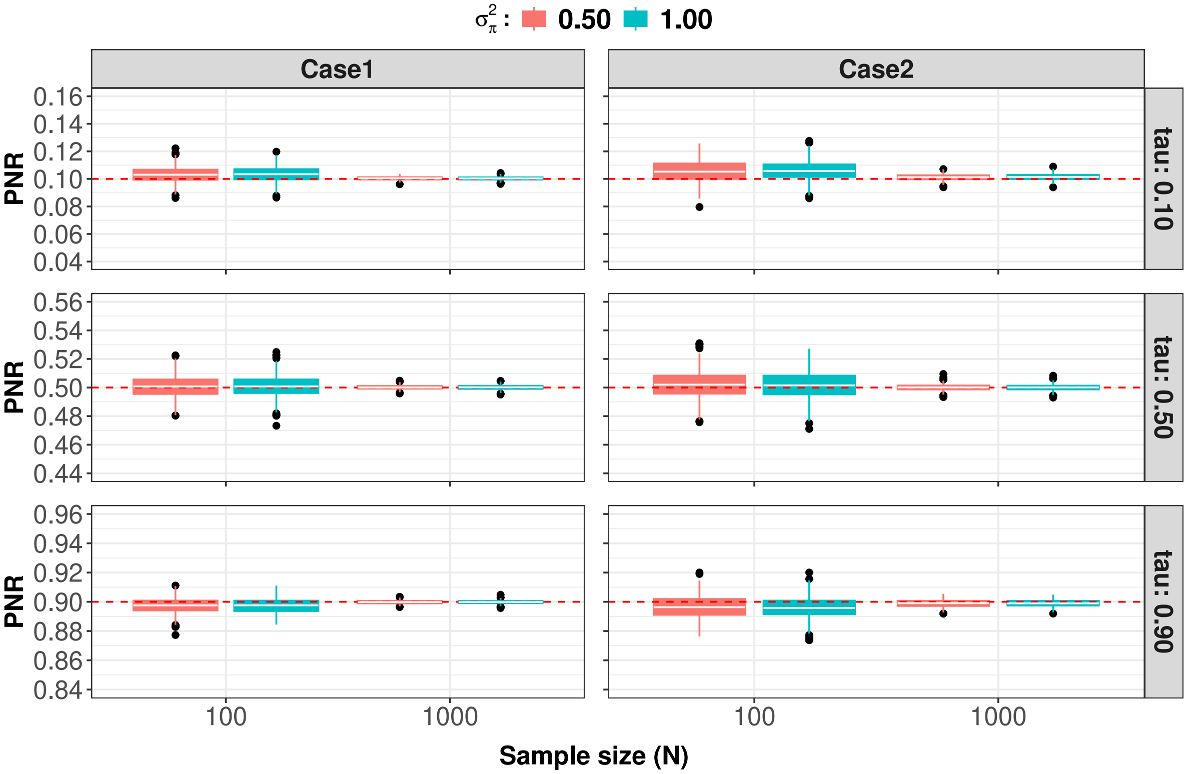


Figure 4.21: The PNR of the AQMM approach with P-splines in Study 4.5 when the simulated data had a medium variance of ω_i ($\sigma^2_\omega = 0.50$). The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

In the scenario where the variance of ω_i (σ^2_ω) was fixed at 1.00 and the variance of π_i (σ^2_π) was varied to be 0.50 and 1.00, Figures 4.22 to 4.24 show trends of MAE, RS, and PNR similar to the scenario of $\sigma^2_\omega = 0.50$. However, the values of MAE and PNR were relatively higher than those of that scenario, whereas the PNR value remained the same. This indicates that AQMM is sensitive when data explicitly show high variance in autocorrelation in subjects or individuals.

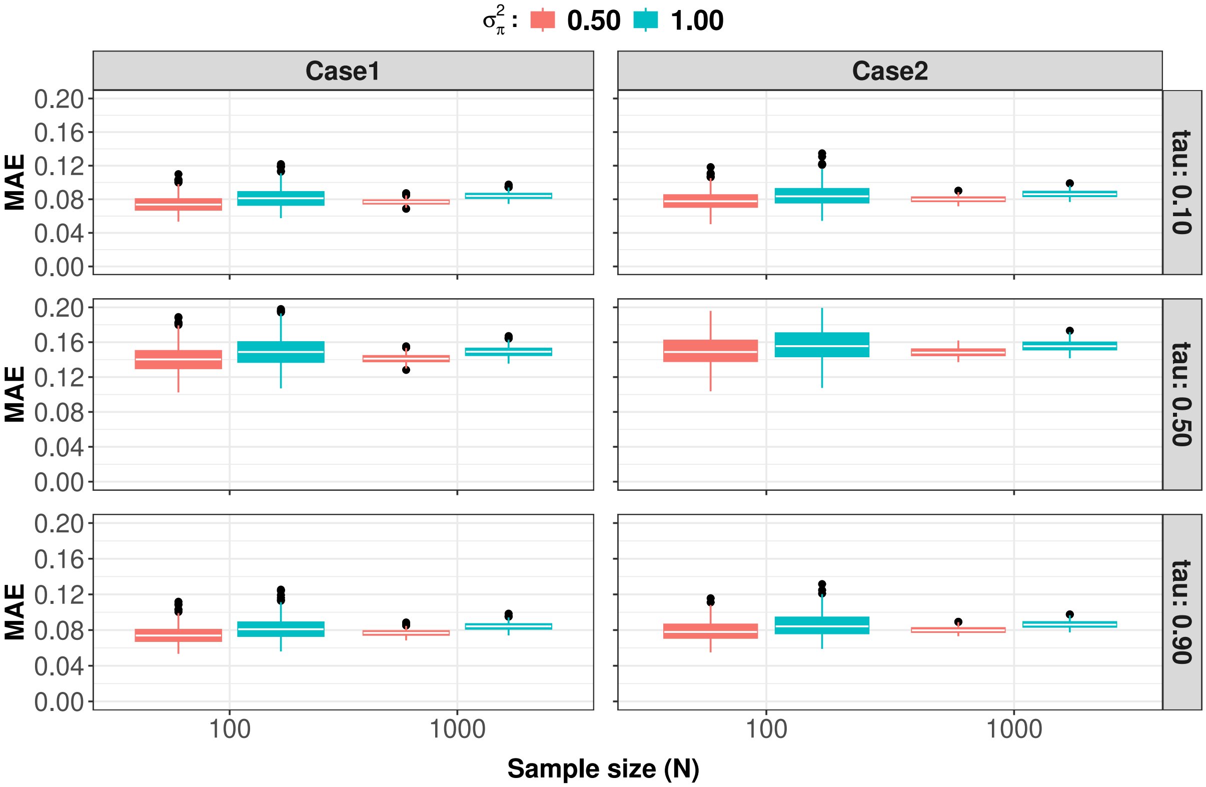


Figure 4.22: The MAE of the AQMM approach with P-splines in Study 4.5 when the simulated data had a large variance of ω_i ($\sigma^2_\omega = 1.0$). The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

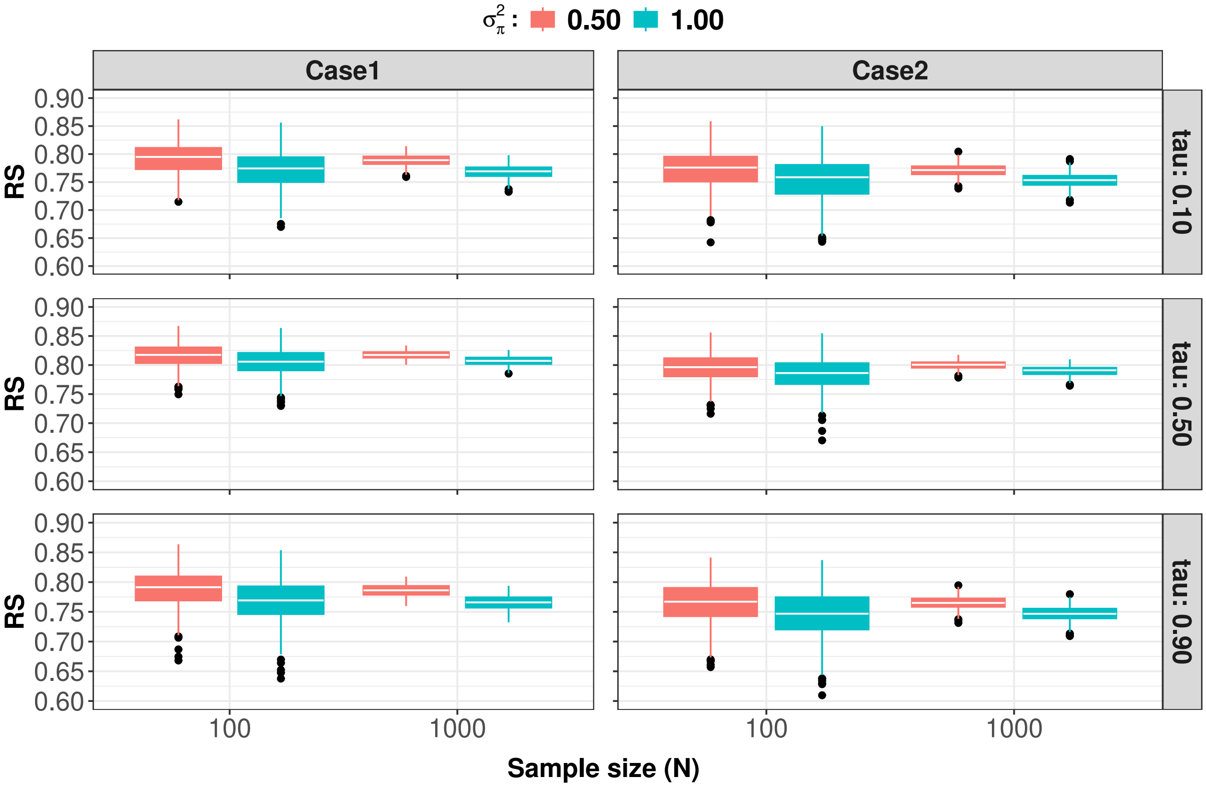


Figure 4.23: The RS of the AQMM approach with P-splines in Study 4.5 when the simulated data had a large variance of ω_i ($\sigma^2_\omega = 1.0$). The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

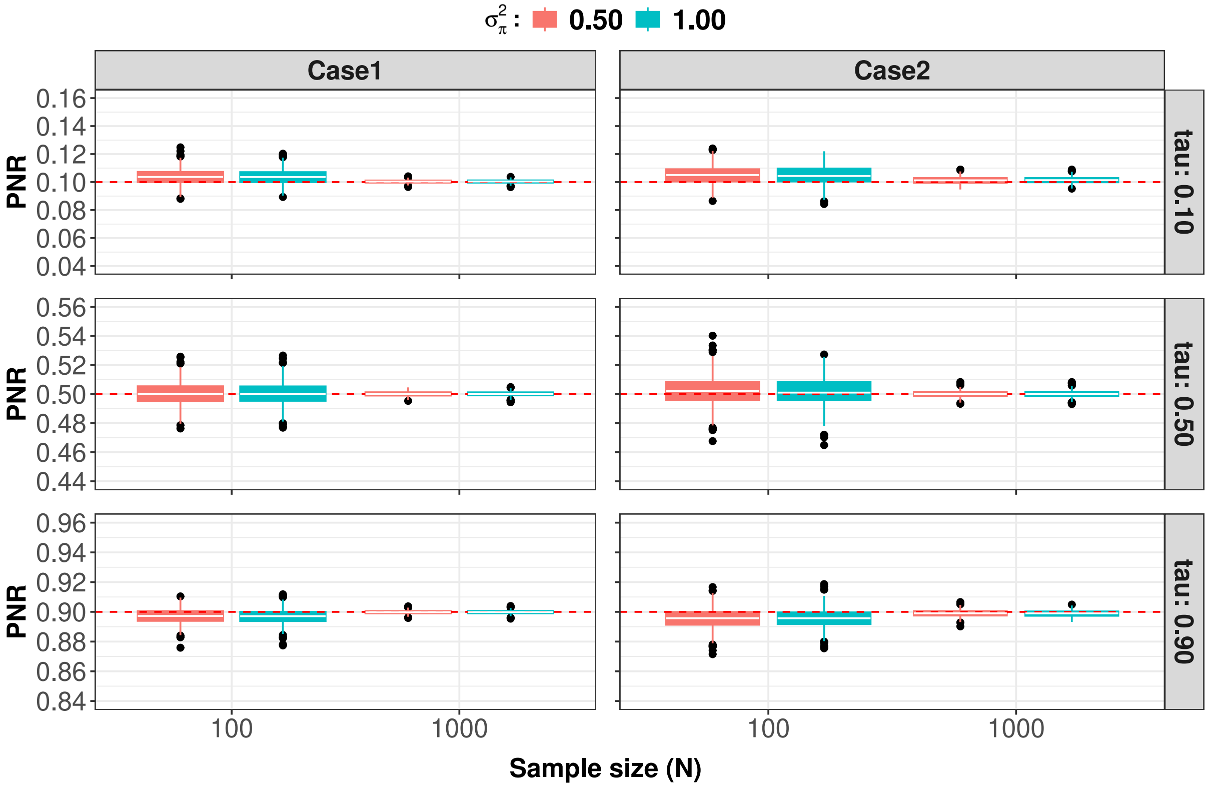


Figure 4.24: The PNR of the AQMM approach with P-splines in Study 4.5 when the simulated data had a large variance of ω_i ($\sigma_\omega^2 = 1.0$). The left column contains the results for the Case 1 scenario while the right column contains the results for the Case 2 scenario. The three rows contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively

Tables 4.21 to 4.23 confirm that both variances of the individual-specific random effect for the residual variance and the autocorrelation influenced the predictive performance of AQMM with P-splines in terms of MAE and R-squared. However, neither design affected the PNR (Table 4.23).

Table 4.21: ANOVA for three quantile models from AQMM with P-splines on MAE of Study 4.5

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	Variance of ω_i ($\sigma^2_\omega = 0.50$ and $\sigma^2_\omega = 1.00$)	1	0.17	3273.13***	0.29
	Variance of π_I ($\sigma^2_\pi = 0.50$ and $\sigma^2_\pi = 1.00$)	1	0.09	1711.16***	0.18
	Data design (Case1 and Case2)	1	0.02	325.14***	0.04
	Sample size ($N = 100$ and $N = 1000$)	1	0.01	151.75***	0.02
	$\sigma^2_\omega * \sigma^2_\pi$	1	0.00	1.52	0.00
	σ^2_ω * Data design	1	0.00	4.77*	0.00
	$\sigma^2_\omega * N$	1	0.00	0.01	0.00
	σ^2_π * Data design	1	0.00	8.35**	0.00
	$\sigma^2_\pi * N$	1	0.00	0.00	0.00
	Data design * N	1	0.00	8.74**	0.00
	Residual	7989	0.41		
0.50	Variance of ω_i (σ^2_ω)	1	0.22	1586.83***	0.17
	Variance of π_I (σ^2_π)	1	0.12	881.45***	0.10
	Data design	1	0.12	883.53***	0.10
	Sample size (N)	1	0.00	6.99**	0.00
	$\sigma^2_\omega * \sigma^2_\pi$	1	0.00	1.94	0.00
	σ^2_ω * Data design	1	0.00	8.66**	0.00
	$\sigma^2_\omega * N$	1	0.00	6.73**	0.00
	σ^2_π * Data design	1	0.00	0.54	0.00
	$\sigma^2_\pi * N$	1	0.00	2.67	0.00
	Data design * N	1	0.00	25.36***	0.00
	Residual	7989	1.09		
0.90	Variance of ω_i (σ^2_ω)	1	0.17	3175.16***	0.28
	Variance of π_I (σ^2_π)	1	0.09	1655.60***	0.17
	Data design	1	0.02	355.48***	0.04
	Sample size (N)	1	0.01	96.78***	0.01
	$\sigma^2_\omega * \sigma^2_\pi$	1	0.00	1.52	0.00
	σ^2_ω * Data design	1	0.00	5.28*	0.00
	$\sigma^2_\omega * N$	1	0.00	0.20	0.00
	σ^2_π * Data design	1	0.00	6.74**	0.00
	$\sigma^2_\pi * N$	1	0.00	0.07	0.00
	Data design * N	1	0.00	14.49***	0.00
	Residual	7989	0.43		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Table 4.22: ANOVA for three quantile models from AQMM with P-splines on R-squared of Study 4.5

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	Variance of ω_i ($\sigma^2_\omega = 0.50$ and $\sigma^2_\omega = 1.00$)	1	0.21	402.92***	0.05
	Variance of π_I ($\sigma^2_\pi = 0.50$ and $\sigma^2_\pi = 1.00$)	1	0.75	1413.65***	0.15
	Data design (Case1 and Case2)	1	0.50	942.34***	0.11
	Sample size ($N = 100$ and $N = 1000$)	1	0.01	24.88***	0.00
	σ^2_ω * σ^2_π	1	0.00	0.08	0.00
	σ^2_ω * Data design	1	0.01	9.51**	0.00
	σ^2_ω * N	1	0.00	0.80	0.00
	σ^2_π * Data design	1	0.00	4.10*	0.00
	σ^2_π * N	1	0.00	0.03	0.00
	Data design * N	1	0.00	2.23	0.00
	Residual	7989	4.22		
0.50	Variance of ω_i (σ^2_ω)	1	0.05	188.88***	0.02
	Variance of π_I (σ^2_π)	1	0.22	798.23***	0.09
	Data design	1	0.65	2342.56***	0.23
	Sample size (N)	1	0.01	49.43***	0.01
	σ^2_ω * σ^2_π	1	0.00	0.45	0.00
	σ^2_ω * Data design	1	0.00	9.81**	0.00
	σ^2_ω * N	1	0.00	5.77*	0.00
	σ^2_π * Data design	1	0.00	0.02	0.00
	σ^2_π * N	1	0.00	2.63	0.00
	Data design * N	1	0.01	21.15***	0.00
	Residual	7989	2.20		
0.90	Variance of ω_i (σ^2_ω)	1	0.21	361.02***	0.04
	Variance of π_I (σ^2_π)	1	0.79	1346.60***	0.14
	Data design	1	0.80	1355.78***	0.15
	Sample size (N)	1	0.00	2.27	0.00
	σ^2_ω * σ^2_π	1	0.00	0.08	0.00
	σ^2_ω * Data design	1	0.01	9.32**	0.00
	σ^2_ω * N	1	0.00	1.33	0.00
	σ^2_π * Data design	1	0.00	2.25	0.00
	σ^2_π * N	1	0.00	0.20	0.00
	Data design * N	1	0.01	8.51**	0.00
	Residual	7989	4.70		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Table 4.23: ANOVA for three quantile models from AQMM with P-splines on PNR of Study 4.5

τ	**Source of Variation**	**DF**	**SS**	**F**	η^2
0.10	Variance of ω_i ($\sigma_{\omega}^2 = 0.50$ and $\sigma_{\omega}^2 = 1.00$))	1	0.00	0.39	0.00
	Variance of π_I ($\sigma_{\pi}^2 = 0.50$ and $\sigma_{\pi}^2 = 1.00$)	1	0.00	1.92	0.00
	Data design (Case1 and Case2)	1	0.00	175.83***	0.02
	Sample size ($N = 100$ and $N = 1000$)	1	0.03	1175.08***	0.13
	σ_{ω}^2 * σ_{π}^2	1	0.00	1.65	0.00
	σ_{ω}^2 * Data design	1	0.00	7.96**	0.00
	σ_{ω}^2 * N	1	0.00	1.22	0.00
	σ_{π}^2 * Data design	1	0.00	0.29	0.00
	σ_{π}^2 * N	1	0.00	0.17	0.00
	Data design * N	1	0.00	20.64***	0.00
	Residual	7989	0.18		
0.50	Variance of ω_i (σ_{ω}^2)	1	0.00	1.28	0.00
	Variance of π_I (σ_{π}^2)	1	0.00	0.14	0.00
	Data design	1	0.00	17.95***	0.00
	Sample size (N)	1	0.00	67.55***	0.00
	σ_{ω}^2 * σ_{π}^2	1	0.00	0.08	0.00
	σ_{ω}^2 * Data design	1	0.00	1.43	0.00
	σ_{ω}^2 * N	1	0.00	1.38	0.00
	σ_{π}^2 * Data design	1	0.00	0.55	0.00
	σ_{π}^2 * N	1	0.00	0.06	0.00
	Data design * N	1	0.00	23.60***	0.00
	Residual	7989	0.32		
0.90	Variance of ω_i (σ_{ω}^2)	1	0.00	4.86*	0.00
	Variance of π_I (σ_{π}^2)	1	0.00	1.57	0.00
	Data design	1	0.00	123.31***	0.02
	Sample size (N)	1	0.02	815.53***	0.09
	σ_{ω}^2 * σ_{π}^2	1	0.00	0.07	0.00
	σ_{ω}^2 * Data design	1	0.00	0.03	0.00
	σ_{ω}^2 * N	1	0.00	5.54*	0.00
	σ_{π}^2 * Data design	1	0.00	0.42	0.00
	σ_{π}^2 * N	1	0.00	1.21	0.00
	Data design * N	1	0.00	7.21**	0.00
	Residual	7989	0.16		

Note: *** $p < 0.001$, ** $p < 0.005$, * $p < 0.05$

Summary

In scenarios where LCGD demonstrated between-individual differences, such as intra-individual variation and autocorrelation features, the predictive capability of the AQMM method declined. This was evidenced by increased MAE and decreased R-squared values across sixteen scenarios and three quantiles. The deterioration in predictive performance was exacerbated when the variances of the two sources of variation doubled. However, the model's PNR remained consistent at the median quantile level, with only minor deviations at the extreme quantiles, which diminished as sample sizes increased. Therefore, while the AQMM method's predictive accuracy was affected by intra-individual variation and autocorrelation features, its PNR was relatively stable.

4.5 Chapter summary

In this chapter, two flexible quantile regression approaches for analysing longitudinal data are reviewed. Subsequently, I carried out simulation studies focused on longitudinal child growth data (LCGD) to evaluate the predictive and parameter estimation performances of these approaches across various scenarios. In the initial two simulation studies, when the LCGD exhibited typical child growth characteristics including autocorrelation among repeated measurements taken from the same child, the AQMM method with two different penalised methods on cubic B-spline bases demonstrated identical performances and outperformed QSAM. AQMM allows users to model nonlinear child growth patterns using various types of splines without the need to determine smoothing parameters, and it helps in identifying risk factors associated with child growth measurements. Although AQMM requires more computational time than QSAM, especially with larger sample sizes, it is still acceptable. Conversely, QSAM has limitations, such as the requirement to specify parameters related to splines.

Furthermore, in the last three simulation studies, I explored scenarios where the LCGD encompassed additional characteristics or features. These included differences between individuals in their baseline (intercept) and growth rate (slope), intra-individual variation and, autocorrelation. Simulation results demonstrate that the AQMM approach with cubic P-splines is more effective in capturing the former characteristics. This indicates that the random effects term in the model can effectively account for these between-individual differences. However, the performance of the AQMM approach diminished when LCGD exhibited the latter two features. In this thesis, the AQMM method will be applied to fit the real LCGD from the “Growing up in Scotland” study in Chapter 6.

While the AQMM approach offers numerous advantages, it may not address all complexities present in longitudinal child growth data. As a result, certain characteristics might not be supported by this approach. Although AQMM allows us to identify risk factors associated with child growth measurement using its fixed effect component through resampling methods such as the bootstrap, it has certain limitations (Kyung et al., 2010) described in Section 5.1. Another key limitation is its inability to determine the appropriate random effects that best represent the data. Therefore, in Chapter 5, I will present a novel method that facilitates the selection of both fixed and random effects within the framework of a quantile mixed model, similar to AQMM.

Chapter 5

Variable selection for quantile mixed models

5.1 Introduction

Fitting a regression model encompasses more than simply incorporating numerous predictors or covariates in the model. It requires the ability to identify the covariates that exhibit associations with the response variable, facilitating a comprehensive understanding of real-world problems, such as evaluating risk factors in the context of child growth measurements. It is important to note that even in the case of a QR model this capacity to identify relevant covariates remains essential and cannot be circumvented. In the previous chapter, it was observed that AQMM is deficient in this regard as it lacks appropriate methods for identifying fixed effects and random effects. This limitation arises from AQMM's reliance on regularisation or penalised methods, specifically the L_2-penalised approach. The distribution of estimates obtained through these methods remains unknown (Revan Özkale & Altuner, 2022), making the estimation of standard errors a challenging task. To address this challenge, the bootstrap method has emerged as a popular approach (Chatterjee & Lahiri, 2011; Vinod, 1995). However, it is important to note that estimates derived from the bootstrap may exhibit inconsistency under certain conditions (Kyung et al., 2010). For example, with the LASSO method, when the true parameter values are close to or exactly zero, bootstrap appears to introduce a bias in a sampling distribution, potentially leading to an unstable standard error for the LASSO estimator (Fu & Knight, 2000; Leeb & Pötscher, 2005). Consequently, the bootstrap method may produce inappropriate results when it comes to selecting fixed effects or random effects.

In recent years, Bayesian LASSO (BLASSO) has been proposed as an alternative method for data analysis, offering a means to address the constraints of conventional frequentist approaches (Kyung et al., 2010; Park & Casella, 2008). This approach offers several notable advantages, including the ability to produce valid standard errors and credible intervals for estimates. Additionally, BLASSO incorporates automated predictor selection and deselection mechanisms within the model, thereby enhancing both its efficiency and interpretability. Furthermore, extended versions of BLASSO, such as Bayesian group LASSO (BGLASSO) (Kyung et al., 2010) and Bayesian sparse group LASSO (BSGLASSO) (Xu & Ghosh, 2015), enable us to effectively handle group structures present in the predictor variables. In the context of the quantile regression model, several relevant types of Bayesian LASSO approaches have been proposed (Alhamzawi & Ali, 2018; Alhamzawi & Yu, 2014; Alhamzawi et al., 2012; Ji & Shi, 2022; Li et al., 2010). Notably, Li et al. (2010) made a significant contribution by introducing the Bayesian LASSO and group LASSO to the field of quantile regression. Additionally, Alhamzawi et al. (2012) leveraged the advantages of the adaptive LASSO (Zou, 2006) and extended it to the Bayesian adaptive LASSO for QR. Furthermore, it is worth noting that Bayesian LASSO types are not limited to the general quantile model. These approaches have been successfully applied to various other regression models and frameworks, including mixed models (Alhamzawi & Yu, 2014; Ji & Shi, 2022).

However, the aforementioned Bayesian LASSO approaches have a significant limitation. Specifically, the estimates obtained from these methods may not converge precisely to zero (Alhamzawi & Ali, 2018; Park & Casella, 2008; Xu & Ghosh, 2015), which can potentially impact the accuracy of variable selection. This issue arises due to the absence of a designed point mass at zero in these approaches. As a result, the estimates may exhibit a certain degree of shrinkage but not exactly reach the zero value. To address this limitation, a spike and slab prior for regression coefficients (fixed effects) has been proposed (Ishwaran & Rao, 2005; Mitchell & Beauchamp, 1988). This prior guarantees that the estimates can be exactly zero. Recently, Xu and Ghosh (2015) applied this prior to a BGLASSO and BSGLASSO in order to select group variables and variables within a group.

Building upon the work of Xu and Ghosh (2015), I aim to extend the existing methodologies by developing BGLASSO and BSGLASSO with spike and slab priors for QR. However, their approaches solely focus on the selection of fixed effects within in the model, ignoring the consideration of random effects. To address this limitation and enable the simultaneous selection of both fixed effects and random effects, a notable approach has been proposed in the Bayesian framework introduced by Kinney and Dunson (2007). This approach employs a reparameterisation of the random effects component within the linear mixed model, thereby allowing for the specification of appropriate priors for the associated random effects parameters. To the best of our knowledge, our work represents the first work to explore BGLASSO and BSGLASSO with spike and slab priors for QR, encompassing both fixed and random effects simultaneously.

This chapter focuses on the development of BGLASSO and BSGLASSO methodologies with spike and slab priors for incorporating both fixed and random effects within the QR model. Section 5.2 provides a detailed description of the group structures present in the predictors. Furthermore, Section 5.3 explores the group and sparse group LASSO, while Section 5.4 delves into the Bayesian group and sparse group LASSO. The utilisation of the spike and slab priors for regression coefficients is discussed in Section 5.5. The following sections, Sections 5.6 and 5.7, introduce the BGLASSO and BSGLASSO approaches for selecting fixed effects within the QR framework, respectively. Section 5.8 explains the process of variable selection for both novel approaches. In addition, Section 5.9 explains the simultaneous selection methods for both fixed and random effects in the quantile mixed models (QMMs). Both Sections 5.10 and 5.11 demonstrate simulation studies through fixed effect selection and simultaneous selection, respectively. Section 5.12 provides a sensitivity analysis for prior specifications. Furthermore, Section 5.13 shows an illustrative analysis from a simulated dataset. Finally, Section 5.14 provides a comprehensive summary of the chapter's key findings and contributions.

5.2 Group structures of predictors

In practical applications, such as child growth development studies, risk factors that influence child growth can take various forms, including continuous variables, categorical variables, and functional variables. As a result, when modelling the associations between these risk factors and growth development outcomes, a diverse array of predictor structures arises. To accommodate categorical variables, it is common practice to construct $d-1$ dummy variables, where d represents the number of levels in the categorical variable. On the other hand, functional variables are commonly represented using spline methods to create basis functions. Consequently, both types of variables can be considered as groups of variables within the QR framework.

Consider the QR model consisting of the covariates X_k, which are partitioned into G non-overlapping groups,

$$\mathbf{y} = \sum_{l=1}^{G} \mathbf{X}_l \boldsymbol{\beta}_{\tau,l} + \boldsymbol{\epsilon}. \tag{5.1}$$

Here, the matrix $\mathbf{X}_l$ has d_l columns and n rows, and it is a sub-matrix of the design matrix $\mathbf{X}$ and $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{ld_l})'$ is the coefficient vector corresponding to the group l. Note that for simplicity of notation, I will omit the subscript τ for $\boldsymbol{\beta}_{\tau,l}$.

For example, consider three variable groups ($G = 3$): the first group consists of a single continuous covariate ($d_1 = 1$); the second consists of three continuous covariates, such as basis functions ($d_2 = 3$); and the third group consiss of two levels of a categorical covariate ($d_3 = 2$). Then the matrix $\mathbf{X}_l$ for three groups can be written as

$$\mathbf{X}_1 = \begin{bmatrix} X_{1,1,1} \\ X_{1,1,1} \\ \vdots \\ X_{1,1,n} \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} X_{2,1,1} & X_{2,2,1} & X_{2,3,1} \\ X_{2,1,2} & X_{2,2,2} & X_{2,3,2} \\ \vdots & \vdots & \vdots \\ X_{2,1,n} & X_{2,2,n} & X_{2,3,n} \end{bmatrix}, \quad \mathbf{X}_3 = \begin{bmatrix} X_{3,1,1} & X_{3,2,1} \\ X_{3,1,2} & X_{3,2,2} \\ \vdots & \vdots \\ X_{3,1,n} & X_{3,2,n} \end{bmatrix}.$$

Hence, the design matrix $\mathbf{X}$ of the model (5.1) is

$$\mathbf{X} = \left[\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\right],$$

Consequently, the coefficient vector associated with the design matrix $\mathbf{X}$ can be defined as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3) = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$, where $\boldsymbol{\beta}_1 = (\beta_1)$, $\boldsymbol{\beta}_2 = (\beta_2, \beta_3, \beta_4)$ and $\boldsymbol{\beta}_3 = (\beta_5, \beta_6)$.

5.3 Group and sparse group LASSO

One regularisation technique commonly used for variable selection in regression models is known as the "Least Absolute Shrinkage and Selection Operator" (LASSO) (Tibshirani, 1996). This technique incorporates an L_1 penalty on the regression coefficients within the loss function, typically least squares, which is defined as

$$\min_{\boldsymbol{\beta}} \left|\left|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda||\boldsymbol{\beta}||_1, \tag{5.2}$$

where $||\mathbf{u}||_2^2 = \sum_{i=1}^n u_i^2$ for a vector $u \in \mathcal{R}^n$ and $||\cdot||_1$ denotes the L_1 norm. The introduction of this penalty results in certain coefficients being effectively shrunk to zero, enabling automatic variable selection. However, it should be noted that LASSO is not well-suited for predictors exhibiting a group structure, as seen in the model (5.1). To address the issue of group variable selection, an alternative variation of LASSO known as the "Group LASSO" (GL) was proposed (Yuan & Lin, 2006). By employing the GL method, the challenges associated with group variable selection can be effectively mitigated.

For the **mean** model, the estimation of the group LASSO estimator can be achieved following the approach proposed by Yuan and Lin (2006):

$$\min_{\boldsymbol{\beta}} \left|\left| \mathbf{y} - \sum_{l=1}^{G} \mathbf{X}_l \boldsymbol{\beta}_l \right|\right|_2^2 + \lambda \sum_{l=1}^{G} ||\boldsymbol{\beta}_l||_2. \tag{5.3}$$

Similarly, the group LASSO approach can be applied to the **quantile** model as proposed by Li et al. (2010). Thus, at the τth quantile, the estimator can be obtained by

$$\min_{\boldsymbol{\beta}} \rho_\tau \left(\mathbf{y} - \sum_{l=1}^{G} \mathbf{X}_l \boldsymbol{\beta}_l \right) + \lambda \sum_{l=1}^{G} ||\boldsymbol{\beta}_l||_2. \tag{5.4}$$

Here, λ represents the regularisation parameter, controlling the amount of shrinkage applied to the coefficients of the regression model. It plays a crucial role in controlling the trade-off between bias and variance in the model.

Note that both (5.3) and (5.4) utilise different penalty terms compared to the standard LASSO (5.2). Their penalty terms are based on the L_2 norm (Euclidean norm) of the coefficients within the l-th group (referred to as the Group LASSO penalty), whereas the standard LASSO employs the L_1 norm (absolute value norm) of the coefficients (Yuan & Lin, 2006). The term $||\boldsymbol{\beta}||_1$ in (5.2) promotes sparsity in the coefficient vector $\boldsymbol{\beta}$, leading to the selection of a subset of relevant features. Meanwhile, the term $\sum_{l=1}^{G} ||\boldsymbol{\beta}_l||_2$ in both (5.3) and (5.4) encourage sparsity at the group level, effectively selecting entire groups of features rather than individual ones. This penalty term combines characteristics of both L_1 and L_2 penalties: the L_1 penalty promotes sparsity, while the L_2 penalty shrinks coefficients towards zero but usually does not set any coefficient exactly to zero. Thus, it is sometimes referred to as *Intermediate penalty* (Meier et al., 2008; Yuan & Lin, 2006) or *Generalised penalty* (Hastie et al., 2015). The term "Group LASSO" is used because the method retains the fundamental principles and objectives of the standard LASSO algorithm, despite using the L_2 norm instead of the L_1 norm for penalty terms. For instance, it promotes sparsity similar to the standard LASSO but at the group level by encouraging entire groups of coefficients to be zero, effectively selecting or excluding whole groups of features.

However, the penalised objective functions (5.3) and (5.4) are limited in their scope, as they solely cater to group-structured variables without considering individual levels within those group variables, thereby lacking the ability to address broader real-world issues. For instance, child growth measurements can be influenced by categorical risk factors that have one or more individual levels that are sparse within the group. To overcome this limitation, the "sparse Group LASSO" (SGL) methodology was introduced (Simon et al., 2013). SGL allows for the selection of variables at both the group and within-group levels. This is achieved by incorporating L_1 and L_2 penalties into the loss function of the linear regression model, SGL facilitates the joint selection of variables at both levels. Consequently, both (5.3) and (5.4) functions can be expanded to take the following form

$$\min_{\boldsymbol{\beta}} \left|\left|\mathbf{y} - \sum_{l=1}^{G} \mathbf{X}_l \boldsymbol{\beta}_l\right|\right|_2^2 + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 \sum_{l=1}^{G} ||\boldsymbol{\beta}_l||_2 \tag{5.5}$$

and

$$\min_{\boldsymbol{\beta}} \rho_\tau \left(\mathbf{y} - \sum_{l=1}^{G} \mathbf{X}_l \boldsymbol{\beta}_l\right) + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 \sum_{l=1}^{G} ||\boldsymbol{\beta}_l||_2, \tag{5.6}$$

respectively, where λ_1 and λ_2 are the regularisation parameters that control the amount of the L_1 and L_2 penalties applied. These parameters control the balance between the within-group levels selection and group selection. Note that both estimations (5.5) and (5.5) incorporate the L_1 and L_2 penalties similar to the elastic net (Zou & Hastie, 2005). However, the elastic net applies regularisation equally to all predictor variables and does not explicitly handle the grouping structure among them. In contrast, the SGL penalises coefficients at both the group level (using the L_2 norm) and within each group (using the L_1 norm).

5.4 Bayesian group and sparse group LASSO

As noted by Tibshirani (1996), it has been established that posterior mode estimates, obtained by specifying an independent and identical Laplace (i.e., double-exponential) prior for each regression coefficient, exhibit equivalence to the estimates obtained through the LASSO method. Consequently, this property can be leveraged to apply to the LASSO types discussed in Section 5.3.

Taking inspiration from the aforementioned property, Kyung et al. (2010) introduced the GL approach within the Bayesian framework. This was accomplished by specifying a multivariate generalisation of the double exponential prior distribution for regression coefficients,

$$p(\boldsymbol{\beta}_l) \propto \exp\left\{ -\frac{\lambda}{\sigma} ||\boldsymbol{\beta}_l||_2 \right\}. \tag{5.7}$$

Furthermore, this concept can be applied similarly to the SGL framework, specifically functions (5.5) and (5.6), as discussed by Xu and Ghosh (2015), where:

$$p(\boldsymbol{\beta}) \propto \exp\left\{ -\frac{\lambda_1}{2\sigma^2}||\boldsymbol{\beta}||_1 - \frac{\lambda_2}{2\sigma^2}||\boldsymbol{\beta}_l||_2 \right\}. \tag{5.8}$$

In practice, the priors (5.7) and (5.8) can be represented as a scale mixture of normals with a conjugate Gamma hyperpriors, allowing for the construction of a fully Bayesian hierarchical model and an efficient Gibbs sampler to address the group LASSO and sparse group LASSO problem (Xu & Ghosh, 2015). A notable advantage of Bayesian analysis lies in its ability to generate standard errors effortlessly through the employment of the Markov Chain Monte Carlo (MCMC) algorithm. In contrast, the LASSO approach faces limitations when it comes to estimating standard errors (Chatterjee & Lahiri, 2011; Fu & Knight, 2000; Xu & Ghosh, 2015).

In the domain of quantile regression, Li et al. (2010) made a seminal contribution by introducing the Bayesian group LASSO methodology. Their work was particularly focused on the integration of the Bayesian group LASSO methodology into quantile regression models, with a specific emphasis on the assumption of errors following the asymmetric Laplace distribution. It is worth noting that, to the best of our knowledge, there has been no previous exploration of the Bayesian sparse group LASSO method in the context of quantile regression.

Nevertheless, Xu and Ghosh (2015) highlighted that estimates derived from this prior, whether they are posterior means or medians, do not precisely reach a value of zero. In other words, the prior does not assign any point-mass specifically on zero. To address this issue, they proposed a hierarchical Bayesian sparse group LASSO model, incorporating an independent spike and slab type prior for both group variable selection and individual variable selection.

5.5 Spike and slab priors for regression coefficients

Spike and slab priors have been emerging as a prominent and widely favored approach for Bayesian variable selection, ever since their introduction by Mitchell and Beauchamp (1988). These priors are typically formulated as mixture priors, often utilising normal mixture priors with two components (Ishwaran & Rao, 2005). One component assigns a probability mass to the regression coefficient β_k being precisely equal to zero (spike component), while the other component allows for non-zero (slab component). This unique feature enables spike and slab priors to effectively identify and select relevant variables, making them suitable for variable selection tasks. As a result, these priors offer a flexible and powerful tool for variable selection within a Bayesian framework.

Considering the linear regression model, it can be expressed as $y_i \sim \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$. In the context of Bayesian variable selection, a widely used version of the spike and slab priors for β_k (Ishwaran & Rao, 2005; L. Zhang et al., 2014) is

$$\begin{aligned}
\beta_k|\gamma_k, \nu_k^2 &\overset{\text{ind}}{\sim} (1-\gamma_k)\delta_0(\beta_k) + \gamma_k N(0, \nu_k^2), \quad k = 1, \dots, p, \\
\gamma_k|\pi &\sim \text{Bernoulli}(\pi), \\
\nu_k^2 &\overset{\text{i.i.d.}}{\sim} \mathcal{G}(\cdot),
\end{aligned} \tag{5.9}$$

where $\gamma_k \in [0, 1]$ corresponds to a latent binary indicator variable functioning as a mixture weight, δ_0 denotes the Dirac delta function which assigns all its mass specifically at the value zero, ν_k^2 is the variance of the slab distribution (or normal scale parameter) and $\mathcal{G}(\cdot)$ is a scale-mixture of normals (e.g. an inverse Gamma (InvGamma) distribution with shape and scale parameters). The π represents the probability that $\gamma_k = 1$, with a common practice of assigning π to $1/2$. The rationale behind this prior specification is that when $\gamma_k = 1$, β_k is assumed to have a normal density with a large value ν_k^2. In this case, β_k is selected as a relevant variable and included in the model. Conversely, when $\gamma_k = 0$, β_k is assumed to follow a point mass density at zero, indicating that β_k is excluded from the model.

Given the computational challenges arising from the presence of numerous parameters in the prior for β_k, of (5.9), a class of continuous bimodal priors has been proposed to address this issue (Ishwaran & Rao, 2005). These priors offer a more computationally efficient option while maintaining the desired characteristics of the model. The specification of this class of priors is

$$\begin{aligned}
\beta_k|\gamma_k, \nu_k^2, &\overset{\text{ind}}{\sim} N(0, \gamma_k\nu_k^2), \quad k = 1, \dots, p, \\
\gamma_k|\pi &\overset{\text{i.i.d.}}{\sim} (1-\pi)\delta_0(\gamma_k) + \pi\delta_1(\gamma_k), \\
\nu_k^{-2} &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(a, b), \\
\pi &\sim \text{Beta}(b_1, b_2).
\end{aligned} \tag{5.10}$$

Indeed, the prior of β_k in (5.10) can be written more compactly as

$$\beta_k|\nu_k^2, \pi \overset{\text{ind}}{\sim} (1-\pi)\delta_0(\beta_k) + \pi N(0, \nu_k^2).$$

In contrast to a manually set bimodal prior, as described in (5.9), this specification results in a continuous bimodal distribution for the variance of β_k $(\gamma_k\nu_k^2)$, featuring a spike at zero and a right continuous tail. These distinctive features play a crucial role in identifying zero and nonzero values for the β_k coefficients, respectively. Additionally, the parameter π serves another purpose as it controls the likelihood of ν_k taking a value of one or zero. Ishwaran and Rao (2005) stated that it acts as a parameter used to control the size of models, known as a complexity parameter.

5.6 Bayesian group LASSO QR with spike and slab prior for fixed effect selection

Motivated by a work of Xu and Ghosh (2015), the representation of the AL distribution as a scale mixture of normals with an exponential mixing density in Section 3.4, as well as the priors specifications detailed in Section 5.5, I develop the Bayesian group LASSO QR with spike and slab priors. This proposed approach combines the Bayesian framework with group LASSO method and incorporates spike and slab priors to enhance the robustness and accuracy of quantile regression (hereafter referred to as BGLSSQR).

Let $\mathbf{y} = (y_1, \ldots, y_n)'$ represent the response variable, $\boldsymbol{\beta}_l$ is a coefficient vector of length d_l, $\mathbf{X}_l$ is an $n \times d_l$ covariate matrix corresponding to the factor $\boldsymbol{\beta}_l (l = 1, 2, \ldots, G)$, and $\mathbf{V} = \text{diag}(v_1^{-1}, \ldots, v_n^{-1})$. Hence, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G)'$ is a coefficients vector of length p $(p = \sum_{l=1}^{G} d_l)$ and $\mathbf{X}$ is an $n \times p$ design (predictors) matrix corresponding to $\boldsymbol{\beta}$. Here, $\boldsymbol{\Psi}_l$ is a known $d_l \times d_l$ positive definite matrix, and π_0 is the probability of $\boldsymbol{\beta}_l = 0$. Therefore, the Bayesian hierarchical QR model can be expressed as

$$\begin{aligned}
\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{v}, \sigma &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + (1-2\tau)\mathbf{V}^{-1}\mathbf{1}_n, 2\sigma\mathbf{V}), \\
\boldsymbol{v}|\sigma &\sim \text{Exp}(\sigma^{-1}\tau(1-\tau)), \\
\boldsymbol{\beta}_l|\eta_l^2 &\overset{\text{ind}}{\sim} (1-\pi_0)\mathcal{N}_{d_l}(\mathbf{0}, \eta_l^2\boldsymbol{\Psi}_l^{-1}) + \pi_0\delta_0(\boldsymbol{\beta}_l), \\
\eta_l^2|\lambda^2 &\overset{\text{ind}}{\sim} \text{Gamma}\Big(\frac{d_l+1}{2}, \frac{\lambda^2}{2}\Big), \\
\sigma &\sim \text{InvGamma}(g_1, g_2), \\
\pi_0 &\sim \text{Beta}(a_1, a_2).
\end{aligned} \tag{5.11}$$

Based on the Bayesian hierarchical model (5.11), the (joint) posterior distribution given observed data is

$$p(\boldsymbol{\beta}, \boldsymbol{\eta}^2, \boldsymbol{v}, \sigma, \pi_0|\mathbf{y}, \mathbf{X}) \propto l(\mathbf{y}|\boldsymbol{\beta}, \sigma, \boldsymbol{v})p(\boldsymbol{\beta}|\sigma, \boldsymbol{v}, \pi_0)p(\boldsymbol{\eta}^2|\lambda^2)p(\boldsymbol{v}|\sigma)p(\sigma)p(\pi_0).$$

The detailed formulation of this posterior distribution is provided in Appendix A.1.

The hyperparameter λ holds considerable importance within any Bayesian LASSO model type, as it plays a pivotal role in the coefficient shrinkage process, thus exerting a substantial influence on its overall effectiveness and predictive capabilities (Park & Casella, 2008; Xu & Ghosh, 2015). One commonly employed approach is to assign a conjugate prior, such as the Gamma(c_1, c_2) prior, to this hyperparameter, which allows for straightforward computation and incorporation of prior beliefs or knowledge regarding coefficient sparsity. However, an alternative method for estimating λ exists. Park and Casella (2008) as well as Xu and Ghosh (2015) employed an empirical Bayes approach utilising marginal maximum likelihood estimation. This data-driven method leverages a Monte Carlo EM algorithm to derive estimates of the marginal likelihood, thereby yielding an estimate for λ based on the observed data. The kth update of the EM algorithm for λ is represented by

$$\lambda^{(k)} = \sqrt{\frac{p+G}{\sum_{l=1}^{G} E_{\lambda^{k-1}}[\eta_l^2|\mathbf{y}]}},$$

where p denotes the number of covariates and G represents the number of group variables. During this EM update, the value of η_l^2 is substituted by the sample average of η_l^2, which is obtained through the Gibbs sampler using the λ from the previous iteration $(k-1)$.

Additionally, hyperparameters such as the shape and scale parameters of an inverse Gamma prior for the scale parameter, σ, can be chosen to be relatively small: $g_1 = 0.1$ and $g_2 = 0.1$. This decision is based on allowing the data to speak for themselves in determining the posterior distribution and ensuring computational stability (Gelman, 2006, 2014). Also, both hyperparameters of a Beta prior for π_0 are set to $a_1 = a_2 = 1$ as default values. This is equivalent to assuming a standard uniform distribution, resulting in a prior mean of 1/2 and allowing a prior spread of π_0 (Xu & Ghosh, 2015).

Gibbs sampler

Let d_l denote the length of the coefficient vector $\boldsymbol{\beta}_l$, and $\xi = (1-2\tau)$. The full conditional distribution of $\boldsymbol{\beta}_l$ follows a multivariate spike and slab distribution, given by

$$\boldsymbol{\beta}_l|\text{rest} \sim (1-q_l)\mathcal{N}_{d_l}(\boldsymbol{\mu}_l, 2\sigma\boldsymbol{\Sigma}_l) + q_l\delta_0(\boldsymbol{\beta}_l), \quad l = 1, \ldots, G,$$

where

$$\boldsymbol{\mu}_l = \boldsymbol{\Sigma}_l\mathbf{X}_l'\mathbf{V}(\mathbf{y} - \mathbf{X}_{(l)}\boldsymbol{\beta}_{(l)} - \xi\boldsymbol{v}),$$

and

$$\boldsymbol{\Sigma}_l = \left(\mathbf{X}_l'\mathbf{V}\mathbf{X}_l + \frac{1}{\eta_l^2}\mathbf{I}_{d_l}\right)^{-1}.$$

Here, q_l is the probability which controls the spike and slab of the distribution, computed as:

$$\begin{aligned} q_l &= P(\boldsymbol{\beta}_l = \mathbf{0}|\text{rest}) \\ &= \frac{\pi_0}{\pi_0 + (1-\pi_0)(\eta_l^2)^{-\frac{d_l}{2}}|\boldsymbol{\Sigma}_l|^{\frac{1}{2}} \exp\left\{\frac{1}{4\sigma}||\boldsymbol{\Sigma}_l^{\frac{1}{2}}\mathbf{X}_l'\mathbf{V}(\mathbf{y} - \mathbf{X}_{(l)}\boldsymbol{\beta}_{(l)} - \xi\boldsymbol{v})||_2^2\right\}}. \end{aligned}$$

It is worth noting that the term $\mathbf{y} - \mathbf{X}_{(l)}\boldsymbol{\beta}_{(l)} - \xi\boldsymbol{v}$ represents the residual vector obtained by excluding the lth factor $\boldsymbol{\beta}_l$ in our quantile regression model. As a result, the elements in $\mathbf{X}_l'(\mathbf{y} - \mathbf{X}_{(l)}\boldsymbol{\beta}_{(l)} - \xi\boldsymbol{v})$ are proportional to the correlation between each predictor in the lth group and this residual vector.

Rather than directly defining the full conditional of η_l^2, an alternative approach is adopted by introducing the parameter $\alpha_l^2 = 1/\eta_l^2$. The full conditional of α_l^2 is subsequently defined under two distinct conditions. In the case when $\boldsymbol{\beta}_l = 0$, the full conditional distribution of α_l^2 is

$$\alpha_l^2|\text{rest} \sim \text{InvGamma}\left(s_1 = \frac{d_l + 1}{2}, s_2 = \frac{\lambda^2}{2}\right),$$

where s_1 and s_2 are the shape and scale parameters of an inverse Gamma distribution, respectively. On the other hand, when $\boldsymbol{\beta}_l \neq 0$,

$$\alpha_l^2|\text{rest} \sim \text{InvGaussian}\left(\mu' = \frac{\sqrt{\lambda^2\sigma}}{||\beta_l||_2}, \lambda' = \lambda^2\right).$$

The full conditional distribution of each v_i is then an inverse Gaussian distribution,

$$v_i|\text{rest} \sim \text{InvGaussian}\left(\mu' = \frac{1}{|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}|}, \lambda' = \frac{1}{2\sigma}\right).$$

The full conditional distribution of σ is an inverse Gamma distribution,

$$\sigma|\text{rest} \sim \text{InvGamma}\left(\frac{3n}{2} + \frac{k}{2} + g_1, \frac{(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})}{2} + \tau(1-\tau)\sum_{i=1}^n v_i + \boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta} + g_2\right),$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \xi\boldsymbol{v}, k = \sum_{l=1}^G d_l \mathbf{1}_{\{\boldsymbol{\beta}_l \neq 0\}}$ and $\mathbf{S} = \text{diag}(\eta_1^{-2}, \ldots, \eta_G^{-2})$.

The full conditional distribution of π_0 is a Beta distribution,

$$\pi_0|\text{rest} \sim \text{Beta}\left(a_1 + \sum_{l=1}^G \mathbf{1}_{\{\boldsymbol{\beta}_l \neq 0\}}, a_2 + \sum_{l=1}^G d_l - \sum_{l=1}^G \mathbf{1}_{\{\boldsymbol{\beta}_l \neq 0\}}\right).$$

5.7 Bayesian sparse group LASSO QR with spike and slab prior for fixed effect selection

Following the reparameterisation of regression coefficients given by Xu and Ghosh (2015), I adopt their approach to QR to handle sparsity at the group level and within the group level:

$$\boldsymbol{\beta}_{\tau,l} = \mathbf{K}_{\tau,l}^{1/2} \boldsymbol{b}_{\tau,l},$$

where $\mathbf{K}_{\tau,l}^{1/2} = \operatorname{diag}\{\theta_{l1}, \dots, \theta_{ld_l}\}, \theta_{lj} \geq 0, l = 1, \dots, G; j = 1, \dots, d_l$. Here, $\boldsymbol{b}_{\tau,l} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_l})$ in case where these are not zeros. This reparameterisation, employed in this context, plays a crucial role in utilising the diagonal elements of $\mathbf{K}_{\tau,l}^{1/2}$ to regulate the magnitude of the elements of $\boldsymbol{\beta}_{\tau,l}$. To maintain the simplicity of notation, the subscript τ for all pertinent parameters will be omitted.

A methodology for the selection of group variables is based on the assumption that each $\boldsymbol{b}_l$ follows the multivariate spike and slab prior:

$$\boldsymbol{b}_l \overset{\text{ind}}{\sim} (1 - \pi_0)\mathcal{N}_{d_l}(\mathbf{0}, \mathbf{I}_{d_l}) + \pi_0 \delta_0(\boldsymbol{b}_l).$$

Consequently, if θ_{lj} is zero, regardless of whether b_{lj} is zero or not, β_{lj} can be dropped from the model.

In a similar way to the selection of group variables, the spike and slab prior is utilised for the within-group level. To ensure $\theta_{lj} \geq 0$, each θ_{lj} is assumed to adhere to this prior distribution:

$$\theta_{lj} \overset{\text{ind}}{\sim} (1 - \pi_1) N^+(0, s^2) + \pi_1 \delta_0(\theta_{lj}).$$

Here, $N^+(0, s^2)$ denotes a truncated normal distribution with mean 0 and variance s^2, bounding the random variable from below at 0.

I combine the Bayesian sparse group LASSO method with the mixture representation of the asymmetric Laplace (AL) distribution and the aforementioned spike and slab priors to develop the Bayesian sparse group LASSO for quantile regression (BSGSSQR). This innovative approach is structured as a Bayesian hierarchical model, integrating the mentioned components to achieve efficient variable selection within the context of quantile regression. Let $\boldsymbol{b} = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_G)'$, and $\mathbf{K}^{1/2} = \operatorname{diag}(\mathbf{K}_1^{1/2}, \dots, \mathbf{K}_G^{1/2})$, then the Bayesian hierarchical QR model is expressed as follows:

$$\begin{aligned}
\mathbf{y}|\boldsymbol{b},\boldsymbol{v},\sigma &\sim \mathcal{N}(\mathbf{X}\mathbf{K}^{1/2}\boldsymbol{b}+(1-2\tau)\mathbf{V}^{-1}\mathbf{1}_n, 2\sigma\mathbf{V}),\\
\boldsymbol{v}|\sigma &\sim \mathrm{Exp}(\sigma^{-1}\tau(1-\tau)),\\
\boldsymbol{b}_l &\overset{\text{ind}}{\sim} (1-\pi_0)\mathcal{N}_{d_l}(\mathbf{0},\mathbf{I}_{d_l})+\pi_0\delta_0(\boldsymbol{b}_l),\\
\theta_{lj} &\overset{\text{ind}}{\sim} (1-\pi_1)N^{+}(0,s^2)+\pi_1\delta_0(\theta_{lj}),\\
s^2|t &\overset{\text{ind}}{\sim} \mathrm{InvGamma}(1,t),\\
\sigma &\sim \mathrm{InvGamma}(g_1,g_2), \quad g_1=0.1, g_2=0.1\\
\pi_0 &\sim \mathrm{Beta}(a_1,a_2), \quad a_1=1, a_2=1\\
\pi_1 &\sim \mathrm{Beta}(c_1,c_2), \quad c_1=1, c_2=1.
\end{aligned} \tag{5.12}$$

Based on the Bayesian hierarchical QR model (5.12) described above, the (joint) posterior distribution given observed data is

$$p(\boldsymbol{b},\boldsymbol{\theta}^2,\boldsymbol{v},\sigma,\pi_0,\pi_1,s^2|\mathbf{y},\mathbf{X}) \propto l(\mathbf{y}|\boldsymbol{b},\sigma,\boldsymbol{v})p(\boldsymbol{b}_l|\pi_0)p(\boldsymbol{\theta}^2|\pi_1)p(\boldsymbol{v}|\sigma)p(\sigma)p(\pi_0)p(\pi_1)p(s^2).$$

The detailed formulation of this posterior distribution is provided in Appendix A.2.

Analogous to the estimation of λ as discussed in Section 5.6, the hyperparameter t assumes a crucial role in the process of coefficient shrinkage. To estimate this parameter, the Monte Carlo EM algorithm is adopted, as described by Casella (2001) and Park and Casella (2008). For the kth EM update,

$$t^{(k)} = \frac{1}{E_{t^{(k-1)}}[\frac{1}{s^2}|\mathbf{y}]},$$

where the posterior expectation of s^2 is substituted by the sample average of s^2. This average is obtained through the Gibbs sampler using the t from the previous iteration $(k-1)$.

Gibbs sampler

The full conditional distribution of $\boldsymbol{b}_l$ is then a multivariate spike and slab distribution

$$\boldsymbol{b}_l|\text{rest} \sim (1-q_l)\mathcal{N}_{d_l}(\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)+q_l\delta_0(\boldsymbol{b}_l),$$

where

$$\boldsymbol{\mu}_l = \left(\frac{1}{2\sigma}\boldsymbol{\Sigma}_l^{\frac{1}{2}}\mathbf{K}_l^{\frac{1}{2}}\mathbf{X}_l'\mathbf{V}(\mathbf{y}-\mathbf{X}_{(l)}\mathbf{K}_{(l)}^{\frac{1}{2}}\boldsymbol{b}_{(l)}^{\frac{1}{2}}-\xi\boldsymbol{v})\right),$$

$$\boldsymbol{\Sigma}_l = \left(\frac{1}{2\sigma} \mathbf{K}_l^{\frac{1}{2}} \mathbf{X}_l' \mathbf{V} \mathbf{X}_l \mathbf{K}_l^{\frac{1}{2}} + \mathbf{I}_{d_l} \right)^{-1},$$

and q_l be the posterior probability of $\boldsymbol{b}_l$ being zero given the remaining parameters and can be computed by

$$\begin{aligned} q_l &= P(\boldsymbol{b}_l = \mathbf{0}|\text{rest}) \\ &= \frac{\pi_0}{\pi_0 + (1-\pi_0)|\boldsymbol{\Sigma}_l|^{\frac{1}{2}} \exp\left\{ \frac{1}{4\sigma^2} ||\boldsymbol{\Sigma}_l^{\frac{1}{2}} \mathbf{K}_l^{\frac{1}{2}} \mathbf{X}_l' \mathbf{V}(\mathbf{y} - \mathbf{X}_{(l)} \mathbf{K}_{(l)}^{\frac{1}{2}} \boldsymbol{b}_{(l)}^{\frac{1}{2}} - \xi \boldsymbol{v})||_2^2 \right\}}. \end{aligned}$$

The full conditional distribution of θ_{lj} is then a spike and slab distribution, with the slab a positive part normal distribution

$$\theta_{lj}|\text{rest} \sim (1 - r_{lj}) N^+(u_{lj}, v_{lj}^2) + r_{lj} \delta_0(\theta_{lj}),$$

where

$$u_{lj} = \frac{1}{2\sigma} v_{lj}^2 (\mathbf{y} - \mathbf{X}_{(lj)} \boldsymbol{\beta}_{(lj)} - \xi \boldsymbol{v})' \mathbf{V} \mathbf{X}_{lj} b_{lj},$$

$$v_{lj}^2 = \left(\frac{1}{s^2} + \frac{1}{2\sigma} \mathbf{X}_{lj}' \mathbf{V} \mathbf{X}_{lj} b_{lj}^2 \right)^{-1},$$

and

$$r_{lj} = p(\theta_{lj} = 0|\text{rest}) = \frac{\pi_1}{\pi_1 + 2(1-\pi_1)(s^2)^{-\frac{1}{2}} (v_{lj}^2)^{\frac{1}{2}} \exp\left\{ \frac{u_{lj}^2}{2 v_{lj}^2} \right\} \left[\Phi(\frac{u_{lj}}{v_{lj}}) \right]}.$$

The full conditional distribution of each v_i is then an inverse Gaussian distribution,

$$v_i|\text{rest} \sim \text{InvGaussian}\left(\mu' = \frac{1}{|y_i - \boldsymbol{x}_i' \boldsymbol{\beta}|}, \lambda' = \frac{1}{2\sigma} \right).$$

The full conditional distribution of σ is given by

$$\sigma|\text{rest} \sim \text{InvGamma}\left(\frac{3n}{2} + g_1, \frac{1}{4} (\boldsymbol{\epsilon} - \xi \boldsymbol{v})' \mathbf{V} (\boldsymbol{\epsilon} - \xi \boldsymbol{v}) + \tau(1-\tau) \sum_{i=1}^{n} v_i + g_2 \right),$$

where $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$.

The full conditional distribution of π_0 and π_1 are given by

$$\begin{aligned} \pi_0|\text{rest} &\sim \text{Beta}(\#(\boldsymbol{b}_l = 0) + a_1, \#(\boldsymbol{b}_l \neq 0) + a_2) \\ \pi_1|\text{rest} &\sim \text{Beta}(\#(\theta_{lj} = 0) + c_1, \#(\theta_{lj} \neq 0) + c_2). \end{aligned}$$

Note that $\#(\cdot)$ denotes the cardinality or count of the set of elements for which the condition in $(\cdot)$ holds true.

The full conditional distribution of s^2 is given by

$$s^2|\text{rest} \sim \text{InvGamma}\left(1 + \frac{1}{2}\#(\theta_{lj} = 0), t + \frac{1}{2}\sum_{l,j}\theta_{lj}^2\right).$$

5.8 Bayesian sparse group LASSO QR with spike and slab prior for fixed and random effect selection

In this section, the primary focus is on extending the BSGSSQR method to enable the simultaneous selection of both fixed and random effects. This emphasis is due to the existing capability of BSGSSQR to select fixed effects at both the group and within-group levels when covariates have a grouped structure. To achieve this objective, the utilisation of linear mixed models (LMMs) is first elucidated, based on a decomposition for the covariance matrix of random effects in quantile mixed models. Subsequently, a comprehensive explanation of the proposed methodology within this context is provided.

5.8.1 Reparameterisation of random effects in LMMs

Recall the linear mixed model,

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\boldsymbol{u}_i + \epsilon_{ij}, \quad i = 1, \ldots, N, j = 1, \ldots, n_i, \tag{5.13}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and $\boldsymbol{u}_i \sim N(0, \Sigma_u)$. Kinney and Dunson (2007) reparameterised the random effects part of the model (5.13) as

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{D}\mathbf{A}\boldsymbol{c}_i + \epsilon_{ij},$$

where $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{iq})', \mathbf{D} = \text{diag}(d_1, \ldots, d_q)'$ and $\mathbf{A}$ is a $q \times q$ lower triangular matrix,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & & & 0 \\ a_{21} & 1 & & & \\ a_{31} & a_{32} & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ a_{q,1} & a_{q,2} & \cdots & a_{q,l'-1} & 1 \end{bmatrix},$$

whose free elements represent the correlations of each random effect. In this model, the vector $\boldsymbol{c}_i$ is assumed to follow $N(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \text{diag}(\omega_1, \ldots, \omega_q)'$. Based on this reparameterisation, the covariance decomposition of the random effects can be implied as

$$\Sigma_u = \mathbf{D}\mathbf{A}\boldsymbol{\Omega}\mathbf{A}'\mathbf{D}.$$

In this case, d_l serves as an analogy to the standard deviations of random effects. When $d_l = 0$, it leads to the exclusion of random effect l from the model.

Motivated by this reparameterisation, the quantile mixed model is proposed as:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}_\tau + \mathbf{z}'_{ij}\mathbf{D}\mathbf{A}\boldsymbol{c}_i + \epsilon_{ij},$$

where $\boldsymbol{\beta}_\tau$ is a vector of coefficients that depend on τ and ϵ_{ij} are independent, with their τth quantile level assumed to be zero. These are also presumed to follow the AL distribution.

Let $\epsilon_{ij} = y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta} - \mathbf{z}'_{ij}\mathbf{D}\mathbf{A}\boldsymbol{c}_i$. Following the scale mixture of normals as outlined in Section 3.4, the likelihood of ϵ_{ij} can be expressed as a scale mixture of normals with exponential mixing density by

$$\prod_{i=1}^{N}\prod_{j=1}^{n_i}\int_0^\infty \frac{1}{\sigma\sqrt{4\pi\sigma v_{ij}}}\exp\left\{-\frac{(y_i - \mathbf{x}'_i\boldsymbol{\beta} - \mathbf{z}'_{ij}\mathbf{D}\mathbf{A}\boldsymbol{c}_i - \xi v_{ij})^2}{4\sigma v_{ij}} - \zeta v_{ij}\right\}dv_{ij},$$

where $\xi = (1-2\tau)$, $\zeta = \tau(1-\tau)/\sigma$ and $v_{ij} \sim \text{Exp}(\zeta)$.

5.8.2 Bayesian sparse group LASSO mixed QR

Based on the BSGSSQR method detailed in Section 5.7 and the reparameterisation of random effects mentioned above, I propose the following l_1-penalised check function:

$$\min_{\boldsymbol{\beta}_\tau, \boldsymbol{c}_\tau} \sum_{i=1}^{N}\sum_{j=1}^{n_i}\rho_\tau(\epsilon_{ij}) + \lambda_1||\boldsymbol{\beta}_\tau||_1 + \lambda_2\sum_{l=1}^{G}||\boldsymbol{\beta}_{l,\tau}||_2 + \lambda_3\sum_{i=1}^{N}\sum_{l'=1}^{q}|c_{il'\tau}|, \tag{5.14}$$

where $\boldsymbol{c}_\tau = (\boldsymbol{c}'_{1,\tau}, \dots, \boldsymbol{c}'_{N,\tau})$ and $\lambda_3 \geq 0$. In this case, I follow Alhamzawi (2013) to restrict λ_3 to being 1. Here, it can be observed that the function (5.14) yields the sparse group LASSO estimates on $\boldsymbol{\beta}_{l,\tau}$ and the LASSO estimates on $\boldsymbol{c}_\tau$. These estimates can be interpreted as posterior mode estimates when $\boldsymbol{\beta}_{l,\tau}$ and $\boldsymbol{c}_\tau$ have independent and identical Laplace priors. Therefore, the Laplace prior $(1/2)\exp\{-|c_{il'\tau}|\}$ on $c_{il'\tau}, l' = 1, \dots, q$ (random effects) (Alhamzawi, 2013), can be employed, while the prior on $\boldsymbol{\beta}_{l,\tau}$ (fixed effects) can be specified in Section 5.7. In reality, a scale mixture of normals of the Laplace prior is exploited in the Gibbs sampler for the Bayesian analysis (Alhamzawi, 2013). This mixture can be represented as:

$$\frac{1}{2}\exp\{-|c_{il'\tau}|\} = \int_0^\infty \mathcal{N}(c_{il'\tau}; 0, \omega_{l'})\mathrm{Exp}\Big(\omega_{l'}; \frac{1}{2}\Big) d\omega_{l'}. \tag{5.15}$$

Henceforth, this approach is referred as Bayesian sparse group LASSO-mixed quantile regression with Spike and Slab (BSGSSMQR).

Priors specification for random effect part

Here, the prior for $\omega_{l'}$ can be assumed to follow Gamma(1,2), which is equivalent to Exp(1/2), as expressed in (5.15). Indeed, this prior is informative and convenient for eliciting a prior of $\omega_{l'}$ (Alhamzawi, 2013). For $d_{l'}$, it is specified to follow a zero-inflated standard half-normal prior, $\mathrm{ZI}-N^{+}(p_{l'0}, \mu_{d_{l'}} = 0, \sigma^2_{d_{l'}} = 1)$, where $p_{l'0} = \Pr(d_{l'} = 0)$. This prior can be expressed as:

$$f(d_l) = \begin{cases} p_{l'0} + (1-p_{l'0})\sqrt{2}e^{-\frac{(d_l)^2}{2}} & \text{if } d_l = 0 \\ (1-p_{l'0})\sqrt{2}e^{-\frac{(d_l)^2}{2}} & \text{if } d_l > 0 \\ 0 & \text{otherwise} \end{cases}$$

For $\boldsymbol{a} = (a_{l'r} : l' = 2, \ldots, q; r = 1, \ldots, l'-1)$, it is assumed to follow the prior $p(\boldsymbol{a}|\boldsymbol{d}) = N(\mathbf{0}, \boldsymbol{\Sigma}_a) \cdot 1(\boldsymbol{a} \in \mathcal{R}_{\boldsymbol{d}})$, where $\boldsymbol{\Sigma}_a$ is the variance-covariance matrix of $\boldsymbol{a}$, $1(\cdot)$ is an indicator function and $\mathcal{R}_{\boldsymbol{d}}$ enforced each element of $\boldsymbol{a}$ to be zero corresponding to random effects that are zero (Kinney & Dunson, 2007).

Gibbs Sampler

Let $i = 1, \ldots, N$ represent a sequence of N individuals, $j = 1, \ldots, n_i$ represent a sequence of repeated observations for individual i, $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})', \mathbf{y} = (y_{11}, \ldots, y_{Nn_N})', \boldsymbol{v} = (v_{11}, \ldots, v_{Nn_N})', \mathbf{V} = \mathrm{diag}(\boldsymbol{v}^{-1}), \boldsymbol{v}_i = (v_{i1}, \ldots, v_{in_i})', R_{ij} = \mathbf{z}'_{ij}\mathbf{D}\mathbf{A}\boldsymbol{c}_i, \mathbf{R} = (R_{11}, \ldots, R_{Nn_N})'$, $\boldsymbol{r}_{1ij} = (c_{il'}d_{m'}z_{ijm} : l' = 1, \ldots, q, m' = l'+1, \ldots, q)', \boldsymbol{r}_{2ij} = (z_{ijl'}(c_{il'} + \sum_{m'=1}^{l'-1} c_{im'}a_{m'l'}) : l' = 1, \ldots, q)', n = \sum_{i=1}^{N} n_i$.

The full conditional distribution of $\boldsymbol{b}_l$ is then a multivariate spike and slab distribution

$$\boldsymbol{b}_l|\text{rest} \sim (1-q_l)\mathcal{N}_{m_l}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) + q_l\delta_0(\boldsymbol{b}_l),$$

where

$$\boldsymbol{\mu}_l = \left(\frac{1}{2\sigma}\boldsymbol{\Sigma}_l^{\frac{1}{2}}\mathbf{K}_l^{\frac{1}{2}}\mathbf{X}_l'\mathbf{V}(\mathbf{y} - \mathbf{X}_{(l)}\mathbf{K}_{(l)}^{\frac{1}{2}}\boldsymbol{b}_{(l)}^{\frac{1}{2}} - \mathbf{R} - \xi\boldsymbol{v})\right),$$

$$\boldsymbol{\Sigma}_l = \left(\frac{1}{2\sigma}\mathbf{K}_l^{\frac{1}{2}}\mathbf{X}_l'\mathbf{V}\mathbf{X}_l\mathbf{K}_l^{\frac{1}{2}} + \mathbf{I}_{m_l}\right)^{-1},$$

and q_l be the posterior probability of $\boldsymbol{b}_l$ being zero given the remaining parameters and can be computed by

$$\begin{aligned} q_l &= P(\boldsymbol{b}_l = \mathbf{0}|\text{rest}) \\ &= \frac{\pi_0}{\pi_0 + (1-\pi_0)|\boldsymbol{\Sigma}_l|^{\frac{1}{2}} \exp\left\{\frac{1}{4\sigma^2}||\boldsymbol{\Sigma}_l^{\frac{1}{2}}\mathbf{K}_l^{\frac{1}{2}}\mathbf{X}_l'\mathbf{V}(\mathbf{y} - \mathbf{X}_{(l)}\mathbf{K}_{(l)}^{\frac{1}{2}}\boldsymbol{b}_{(l)}^{\frac{1}{2}} - \mathbf{R} - \xi\boldsymbol{v})||_2^2\right\}}. \end{aligned}$$

The full conditional distribution of θ_{lj} is then a spike and slab distribution, with the slab a positive part normal distribution

$$\theta_{lj}|\text{rest} \sim (1-r_{lj})N^+(u_{lj}, v_{lj}^2) + r_{lj}\delta_0(\theta_{lj}),$$

where

$$u_{lj} = \frac{1}{2\sigma}v_{lj}^2(\mathbf{y} - \mathbf{X}_{(lj)}\boldsymbol{\beta}_{(lj)} - \mathbf{R} - \xi\boldsymbol{v})'\mathbf{V}\mathbf{X}_{lj}b_{lj},$$

$$v_{lj}^2 = \left(\frac{1}{s^2} + \frac{1}{2\sigma}\mathbf{X}_{lj}'\mathbf{V}\mathbf{X}_{lj}b_{lj}^2\right)^{-1},$$

and

$$r_{lj} = p(\theta_{lj} = 0|\text{rest}) = \frac{\pi_1}{\pi_1 + 2(1-\pi_1)(s^2)^{-\frac{1}{2}}(v_{lj}^2)^{\frac{1}{2}}\exp\left\{\frac{u_{lj}^2}{2v_{lj}^2}\right\}\left[\Phi(\frac{u_{lj}}{v_{lj}})\right]}.$$

The full conditional distribution of $\boldsymbol{c}_i$ is a multivariate normal distribution

$$\boldsymbol{c}_i \sim \mathcal{N}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}),$$

where

$$\boldsymbol{\mu}_{c_i} = \frac{\boldsymbol{\Sigma}_{c_i}\mathbf{A}'\mathbf{D}\mathbf{Z}_i'\mathbf{V}_i}{2\sigma}\Big(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta} - \xi\boldsymbol{v}_i\Big)$$

and

$$\boldsymbol{\Sigma}_{c_i} = \left(\frac{\mathbf{A}'\mathbf{D}\mathbf{Z}_i'\mathbf{V}_i\mathbf{Z}_i\mathbf{D}\mathbf{A}}{2\sigma} + \boldsymbol{\Omega}^{-1}\right)^{-1}.$$

The full conditional distribution of $\boldsymbol{a}$ is a multivariate normal distribution

$$\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\Sigma}_{\boldsymbol{a}}) \cdot 1(\boldsymbol{a} \in \mathcal{R}_{\boldsymbol{d}}),$$

where

$$\boldsymbol{\mu}_{\boldsymbol{a}} = \boldsymbol{\Sigma}_{\boldsymbol{a}}\left(\sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{\boldsymbol{r}_{1ij}(y_{ij} - \mathbf{x}_{ij}'\boldsymbol{\beta} - \xi v_{ij})}{2\sigma v_{ij}}\right)$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{a}} = \left(\sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{\boldsymbol{r}_{1ij}\boldsymbol{r}_{1ij}'}{2\sigma v_{ij}} + \mathbf{A}_0^{-1}\right)^{-1}.$$

The full conditional distribution of each $d_{l'}(l' = 1, \ldots, q)$ is a zero-inflated half-normal distribution

$$d_{l'} \sim \mathrm{ZI} - N^{+}(\hat{p}_{l'}, \mu_{d_{l'}}, \sigma^2_{d_{l'}}),$$

where

$$\mu_{d_{l'}} = \sigma^2_{d_{l'}} \left(\sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{r_{2ij}(y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta} - \sum_{s \neq l'} r_{2ijs} d_s - \xi v_{ij})}{2\sigma v_{ij}} \right),$$

$$\sigma^2_{d_{l'}} = \left(\sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{r^2_{2ij}}{2\sigma v_{ij}} + 1 \right)^{-1},$$

and

$$\hat{p}_{l'} = \left(1 + \frac{(1 - p_{l'0}) N(0; 0, 1)(1 - \Phi(0; \mu_{d_{l'}}, \sigma^2_{d_{l'}}))}{p_{l'0} N(0; \mu_{d_{l'}}, \sigma^2_{d_{l'}})(1 - \Phi(0; 0, 1)} \right)^{-1}.$$

Here, $\Phi(\cdot)$ is the cumulative normal distribution function.
The full conditional distribution of each v_i is then an inverse Gaussian distribution,

$$v_i | \text{rest} \sim \text{InvGaussian}\left(\mu' = \sqrt{\frac{1}{(y_i - \boldsymbol{x}'_{ij}\beta - \boldsymbol{z}'_{ij}\mathbf{D}\mathbf{A}\boldsymbol{c}_i)^2}}, \lambda' = \frac{1}{2\sigma} \right).$$

The full conditional distribution of σ is

$$\sigma | \text{rest} \sim \text{InvGamma}\left(\frac{3n}{2} + g_1, \frac{1}{4}(\boldsymbol{\epsilon} - \xi \boldsymbol{v})' \mathbf{V} (\boldsymbol{\epsilon} - \xi \boldsymbol{v}) + \tau(1 - \tau) \sum_{i=1}^{n} v_i + g_2 \right),$$

where $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{R}$.

The full conditional distribution of each $\omega_{l'}(l' = 1, \ldots, q)$ is then a generalised inverse Gaussian distribution (GIG),

$$\omega_{l'} \sim \text{GIG}(\tilde{\nu}, \tilde{\chi}, \tilde{\psi}),$$

where $\tilde{\nu} = -(N + 2)/2, \tilde{\chi}^2 = \sum_{i=1}^{N} c^2_{il'}$ and $\tilde{\psi}^2 = N$.

The full conditional distribution of π_0 and π_1 is given by

$$\pi_0 | \text{rest} \sim \text{Beta}(\#(\boldsymbol{b}_l = 0) + a_1, \#(\boldsymbol{b}_l \neq 0) + a_2),$$
$$\pi_1 | \text{rest} \sim \text{Beta}(\#(\theta_{lj} = 0) + c_1, \#(\theta_{lj} \neq 0) + c_2).$$

The full conditional distribution of s^2 is given by

$$s^2 | \text{rest} \sim \text{InvGamma}\left(1 + \frac{1}{2}\#(\theta_{lj} = 0), t + \frac{1}{2} \sum_{l,j} \theta^2_{lj} \right).$$

5.9 Variable selection process

The Bayesian hierarchical QR models discussed earlier typically yield posterior mean estimators. Nevertheless, as highlighted by Xu and Ghosh (2015), these estimators are not inherently effective in sufficiently shrinking the regression coefficients toward zero. Consequently, these posterior mean estimators are less suited for the specific purpose of variable selection. While methods such as the utilisation of posterior credible intervals (L. Zhang et al., 2014) can potentially address this limitation, they do introduce additional complexity to the analysis. To avoid this complexity, an alternative way is to employ the posterior median (Xu & Ghosh, 2015) as an adaptive thresholding estimator. This approach presents distinct merits in terms of both selection and estimation. Xu and Ghosh (2015) showed that the use of posterior median thresholding possesses an oracle property under orthogonal designs, meaning that this thresholding correctly identifies which predictors should be included in a model and which should be excluded. Therefore, within this thesis, I advocate for the adoption of the posterior median estimator as the preferred method for the variable selection process within the domain of quantile regression.

5.10 Simulation study for fixed effects selection

To assess the efficacy of the proposed methodologies in Sections 5.6 and 5.7, a series of simulation studies were conducted. These studies utilised only independent generated data, not correlated data or longitudinal data presented in Chapter 4, and considered only homogeneous variance of errors. However, some studies, specifically the second and fourth, shared a commonality with Chapter 4, such as the presence of a non-linear component in the model. Furthermore, the predictors were grouped in two scenarios: independent and correlated. In this context, the number of group variables (G) was adjusted based on the specific contextual requirements of each simulation study. For each study, a sample size of 100 was employed. Subsequently, the model was fitted at three different quantiles, specifically $\tau \in (0.10, 0.50, 0.90)$. Concerning the prior specifications of our proposed approaches (BGLSSQR and BSGSSQR), I chose to set $a_1 = a_2 = c_1, = c_2 = 1$ for the beta priors on both π_0 and π_1. Additionally, the inverse Gamma priors applied to σ were configured with $g_1 = g_2 = 0.1$. For all Bayesian approaches, a Gibbs sampler was employed, running for a total of 50,000 iterations. To account for convergence issues, an initial burn-in period of 10,000 iterations was implemented. In this simulation study, thinning is not taken into consideration, as described in Chapter 3, since it can result in a loss of information and reduce the precision of the MCMC algorithm.

As is commonly known, LASSO introduces bias into parameter estimates by shrinking the coefficients of some variables toward zero. Thus, I would like to investigate this property in our proposed method as well. Consequently, the first two simulation studies aim to examine the estimation performance of two proposed approaches (BGLSSQR and BSGSSQR) in comparison with both existing frequentist and Bayesian methods. In terms of the existing frequentist approaches, I considered the classical quantile regression (CQR) as introduced by Koenker (2005), and both LASSO types for quantile regression, i.e. LASSO (LASSOQR) and group LASSO (GLASSOQR) approaches, as presented by Sherwood et al. (2023). The CQR was implemented using the `rq` function from the R package `quantreg` (Koenker, 2021), while the LASSOQR and GLASSOQR approaches were implemented using the `rq.pen(penalty = "LASSO")` function and `rq.group.pen(penalty = "gLASSO")` function, respectively, from the R package `rqPen` (Sherwood et al., 2023). It is important to note that the tuning parameter (λ) for both the LASSO and GLASSOQR methods was chosen based on the Bayesian information criterion (BIC) using the `qic.select(method="BIC")` function (Sherwood et al., 2023). Regarding the existing Bayesian approaches, I considered the Bayesian LASSO quantile regression (BLASSOQR) as reported in Alhamzawi and Ali (2020). The `BLqr` function from the R package `Brq` (Alhamzawi, 2020) was employed to perform the BLASSOQR approach. For both studies, three distinct error $\boldsymbol{\epsilon}$ distributions were considered: a standard normal distribution, a student-t distribution with $\nu = 3$ ($t_{\nu=3}$) and a Chi-squared distribution with $\nu = 3$ ($\chi^2_{\nu=3}$). I summarised the performance of each approach using two metrics in 500 simulations, Mean Bias Error (MBE) and Root-Mean-Square Error (RMSE):

$$\text{MBE}(\beta_{\tau,h}) = \frac{1}{500}\sum_{r=1}^{500}\left(\hat{\beta}_{\tau,h} - \beta_{\tau,h}\right), \quad h = 1, \ldots, p, \ \tau = 0.10, 0.50, 0.90,$$

and

$$\text{RMSE}(\beta_{\tau,h}) = \sqrt{\frac{1}{500}\sum_{r=1}^{500}\left(\hat{\beta}_{\tau,h} - \beta_{\tau,h}\right)^2}, \quad h = 1, \ldots, p, \ \tau = 0.10, 0.50, 0.90,$$

where $\hat{\beta}_h$ and β_h are estimates and true coefficients in three quantile models.

Our focus now shifts to the evaluation of subset selection and predictive performance in the remaining studies, where I compared our proposed approaches with existing variable selection methods. In particular, I began with two existing frequentist approaches, i.e. LASSOQR and GLASSOQR. Subsequently, I delved into the comparison of relevant existing Bayesian variable selection methods. One such existing method is the model selection based on credible intervals (MSCI) employed for the BLASSOQR approach. The MSCI method was implemented using the `model()` function provided by the R package `Brq` (Alhamzawi, 2020). In these simulation studies, I restricted our study to the standard normal distribution for the errors.

Table 5.1: Confusion Matrix

		True	
		# ($\beta_p \neq 0$)	# ($\beta_p = 0$)
Predicted	# ($\hat{\beta}_p \neq 0$)	True Positive (TP)	False Positive (FP)
	# ($\hat{\beta}_p = 0$)	False Negative (FN)	True Negative (TN)

$$\textbf{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \textbf{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \qquad F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

In the context of subset selection, the assessment of each approach was based on two rates and one score: the true positive rate (TPR), the false positive rate (FPR) and the F_1 score. The TPR, also known as recall, measures the proportion of actual positives correctly identified by the model, while the FPR measures the proportion of actual negatives incorrectly identified as positives. Meanwhile, the F_1 score is the harmonic mean of precision (the proportion of true positive predictions among all positive predictions) and recall. As is known, both precision and recall represent different aspects of a classification model's performance, and sometimes one may be more important than the other depending on the specific application or problem. Hence, the F_1 score provides a single metric that balances both precision and recall. The TPR, FPR, F_1 score were calculated by utilising the information encapsulated within the confusion matrix, as presented in Table 5.1. Additionally, to assess the predictive performance, I employed the mean weighted absolute errors or MAE as described in Chapter 4.

5.10.1 Study 5.1

Aim

This simulation study was conducted to investigate the behaviour of parameter estimates obtained by using the proposed methods, compared with existing approaches, under the simplest quantile models (linear) where each predictor was designed to be independent of the others. Furthermore, I considered to assess additional performance of BGLSSQR and BSGSSQR when choosing the number of updates and iterations for estimating the hyperparameters λ and t through a Monte Carlo EM algorithm in two distinct scenarios: 100 and 1000. In this study, I only considered groups of continuous predictors as an initial investigation.

Data generation

In this simulation study, I considered the true model to be represented by the model (5.1), which comprised a structure of group variables. Specifically, I set the number of variable groups (G) to be 3. Among these groups, the first one was designed as the "active group", where each coefficient within the group took a value of one, indicating its significance. In contrast, the last group was designed as the "inactive group", wherein each coefficient within the group took a value of zero, indicating its lack of contribution to the response variable. Meanwhile, the second group was designed to contain at least one inactive coefficient. More specifically, the active coefficients were established with magnitudes categorised as weak ($\beta = 0.30$), medium ($\beta = 0.5$) and strong ($\beta = 1$). Thus, the true regression coefficients were set to

$$\boldsymbol{\beta} = (\mathbf{1}, (0.5, 0.3, 0), \mathbf{0}),$$

where $\mathbf{1}$ and $\mathbf{0}$ were the 1 and 0 vectors of length 3, respectively. A group structure with a length of 3 was chosen as the minimal structure capable of accommodating three magnitudes of coefficients. Each predictor was generated to follow the i.i.d. standard normal distribution.

Results

In scenarios involving independent predictors (Study 1), Figures 5.1 to 5.3 show the violin plot of estimates for eight methods. The figures indicate that classical quantile regression (CQR) performed exceptionally well, consistently providing estimates closer to the true parameters than all other methods. This evidence further confirmed by Figure 5.4, which indicates that the Mean Bias Error (MBE) of CQR approaches close to zero. This result remained consistent across three distinct quantile models, various error distributions, and coefficients of varying magnitudes (weak, medium, and strong). Both the frequentist and Bayesian LASSO methods, including the proposed methods (BGLSSQR and BSGSSQR), tended to produce values that were consistently underestimated.

All methods were sensitive to the assumptions of the residual distribution, especially in the case of non-negative errors like those of the Chi-squared distribution. However, they exhibited relatively robust performance with heavy-tailed errors, such as the t distributions.

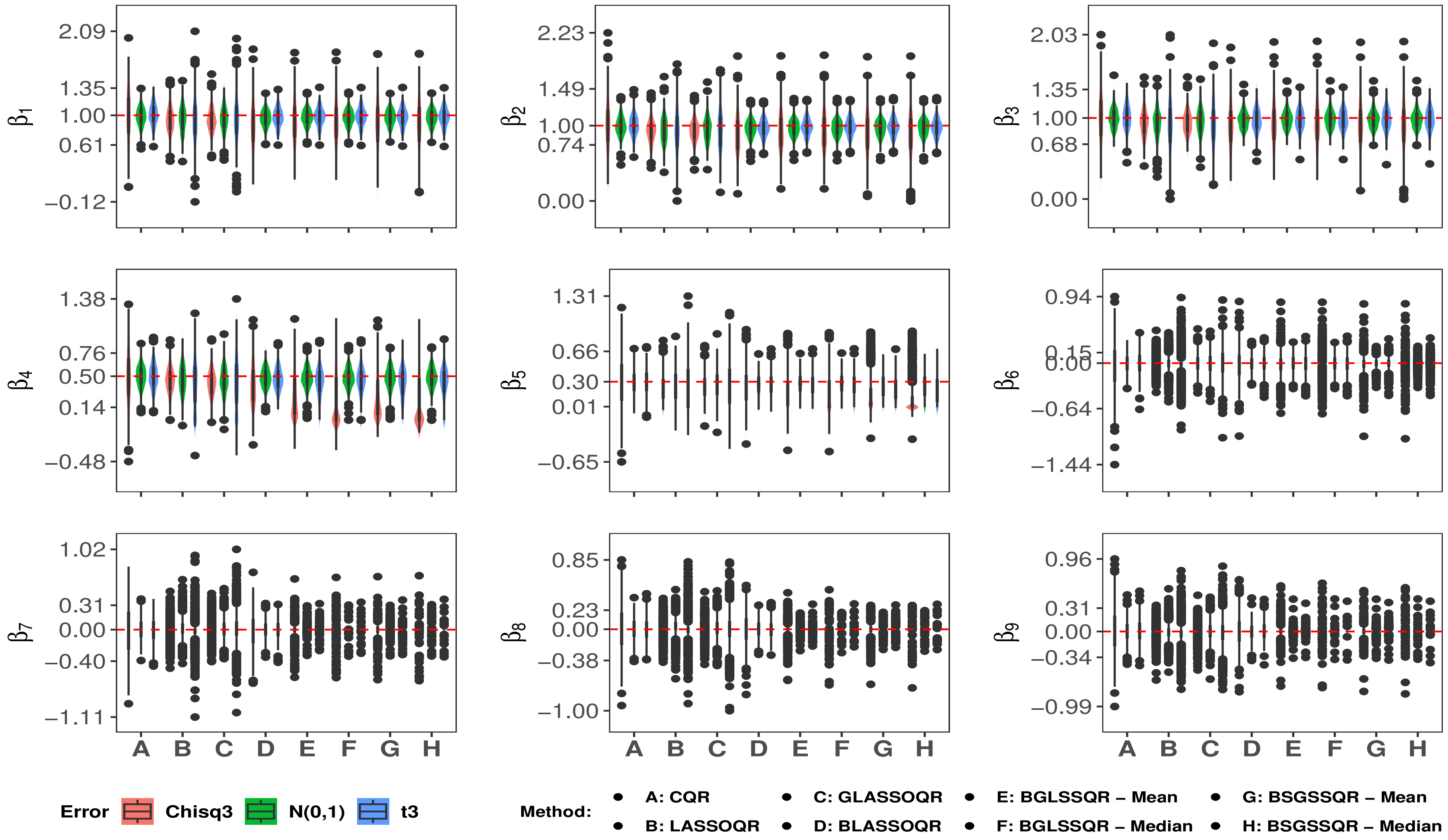


Figure 5.1: Estimates or posterior mean (median) estimates for the 0.10th quantile model obtained using eight different methods across three distinct scenarios of error distribution under the Study 5.1 setting.

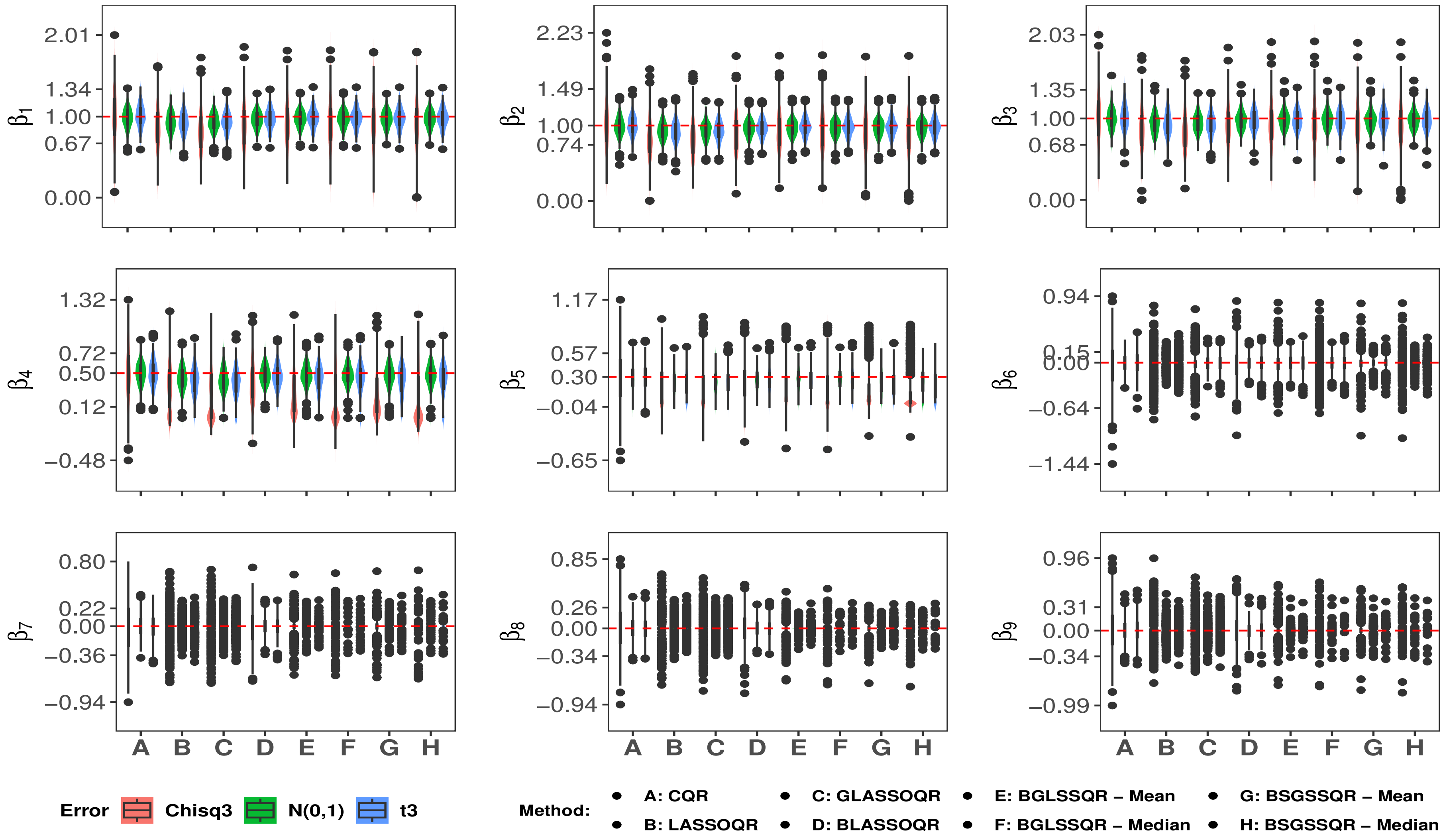


Figure 5.2: Estimates or posterior mean estimates for the 0.50th quantile model obtained using eight different methods across three distinct scenarios of error distribution under the Study 5.1 setting.

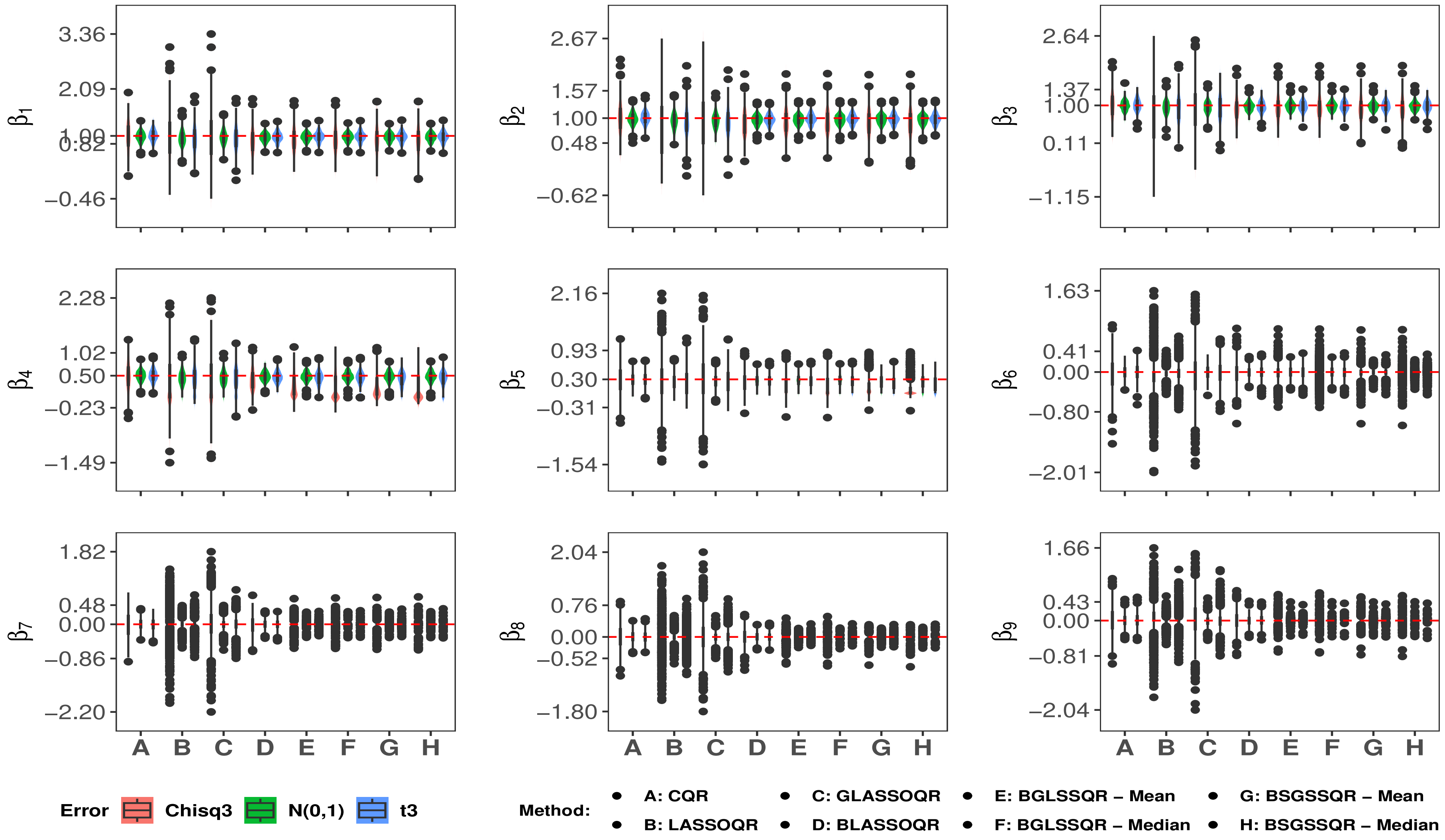


Figure 5.3: Estimates or posterior mean estimates for the 0.90th quantile model obtained using eight different methods across three distinct scenarios of error distribution under the Study 5.1 setting.

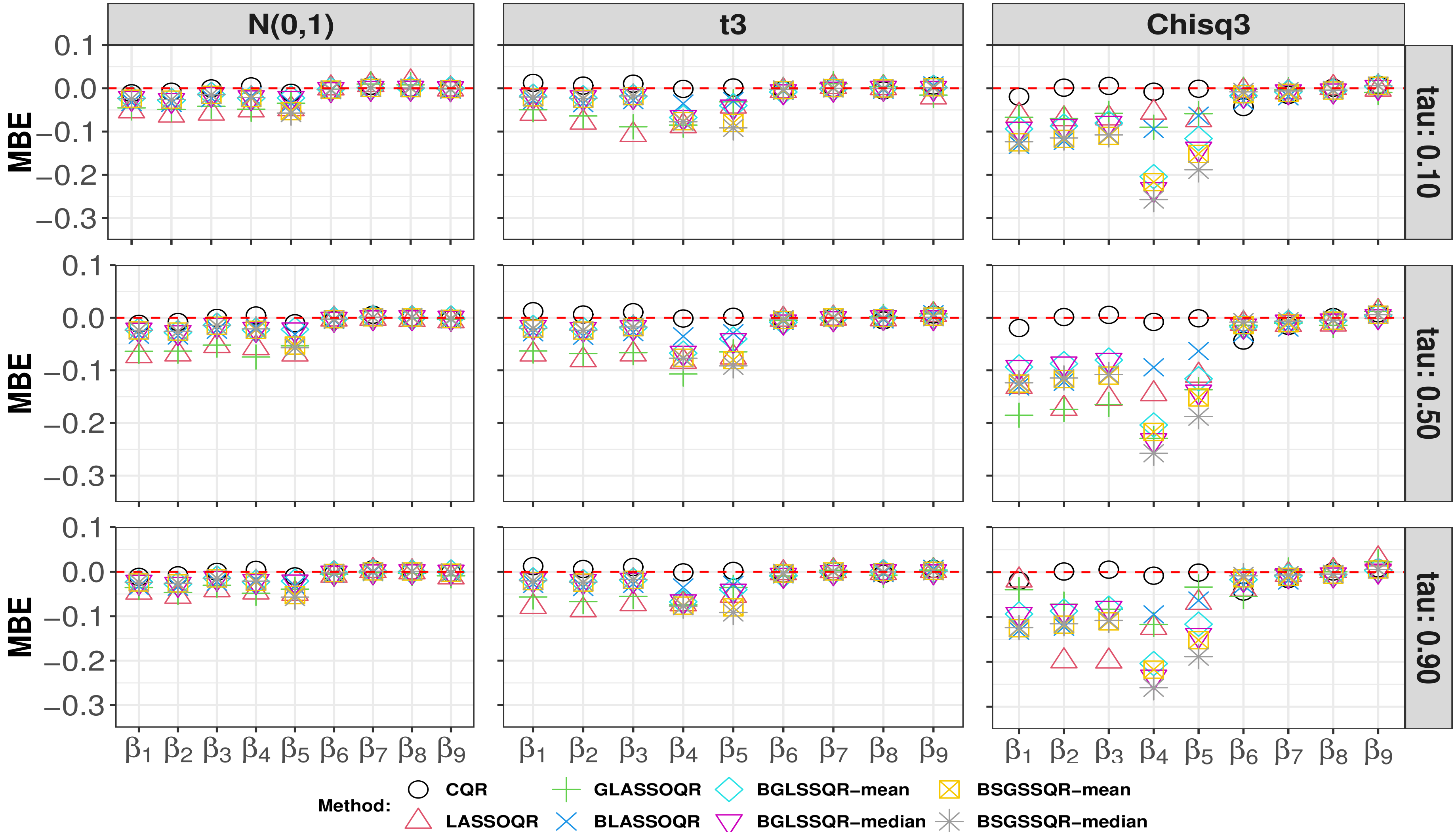


Figure 5.4: Mean Bias error (MBE) obtained from eight different methods across three quantiles and three scenarios of error distribution under the Study 5.1 setting.

Furthermore, each method performed well with a group of variables that only contained a strong magnitude ($\boldsymbol{\beta} = 1$) of active coefficients, with the exception of the LASSOQR and GLASSOQR methods. These two methods tended to introduce greater variability in the estimates compared to other approaches. When a group of variables included a mix of medium ($\boldsymbol{\beta} = 0.50$) and small ($\boldsymbol{\beta} = 0.30$) magnitudes of active coefficients, as well as inactive coefficient ($\boldsymbol{\beta} = 0$), the proposed methods (BGLASSQR and BSGSSQR) still performed commendably with medium magnitudes. Nevertheless, they underestimated small active coefficients, often estimating them closer to zero rather than their true values. Notably, the BSGSSQR method, particularly with posterior median estimates, outperformed all other methods in estimating inactive coefficients. Regarding a group of variables containing only inactive coefficients, both the BGLASSQR and BSGSSQR methods predominantly estimated inactive coefficients as zero, while others tended to estimate coefficients close to zero. Both frequentist LASSO methods (LASSOQR and GLASSOQR) did not succeed in this respect. The numerical results involving this simulation study are presented in Tables C.1 to C.6 in Appendix C.

Moreover, there was no discernible difference in performance between 100 and 1000 iterations when estimating λ and t via the Monte Carlo EM method, particularly in instances of a standard normal error distribution and a χ^2_3 error distribution. Nonetheless, slight differences in estimated bias values were noted between 100 and 1000 iterations in case of a t_3 error distribution, with the latter seemingly providing estimates closer to the true parameter values than the former. The corresponding results are presented in Tables C.7 - C.12 in Appendix C.

Summary

When focusing on linear quantile models with independent predictors, all methods exhibited sensitivity to non-negative errors resulting in biased estimates with high variability. Nevertheless, they demonstrated robustness against heavy-tailed errors. The classical quantile regression (CQR) consistently provided estimates that were closest to the true parameters across various scenarios. Despite this, most methods, including the proposed ones (BGLSSQR and BSGSSQR), tended to underestimate values. This trend of underestimation persisted across a range of situations, including different error distributions and coefficient magnitudes, and it prevailed even with a mix of active and inactive coefficients.

5.10.2 Study 5.2

Aim

The aim of this study was to investigate the behaviour of parameter estimates in a manner similar to Study 5.1, but under a more complex model, such as a non-linear model with polynomial and dummy predictors for categorical predictors, where some predictors were assumed to exhibit multicollinearity.

Data generation

Within this simulation framework, I adopted one of the simulation designs outlined by Yuan and Lin (2006) to define the true model. The true model encompassed a group of polynomial basis predictors representing the continuous covariate, as well as a group of dummy variables representing the multi-level categorical covariate. Analogous to Study 1, I classified the magnitude of the active coefficients into categories of weak ($\beta \in (0.30, 0.33)$), medium ($\beta \in (0.50, 0.67)$) and strong ($\beta \in (1, 2)$). Thus, the true model was

$$\mathbf{y} = \frac{2}{3}X_1 - X_1^2 + \frac{1}{3}X_1^3 + \frac{1}{2}X_{2,1} + \frac{3}{10}X_{2,2} + 2I(X_4 = 1) + I(X_4 = 2).$$

Here, X_1 and X_3 were generated by $X_p = \frac{Z_p + W}{\sqrt{2}}, Z_p \sim N(0,1)$, and $W \sim N(0,1)$. This implies that both X_1 and X_3 were correlated with each other. Each predictor of X_2 (i.e. $X_{2,1}$, $X_{2,2}$, and $X_{2,3}$) was generated in the same way as in Study 5.1. X_4 was generated by trichotomising X_1 as 0, 1, 2 under the conditions that $X_1 < \Phi^{-1}(1/3) = 0, X_1 > \Phi^{-1}(2/3) = 1$ and otherwise, it is to 2. Therefore, the true regression coefficients of this model were

$$\boldsymbol{\beta} = \left(\left(\frac{2}{3}, -1, \frac{1}{3}\right), \left(\frac{1}{2}, \frac{3}{10}, 0\right), \mathbf{0}, (2, 1) \right).$$

Results

In case of a multicollinearity setting, Figures 5.5 to 5.7 illustrate that the classical quantile regression (CQR) consistently performed well, yielding less biased estimators (further confirmed by Figure 5.8) compared to other methods across three distinct quantile models, three error distributions, and three levels of coefficient magnitudes. Similar to Study 5.1, the estimates from both the frequentist (LASSOQR and GLASSOQR) and Bayesian LASSO (BLASSOQR, BGLSSQR and BSGSSQR) methods exhibited bias.

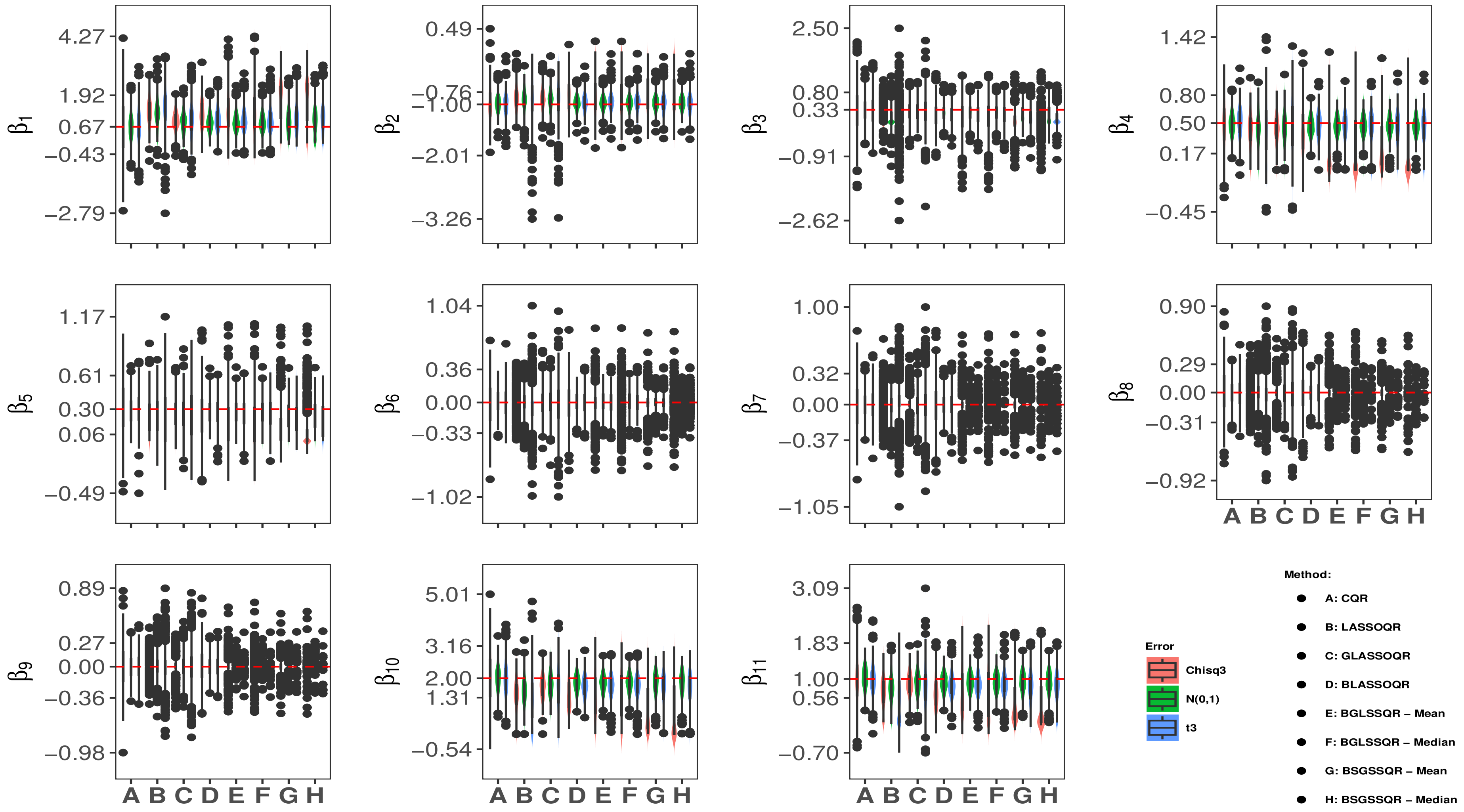


Figure 5.5: Estimates or posterior mean (median) estimates for the 0.10th quantile model obtained using six different methods across three distinct scenarios of error distribution under the Study 5.2 setting.

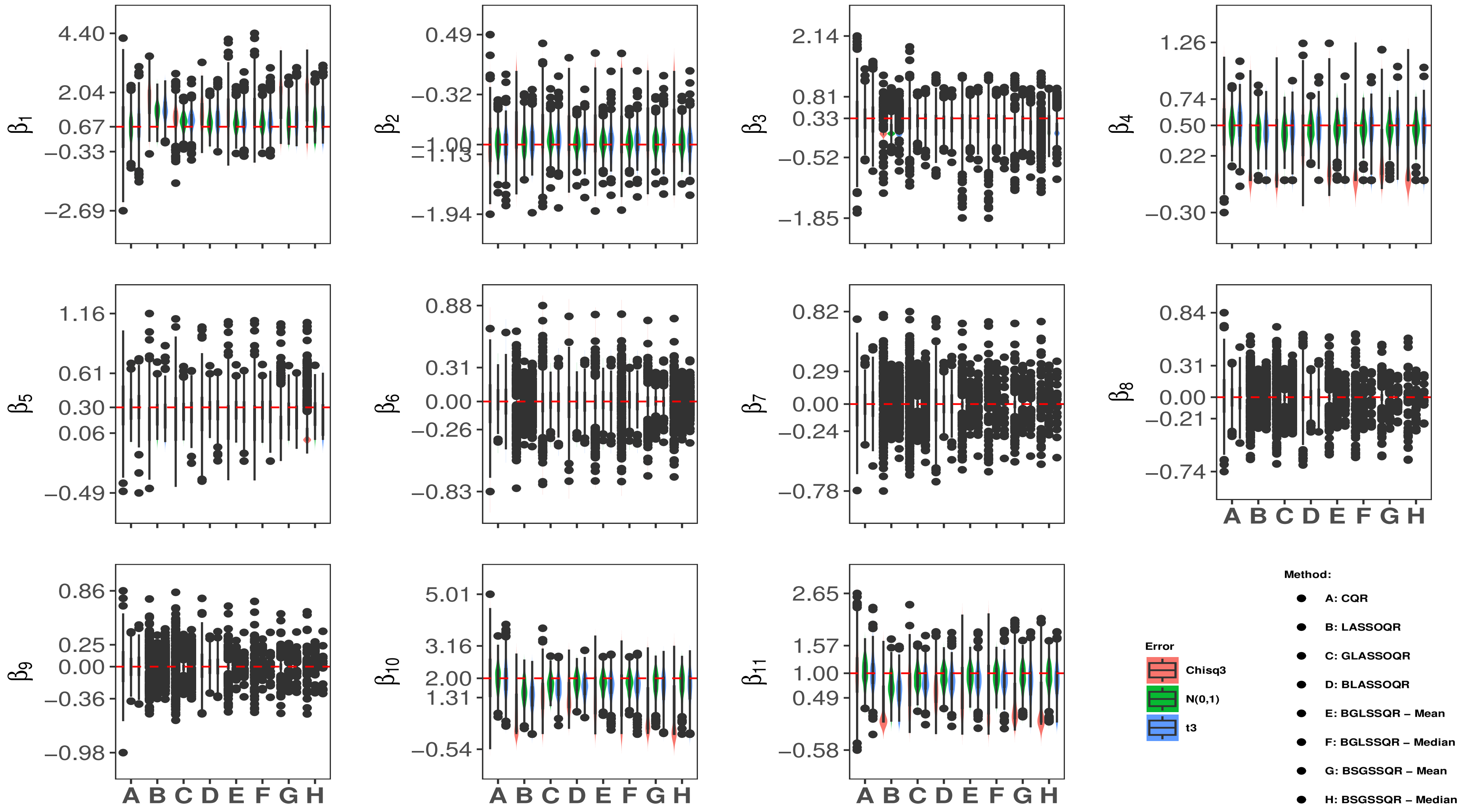


Figure 5.6: Estimates or posterior mean (median) estimates for the 0.50th quantile model obtained using six different methods across three distinct scenarios of error distribution under the Study 5.2 setting.

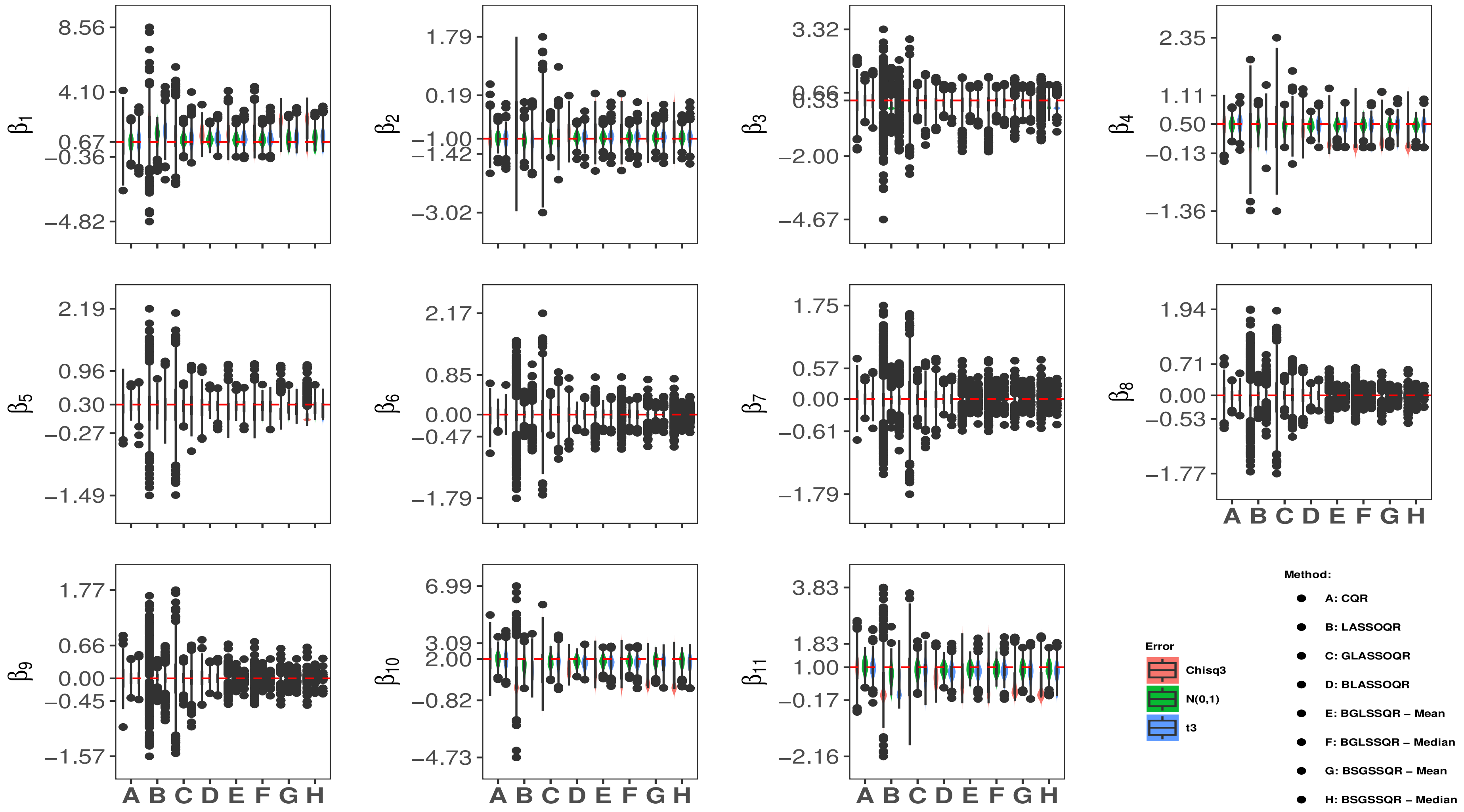


Figure 5.7: Estimates or posterior mean (median) estimates for the 0.90th quantile model obtained using six different methods across three distinct scenarios of error distribution under the Study 5.2 setting.

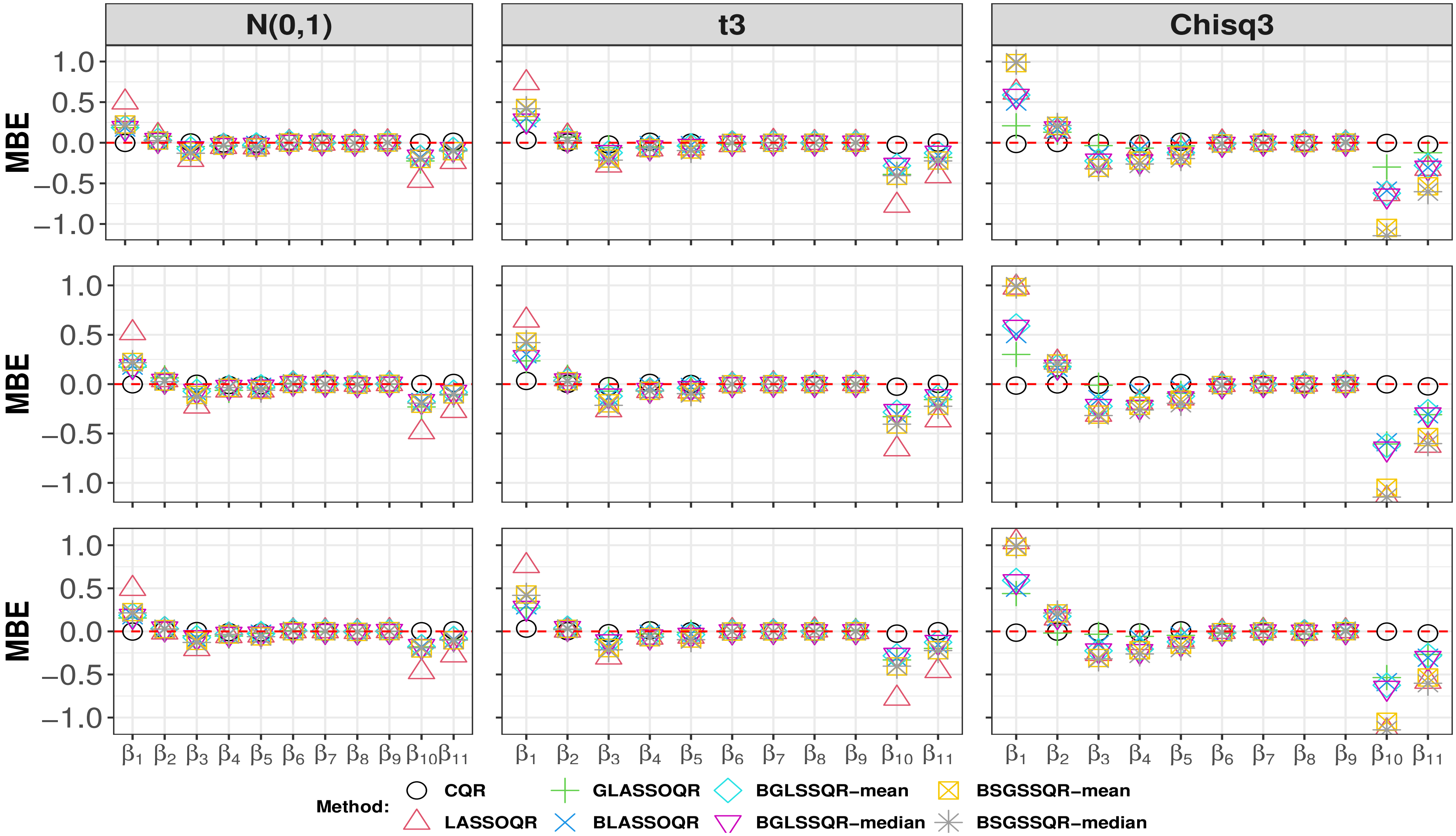


Figure 5.8: Mean Bias error (MBE) obtained from eight different methods across three quantiles and three scenarios of error distribution under the Study 5.2 setting.

All methods remained sensitive to the case of non-negative errors ($\chi^2_{\nu=3}$). Estimates from this scenario for each method showed a large bias and greater variability compared to other error types, such as standard normal and heavy-tailed errors ($t_{\nu=3}$).

When examining a scenario involving a group of variables with a polynomial degree of 3, all methods yielded varying results, depending on magnitudes rather than the sign of active coefficients. For a strong magnitude ($\beta = -1$), each method provided relatively similar estimates. In contrast, for a medium magnitude ($\beta = 0.67$), most estimates, particularly those from the frequentist and Bayesian LASSO methods, including the proposed methods (BGLSSQR and BSGSSQR), were predominantly overestimated. Moreover, in the case of a small magnitude ($\beta = 0.33$), the estimates from both the frequentist and Bayesian LASSO methods were underestimated.

Furthermore, in a scenario where a group of variables was linear and included both active and inactive coefficients with medium and small magnitudes, the frequentist and Bayesian LASSO methods underestimated both magnitudes of active coefficients. Notably, the BSGSSQR method, particularly with posterior median estimates, outperformed all other methods in estimating inactive coefficients.

In the case of dummy predictors representing categorical predictors, all methods consistently yielded estimates close to true values, except for the LASSOQR method. Notably, most estimates from all Bayesian methods, including the proposed methods, were mostly underestimated.

Additionally, when a group of variables encompassed only inactive coefficients, both the BGLSSQR and BSGSSQR methods, particularly with their posterior median estimates, outperformed other methods.

All corresponding numerical outcomes of this study are reported in Tables C.13 to C.18 in Appendix C.

Summary

In scenarios involving non-linear quantile models with mixed predictor types and exhibiting multicollinearity, all methods remained sensitive to non-negative errors whilst demonstrating robustness against heavy-tailed errors. Similar to Study 5.1, classical quantile regression (CQR) consistently produced estimates closest to the true parameters across various scenarios. Both BGLSSQR and BSGSSQR continued to yield underestimates across a range of situations, including different error distributions and coefficient magnitudes, and this trend prevailed even with a mix of active and inactive coefficients. This suggests that the proposed methods, akin to other regularisation methods, provided biased estimates as anticipated.

5.10.3 Study 5.3

Aim

This study was specifically conducted to assess the performance of both subset selection and prediction when the true quantile models were assumed to be linear, with multicollinearity present among a high number of predictors within group variables.

Data generation

In this particular scenario, I considered an expansion of the variable groups to 4, with each group comprising 5 predictors. The true regression coefficients were denoted as

$$\boldsymbol{\beta} = ((0.3, -1, 0, 0.5, 0.01), \mathbf{0}, (0.8, 0.8, 0.8, 0.8, 0.8), \mathbf{0}),$$

where $\mathbf{0}$ was the zero-vector of length 5. Predictors were generated to follow a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance $\boldsymbol{\Sigma}_x$, where off-diagonal elements of $\boldsymbol{\Sigma}_x$ was defined to $0.5^{|i-j|}$ for $i \neq j$. This implies that predictors closer in sequence have a higher correlation.

Results

From Table 5.2, in terms of predictive performance, the BLASSOQR method outperformed other methods across each linear quantile models. In particular, the predictive capabilities of both BGLSSQR and BSGSSQR methods, whether based on posterior mean or median, appeared to be on par with those of the BLASSOQR method. In contrast, the frequentist approaches (LASSOQR and GLASSOQR) yielded higher MAE values at the extreme quantile models compared to the performance of the 0.50th quantile model (median model).

In summarising the model selection accuracy, Bayesian approaches with spike and slab priors (BGLSSQR and BSGSSQR) exhibited superior performance compared to other methods across three quantile models, yielding higher TPR and F_1 scores, along with acceptable FPR. Notably, BGLSSQR outperformed all Bayesian methods in terms of TPR but exhibited a higher FPR compared to BSGSSQR. Conversely, BSGSSQR showcased a lower FPR and higher F_1 score while maintaining an acceptable TPR. Interestingly, both LASSOQR and GLASSOQR demonstrated high TPR values, but they exhibited higher FPRs and lower F_1 scores compared to the Bayesian methods. This variation in performance could be attributed to the choice of the tuning parameter.

Table 5.2: Mean MAE, true positive (TPR) rate, false positive (FPR) rate and F_1 score for five different methods under Study 5.3

τ		**LASSOQR**	**GLASSOQR**	**BLASSOQR**	**BGLSSQR**		**BSGSSQR**	
					mean	median	mean	median
0.10	MAE (SD)	0.42 (0.04)	0.41 (0.04)	**0.35 (0.03)**	0.36 (0.03)	0.37 (0.03)	0.36 (0.03)	0.37 (0.03)
	TPR	0.94	**1.00**	0.83	**1.00**		0.94	
	FPR	0.55	0.78	**0.04**	0.11		0.06	
	F_1	0.73	0.70	0.88	**0.94**		**0.94**	
0.50	MAE (SD)	0.36 (0.03)	0.37 (0.04)	**0.35 (0.03)**	0.36 (0.03)	0.37 (0.03)	0.36 (0.03)	0.37 (0.03)
	TPR	0.93	**1.00**	0.83	**1.00**		0.94	
	FPR	0.37	0.39	**0.04**	0.11		0.06	
	F_1	0.79	0.83	0.88	**0.94**		**0.94**	
0.90	MAE (SD)	0.42 (0.04)	0.41 (0.04)	**0.35 (0.03)**	0.36 (0.03)	0.37 (0.03)	0.36 (0.03)	0.37 (0.03)
	TPR	0.94	**1.00**	0.83	**1.00**		0.94	
	FPR	0.57	0.76	**0.04**	0.11		0.06	
	F_1	0.72	0.70	0.88	**0.94**		**0.94**	

Summary

In scenarios focusing on linear quantile models with multicollinearity among a high number of predictors within group variables, the BLASSOQR method excelled in predictive performance across all quantile models in diverse simulation settings. It was closely followed by BGLSSQR and BSGSSQR methods. In contrast, frequentist methods LASSOQR and GLASSOQR, displayed higher prediction errors, especially in extreme quantile models. Regarding model selection accuracy, the proposed methods outperformed other approaches across various quantile models and simulation studies. Specifically, BSGSSQR balanced a lower FPR with a superior F_1 score, while BGLSSQR exhibited an elevated FPR. The frequentist methods, LASSOQR and GLASSOQR, failed to control FPR effectively, consequently yielding an unacceptable F_1 score.

5.10.4 Study 5.4

Aim

This simulation study aimed to assess the performance of subset selection and prediction, similar to that in Study 5.3. However, unlike in previous studies, the true model here was more complex and exhibited high sparsity.

Data generation

For the high sparsity design, I expanded the number of variable groups to 10. The true model was the same as Study 5.2. Thus, the true regression coefficients of this model were

$$\boldsymbol{\beta} = \Bigg(\underbrace{\left(\frac{2}{3}, -1, \frac{1}{3}\right)}_{X_1}, \underbrace{\left(\frac{1}{2}, \frac{3}{10}, 0\right)}_{X_2}, \underbrace{\mathbf{0}}_{X_3}, \underbrace{(2,1)}_{X_4}, \underbrace{\mathbf{0}}_{X_5}, \underbrace{\mathbf{0}}_{X_6}, \underbrace{\mathbf{0}}_{X_7}, \underbrace{\mathbf{0}}_{X_8}, \underbrace{\mathbf{0}}_{X_9}, \underbrace{\mathbf{0}}_{X_{10}} \Bigg).$$

where $\mathbf{0}$ denotes a zero vector of length 3 for continuous variables and length 2 for categorical variables. I generated each predictor as follow: For a third-degree polynomial of X_1, X_5, X_7, I generated them by

$$X_p = \frac{Z_p + W}{\sqrt{2}},$$

where $Z_p \sim N(0,1)$ and $W \sim N(0,1)$. Specially, each predictor of X_2, X_3, X_9 and X_{10} was generated from $\mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma}_{X.})$, where $\boldsymbol{\Sigma}_{X.} = \text{diag}(1,1,1)$. For each dummy predictor of X_4, X_6 and X_8, I trichotomised X_1, X_5, X_7 to meet the conditions that $X_p < \Phi^{-1}(1/3) = 0, X_p > \Phi^{-1}(2/3) = 1$, and otherwise, equal to 2.

Results

Table 5.3: Mean MAE, true positive (TPR) rate, false positive (FPR) rate and F_1 score for five different methods under Study 5.4

τ		LASSOQR	GLASSOQR	BLASSOQR	BGLSSQR		BSGSSQR	
					mean	median	mean	median
0.10	MAE (SD)	0.42 (0.04)	0.42 (0.05)	**0.33 (0.03)**	0.36 (0.03)	0.37 (0.03)	0.36 (0.03)	0.37 (0.03)
	TPR	0.87	0.98	0.77	**0.99**		0.92	
	FPR	0.48	0.50	**0.03**	0.08		**0.03**	
	F_1	0.55	0.62	0.83	0.90		**0.91**	
0.50	MAE (SD)	0.36 (0.04)	0.39 (0.04)	**0.33 (0.03)**	0.36 (0.03)	0.37 (0.03)	0.36 (0.03)	0.37 (0.03)
	TPR	0.84	0.98	0.77	**0.99**		0.92	
	FPR	0.23	0.20	**0.03**	0.07		**0.03**	
	F_1	0.65	0.79	0.83	0.90		**0.91**	
0.90	MAE (SD)	0.42 (0.04)	0.42 (0.04)	**0.33 (0.03)**	0.36 (0.03)	0.37 (0.03)	0.36 (0.03)	0.37 (0.03)
	TPR	0.87	**0.99**	0.77	**0.99**		0.92	
	FPR	0.50	0.55	**0.03**	0.08		**0.03**	
	F_1	0.53	0.60	0.83	0.90		**0.91**	

In Table 5.3, consistent with the findings of Study 5.3, the BLASSOQR method consistently demonstrated superior predictive performance compared to other methods across the three quantile models. On average, both BGLSSQR and BSGSSQR methods produced closely alighted prediction error values, or MAE. In the model selection accuracy, BGLSSQR remained the best among all Bayesian methods in terms of TPR, while BSGSSQR showcased a lower FPR and higher F_1 score. In contrast, both LASSOQR and GLASSOQR struggled to manage their FPR and F_1 effectively.

Summary

As the focus shifted towards more complex scenarios characterised by higher sparsity, the BLASSOQR method maintained its superior predictive performance. The proposed methods, BGLSSQR and BSGSSQR, closely followed this respect. With regard to model selection accuracy, there was a slight but not considerable decline compared to Study 5.3. BSGSSQR exhibited commendable performance across all three quantile models with acceptable three metrics.

5.11 Simulation study for simultaneous selection

Subsequently, our attention shifted towards evaluating the effectiveness of simultaneous selection using our proposed method (BSGSSMQR) in Section 5.8 through a comprehensive simulation study. This study simulated longitudinal data similar to that presented in Chapter 4, with independent variables containing grouped structures. I thoroughly investigate the method's performance in prediction and subset selection, with specific emphasis on both fixed and random effects. Additionally, I compared this proposed method with AQMM, as outlined in Chapter 4.

5.11.1 Data generation

The simulated data were generated by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + u_{i1}z_{ij1} + u_{i2}z_{ij2} + u_{i3}z_{ij3} + u_{i4}z_{ij4} + \epsilon_{ij}, \quad i = 1, \ldots, 50, j = 1, \ldots, 10,$$

where the fixed effect component follows the specifications outlined in Study 5.4. The residuals ϵ_{ij} were generated from $N(0, \sigma_\epsilon^2)$ with σ_ϵ^2 set to two scenarios: $\sigma_\epsilon^2 = 1$ and $\sigma_\epsilon^2 = 9$. Specifically, z_{ij1} took the form of a column vector of ones with dimension $\mathbf{1}_n'$. Additionally, $z_{ijl'}$, with l' ranging from 2 to 4, were generated from a uniform distribution $\mathcal{U}(-2, 2)$. I set $\boldsymbol{u} = (u_{i1}, u_{i2}, u_{i3}, u_{i4})' \sim N(\mathbf{0}, \boldsymbol{\Sigma_u})$, where

$$\boldsymbol{\Sigma_u} = \begin{pmatrix} 0.90 & 0.40 & 0.06 & 0.00 \\ 0.40 & 0.50 & 0.10 & 0.00 \\ 0.06 & 0.10 & 0.20 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \end{pmatrix}.$$

5.11.2 Fitting the simulated data

In this study, the quantile models with $\tau \in (0.10, 0.50, 0.90)$ were taken into account during the simulation process. For the Beta priors for π_0 and π_1, and inverse Gamma priors for σ, I followed the same prior specifications as mentioned in Section 5.10. The number of updates and iterations for estimating the hyperparameter t of s^2 through a Monte Carlo EM algorithm was designated as 100. Furthermore, I fixed $p_{l'0}$ of $d_{l'}$ to 0.5, while the prior mean and variance of $\boldsymbol{a}$ were set as $\mathbf{0}$ and $0.5\mathbf{I}$. I employed a Gibbs sampler for a total of 25,000 iterations, with an initial burn-in period of 5,000 iterations. Following the simulations, I summarised each performance using numerical metrics, considering a total of 200 simulations. Similar to Simulation studies in Section 5.10, thinning is not taken into consideration in this study.

Regarding AQMM, it should be noted that the `aqmm` function from the R package `aqmm` allows only for the fitting of non-linear terms through the `smooth.terms` specification in the `mgcv` package. Consequently, variables associated with a third-degree polynomial, such as X_1, X_5, and X_7, were modelled as cubic P-splines on 40 equidistant knots distributed across the rage of each variable. Other variables were assumed to be linear predictors. To select the variables or predictors, the p-values for a two-tailed t-test from the bootstrap method with 50 replications were employed, on the understanding that if this p-value is less than the significance level of 0.05, the variable will be included in the model.

5.11.3 Summarising the results

To evaluate the predictive performance of the proposed approach, I employed the MAE metric as prediction error, as outlined in Section 4.4. Additionally, posterior median was used as the thresholding estimator in subset selection. Subsequently, for assessing the performance of subset selection, I utilised several metrics. These include the true positive rate (TPR), the false positive rate (FPR) and the F_1 score, as detailed in Table 5.1. To ensure fairness in comparing BSGSSMQR and AQMM, I excluded predictors from the group variables X_1, X_5 and X_7 when measuring subset selection performance. This decision was made because the bootstrap testing in AQMM is limited to linear predictors and does not support any smooth terms.

5.11.4 Results

Table 5.4 summarises the mean MAE for two error variance (σ_ϵ^2) scenarios, using two proposed methods (BSGSSMQR with posterior mean and BSGSSMQR with posterior median) and compared to AQMM. Generally, the results indicate that the prediction errors from both the posterior mean and posterior median estimators were relatively similar across three quantile models and two error variances. When the noise level was high, the prediction errors exhibited higher values. In comparison with AQMM, both methods yielded slightly higher MAE across the three quantiles and in both error variance scenarios.

In Table 5.5, I summarised the model selection accuracy of two approaches in two error variance scenarios. Compared to AQMM, BSGSSMQR demonstrated superior accuracy across all three quantiles and in both error variance scenarios. When evaluating the simultaneous selection performance of BSGSSMQR, it exhibited a very low false positive rate and a high true positive rate, along with a high F_1 score, particularly in scenarios with minimal noise. However, in cases of too much noise, all three metrics showed a decline.

Table 5.4: Mean MAE for AQMM and BSGSSMQR with two posterior estimators in two error variances

τ		**AQMM**	**BSGSSMQR**$_{mean}$	**BSGSSMQR**$_{median}$
0.10	$\sigma_\epsilon^2 = 1$	0.1530 (0.0085)	0.2302 (0.0175)	0.2393 (0.0167)
	$\sigma_\epsilon^2 = 9$	0.4915 (0.0272)	0.9549 (0.0468)	0.9724 (0.0472)
0.50	$\sigma_\epsilon^2 = 1$	0.3303 (0.0163)	0.3524 (0.0160)	0.3582 (0.0151)
	$\sigma_\epsilon^2 = 9$	1.0840 (0.0559)	1.0061 (0.0416)	1.0221 (0.0418)
0.90	$\sigma_\epsilon^2 = 1$	0.1469 (0.0087)	0.2304 (0.0171)	0.2393 (0.0167)
	$\sigma_\epsilon^2 = 9$	0.4771 (0.0262)	0.9544 (0.0449)	0.9724 (0.0472)

Table 5.5: Mean true positive rate (TPR), false positive rate (FPR) and F_1 score for AQMM and BSGSSMQR in two error variances

τ		$\sigma_\epsilon^2 = 1$			$\sigma_\epsilon^2 = 9$		
		AQMM	**BSGSSMQR**$_{fixed}$	**BSGSSMQR**$_{all}$	**AQMM**	**BSGSSMQR**$_{fixed}$	**BSGSSMQR**$_{all}$
0.10	TPR	0.53	1.00	1.00	0.27	0.80	0.88
	FPR	0.14	0.02	0.03	0.15	0.09	0.15
	F_1	0.54	0.97	0.97	0.43	0.75	0.80
0.50	TPR	0.66	1.00	1.00	0.34	0.80	0.88
	FPR	0.06	0.02	0.03	0.05	0.09	0.16
	F_1	0.71	0.97	0.97	0.49	0.75	0.80
0.90	TPR	0.77	1.00	1.00	0.45	0.80	0.88
	FPR	0.07	0.02	0.03	0.11	0.09	0.15
	F_1	0.77	0.97	0.97	0.54	0.75	0.80

Note:

- The term *fixed* represents only fixed effects were considered.
- The term *all* represents both fixed and random effects were considered.

5.11.5 Summary

The posterior mean estimator and the posterior median estimator yielded similar prediction errors in both cases involving two error variances. Both estimators exhibited slightly higher prediction error compared to AQMM in the median model, and were relatively higher in both extreme quantiles. Regarding the simultaneous selection performance, the posterior median estimator remained a robust choice, effectively serving as a threshold for selecting or deselecting predictors. As anticipated, BSGSSMQR exhibited a drop in its performance when the data contained excessive noise. In contrast, AQMM appeared to underperform across all accuracy matrices in this respect.

5.12 Sensitivity analysis

In this section, I aimed to investigate the behaviour of the prior parameter specifications concerning the selection of fixed and random effects. To achieve this, I employed the simulation setting described in Section 5.11, where the errors follow standard normal distribution. I varied hyperpameters of π_0, π_1, and $d_{l'}$ into nine distinct scenarios.

Table 5.6: Sensitivity analysis for BSGSSMQR

τ	$p_{l'0}$	**Hyperprior**								
		$a_1, a_2, c_1, c_2 = 0.50$			$a_1, a_2, c_1, c_2 = 1.00$			$a_1, a_2, c_1, c_2 = 1.50$		
		TPR	FPR	F_1	TPR	FPR	F_1	TPR	FPR	F_1
0.10	0.30	0.96	0.04	0.94	0.96	0.04	0.94	0.96	0.04	0.94
	0.50	0.96	0.03	0.95	0.96	0.03	0.95	0.96	0.02	0.95
	0.80	0.96	0.02	0.96	0.96	0.02	0.96	0.96	0.02	0.96
0.50	0.30	0.97	0.03	0.95	0.97	0.03	0.95	0.97	0.03	0.95
	0.50	0.97	0.03	0.96	0.98	0.03	0.96	0.97	0.02	0.96
	0.80	0.97	0.02	0.97	0.97	0.02	0.97	0.97	0.02	0.97
0.90	0.30	0.97	0.04	0.95	0.97	0.04	0.95	0.97	0.04	0.95
	0.50	0.97	0.03	0.96	0.97	0.03	0.96	0.97	0.02	0.96
	0.80	0.97	0.02	0.96	0.97	0.02	0.97	0.97	0.02	0.97

Table 5.6 summarises the mean of TPR, FPR and F_1 for nine distinct scenarios of hyperpameters. The results indicate that the BSGSSMQR method was very stable in all three metrics for most of the nine variables across three quantile models.

5.13 Illustrative analysis

I employed two simulated datasets, corresponding to two data variations with $\sigma_\epsilon^2 = 1$ and $\sigma_\epsilon^2 = 9$, respectively, obtained from Section 5.11, to conduct an illustrative analysis using our proposed methodology. Three quantile models at $\tau = 0.10, 0.50$ and 0.90 were fitted, adhering to the specifications outlined in Section 5.11.

In Table 5.7, the results indicate that BSGSSMQR yielded zero posterior median estimates for the fixed effects and accurately identified the three active groups of fixed effects in both cases. However, in the instance of high noise ($\sigma_\epsilon^2 = 9$), BSGSSMQR tended to shrink some

Table 5.7: Posterior mean, standard deviation (SD), 95% credible intervals (CrI), posterior median for both the fixed (FEs) and random (REs) effects, and computational times (in minutes) of the **0.10th quantile** model under **BSGSSMQR** with two different error variances

				$\sigma_\epsilon^2 = 1$				$\sigma_\epsilon^2 = 9$	
	True	**Mean**	**SD**	**CrI**	**Median**	**Mean**	**SD**	**CrI**	**Median**
FEs									
β_1	0.67	0.827	0.297	0.194, 1.372	**0.837**	0.449	0.498	-0.280, 1.526	**0.390**
β_2	-1.00	-0.997	0.099	-1.193, -0.805	**-0.996**	-1.082	0.248	-1.571, -0.591	**-1.081**
β_3	0.33	0.283	0.195	0.000, 0.656	**0.296**	0.159	0.290	-0.339, 0.822	0
β_4	0.50	0.488	0.087	0.319, 0.663	**0.487**	0.612	0.244	0.000, 1.074	**0.620**
β_5	0.30	0.361	0.083	0.195, 0.521	**0.362**	0.288	0.224	0.000, 0.730	**0.292**
β_6	0	-0.004	0.039	-0.112, 0.081	0	-0.103	0.167	-0.516, 0.069	0
β_7	0	-0.007	0.040	-0.125, 0.069	0	-0.117	0.173	-0.539, 0.032	0
β_8	0	-0.046	0.071	-0.218, 0.000	0	-0.074	0.147	-0.459, 0.090	0
β_9	0	0.003	0.039	-0.089, 0.114	0	0.035	0.134	-0.216, 0.403	0
β_{10}	2.00	1.873	0.198	1.512, 2.292	**1.864**	2.359	0.411	1.466, 3.083	**2.387**
β_{11}	1.00	0.939	0.117	0.719, 1.179	**0.935**	1.026	0.303	0.390, 1.579	**1.038**
β_{12}	0	0.050	0.105	-0.052, 0.350	0	-0.089	0.209	-0.616, 0.244	0
β_{13}	0	0.017	0.052	-0.051, 0.173	0	0.149	0.206	-0.053, 0.639	0
β_{14}	0	-0.053	0.100	-0.334, 0.021	0	-0.067	0.190	-0.541, 0.283	0
β_{15}	0	0.000	0.050	-0.124, 0.127	0	-0.010	0.120	-0.316, 0.259	0
β_{16}	0	0.040	0.068	0.000, 0.217	0	0.032	0.126	-0.202, 0.384	0
β_{17}	0	-0.040	0.088	-0.290, 0.045	0	-0.129	0.228	-0.718, 0.130	0
β_{18}	0	-0.002	0.043	-0.120, 0.103	0	0.066	0.156	-0.150, 0.489	0
β_{19}	0	-0.016	0.063	-0.183, 0.104	0	0.056	0.181	-0.249, 0.553	0
β_{20}	0	0.022	0.062	-0.051, 0.206	0	-0.039	0.149	-0.440, 0.245	0
β_{21}	0	-0.022	0.057	-0.183, 0.039	0	-0.160	0.203	-0.624, 0.009	0
β_{22}	0	0.008	0.041	-0.058, 0.137	0	-0.051	0.133	-0.414, 0.136	0
β_{23}	0	-0.010	0.039	-0.133, 0.041	0	0.002	0.106	-0.263, 0.276	0
β_{24}	0	-0.004	0.037	-0.114, 0.073	0	0.031	0.123	-0.203, 0.378	0
β_{25}	0	-0.014	0.044	-0.146, 0.040	0	-0.042	0.127	-0.391, 0.168	0
β_{26}	0	0.039	0.064	0.000, 0.199	0	0.052	0.131	-0.138, 0.414	0
β_{27}	0	0.020	0.048	-0.021, 0.160	0	-0.108	0.159	-0.491, 0.022	0
REs									
σ_{11}	0.90	0.866	0.213	0.529, 1.358	**0.839**	1.129	0.346	0.574, 1.924	**1.087**
σ_{22}	0.50	0.470	0.110	0.293, 0.721	**0.457**	0.533	0.237	0.052, 1.045	**0.520**
σ_{33}	0.20	0.116	0.048	0.035, 0.224	**0.112**	0.441	0.231	0.074, 0.992	**0.410**
σ_{44}	0	0.002	0.006	0.000, 0.019	0	0.187	0.195	0.000, 0.709	0.130
Time				124				114	

coefficients of small magnitude towards zero, particularly within the first group of predictors (β_3). Regarding the random effects, BSGSSMQR produced a zero posterior median estimate for the last random effect and correctly pinpointed the three most pivotal random effects only in the scenario with $\sigma_\epsilon^2 = 1$. Conversely, this method failed to produce a zero posterior median estimate for the last random effect when the simulated data contained excessive noise ($\sigma_\epsilon^2 = 9$).

Analogous to the 0.10th quantile model, the results presented in Table 5.8 for the 0.50th

Table 5.8: Posterior mean, standard deviation (SD), 95% credible intervals (CrI), posterior median for both the fixed (FEs) and random (REs) effects, and computational times (in minutes) of the **0.50th quantile** model under **BSGSSMQR** with two different error variances

		$\sigma_\epsilon^2 = 1$				$\sigma_\epsilon^2 = 9$			
	True	**Mean**	**SD**	**CrI**	**Median**	**Mean**	**SD**	**CrI**	**Median**
FEs									
β_1	0.67	0.754	0.275	0.194, 1.285	**0.751**	0.443	0.502	-0.283, 1.519	**0.367**
β_2	-1.00	-1.000	0.092	-1.182, -0.818	**-1.000**	-1.079	0.252	-1.574, -0.589	**-1.079**
β_3	0.33	0.323	0.183	0.000, 0.658	**0.338**	0.160	0.292	-0.333, 0.823	0
β_4	0.50	0.484	0.081	0.327, 0.644	**0.484**	0.613	0.240	0.012, 1.070	**0.618**
β_5	0.30	0.359	0.079	0.202, 0.511	**0.360**	0.285	0.227	0.000, 0.731	**0.289**
β_6	0	-0.005	0.037	-0.110, 0.073	0	-0.103	0.166	-0.523, 0.064	0
β_7	0	-0.005	0.037	-0.115, 0.068	0	-0.111	0.171	-0.537, 0.048	0
β_8	0	-0.036	0.063	-0.202, 0.000	0	-0.075	0.149	-0.468, 0.094	0
β_9	0	0.002	0.038	-0.087, 0.108	0	0.034	0.135	-0.222, 0.405	0
β_{10}	2.00	1.930	0.183	1.578, 2.306	**1.933**	2.370	0.419	1.467, 3.111	**2.400**
β_{11}	1.00	0.971	0.111	0.757, 1.194	**0.971**	1.038	0.303	0.403, 1.589	**1.052**
β_{12}	0	0.058	0.114	-0.028, 0.376	0	-0.091	0.206	-0.612, 0.243	0
β_{13}	0	0.017	0.050	-0.041, 0.170	0	0.150	0.208	-0.054, 0.642	0
β_{14}	0	-0.059	0.106	-0.344, 0.007	0	-0.070	0.192	-0.549, 0.283	0
β_{15}	0	-0.002	0.052	-0.142, 0.123	0	-0.009	0.115	-0.317, 0.257	0
β_{16}	0	0.033	0.063	-0.003, 0.205	0	0.034	0.128	-0.206, 0.387	0
β_{17}	0	-0.043	0.093	-0.310, 0.034	0	-0.124	0.224	-0.708, 0.141	0
β_{18}	0	-0.002	0.039	-0.109, 0.094	0	0.068	0.156	-0.146, 0.486	0
β_{19}	0	-0.014	0.061	-0.178, 0.100	0	0.054	0.179	-0.255, 0.553	0
β_{20}	0	0.025	0.068	-0.048, 0.233	0	-0.043	0.150	-0.456, 0.221	0
β_{21}	0	-0.021	0.055	-0.179, 0.036	0	-0.160	0.204	-0.631, 0.012	0
β_{22}	0	0.008	0.039	-0.051, 0.131	0	-0.049	0.130	-0.410, 0.133	0
β_{23}	0	-0.009	0.037	-0.128, 0.026	0	0.001	0.106	-0.271, 0.274	0
β_{24}	0	-0.008	0.037	-0.125, 0.038	0	0.028	0.123	-0.220, 0.367	0
β_{25}	0	-0.014	0.041	-0.141, 0.031	0	-0.041	0.127	-0.389, 0.186	0
β_{26}	0	0.037	0.063	0.000, 0.196	0	0.050	0.130	-0.131, 0.409	0
β_{27}	0	0.022	0.049	-0.012, 0.163	0	-0.106	0.159	-0.491, 0.025	0
REs									
σ_{11}	0.90	0.797	0.198	0.486, 1.253	**0.772**	1.095	0.349	0.525, 1.888	**1.057**
σ_{22}	0.50	0.473	0.109	0.297, 0.721	**0.460**	0.537	0.249	0.048, 1.066	**0.521**
σ_{33}	0.20	0.107	0.047	0.027, 0.213	**0.102**	0.474	0.259	0.094, 1.130	**0.431**
σ_{44}	0	0.002	0.006	0.000, 0.021	0	0.206	0.210	0.000, 0.759	0.143
Time				132				128	

quantile model indicate a consistent trend. BSGSSMQR was able to identify the correct active groups of fixed effects in both error variance cases. In instances of high noise, this method could not produce a zero posterior median estimate for the inactive random effect. For the 0.90th quantile model, Table 5.9 demonstrates a trend analogous to the two previous quantile models. Each active group and predictor of fixed effects was correctly identified. However, in instances of high noise, the last predictor in the first active group tended to have a posterior median estimate close to zero. Notably, BSGSSMQR, regarding the posterior median estimate, still incorrectly identified the last (inactive) random effect

Table 5.9: Posterior mean, standard deviation (SD), 95% credible intervals (CrI), posterior median for both the fixed (FEs) and random (REs) effects, and computational times (in minutes) of the **0.90th quantile** model under **BSGSSMQR** with two different error variances

		$\sigma_\epsilon^2 = 1$				$\sigma_\epsilon^2 = 9$			
	True	**Mean**	**SD**	**CrI**	**Median**	**Mean**	**SD**	**CrI**	**Median**
FEs									
β_1	0.67	0.745	0.293	0.118, 1.312	**0.744**	0.432	0.508	-0.326, 1.536	**0.345**
β_2	-1.00	-0.991	0.098	-1.184, -0.797	**-0.992**	-1.079	0.251	-1.574, -0.593	**-1.079**
β_3	0.33	0.327	0.196	0.000, 0.690	**0.342**	0.170	0.297	-0.329, 0.838	**0.014**
β_4	0.50	0.478	0.083	0.316, 0.644	**0.477**	0.614	0.238	0.036, 1.067	**0.621**
β_5	0.30	0.365	0.082	0.203, 0.524	**0.365**	0.276	0.225	0.000, 0.728	**0.279**
β_6	0	-0.005	0.039	-0.117, 0.082	0	-0.104	0.166	-0.517, 0.067	0
β_7	0	-0.006	0.039	-0.125, 0.070	0	-0.111	0.168	-0.518, 0.038	0
β_8	0	-0.031	0.060	-0.196, 0.000	0	-0.076	0.148	-0.464, 0.092	0
β_9	0	0.002	0.039	-0.098, 0.110	0	0.034	0.136	-0.234, 0.407	0
β_{10}	2.00	1.931	0.192	1.573, 2.336	**1.930**	2.380	0.418	1.491, 3.098	**2.415**
β_{11}	1.00	0.977	0.115	0.755, 1.212	**0.976**	1.040	0.302	0.414, 1.586	**1.055**
β_{12}	0	0.055	0.111	-0.032, 0.368	0	-0.095	0.214	-0.632, 0.250	0
β_{13}	0	0.017	0.052	-0.040, 0.177	0	0.151	0.208	-0.053, 0.643	0
β_{14}	0	-0.053	0.102	-0.343, 0.017	0	-0.068	0.196	-0.543, 0.307	0
β_{15}	0	-0.001	0.050	-0.134, 0.120	0	-0.007	0.121	-0.316, 0.271	0
β_{16}	0	0.031	0.061	-0.002, 0.200	0	0.036	0.128	-0.191, 0.392	0
β_{17}	0	-0.039	0.093	-0.308, 0.046	0	-0.129	0.225	-0.713, 0.126	0
β_{18}	0	-0.002	0.042	-0.116, 0.104	0	0.070	0.158	-0.149, 0.491	0
β_{19}	0	-0.016	0.066	-0.187, 0.114	0	0.051	0.178	-0.263, 0.538	0
β_{20}	0	0.024	0.067	-0.043, 0.225	0	-0.038	0.145	-0.431, 0.232	0
β_{21}	0	-0.019	0.054	-0.177, 0.042	0	-0.156	0.202	-0.619, 0.009	0
β_{22}	0	0.006	0.040	-0.071, 0.127	0	-0.052	0.133	-0.422, 0.131	0
β_{23}	0	-0.011	0.039	-0.138, 0.028	0	0.000	0.106	-0.275, 0.269	0
β_{24}	0	-0.009	0.040	-0.137, 0.046	0	0.030	0.120	-0.190, 0.366	0
β_{25}	0	-0.018	0.047	-0.156, 0.027	0	-0.041	0.128	-0.393, 0.175	0
β_{26}	0	0.038	0.063	0.000, 0.199	0	0.054	0.133	-0.139, 0.415	0
β_{27}	0	0.028	0.056	-0.006, 0.184	0	-0.106	0.158	-0.490, 0.032	0
REs									
σ_{11}	0.90	0.883	0.212	0.551, 1.386	**0.855**	1.103	0.347	0.543, 1.894	**1.061**
σ_{22}	0.50	0.480	0.110	0.302, 0.728	**0.468**	0.553	0.245	0.072, 1.089	**0.538**
σ_{33}	0.20	0.121	0.050	0.036, 0.232	**0.117**	0.465	0.254	0.074, 1.102	**0.426**
σ_{44}	0	0.003	0.007	0.000, 0.021	0	0.196	0.199	0.000, 0.721	0.137
Time				133				120	

in the presence of high noise in the simulated data.

Additionally, I fitted the same simulated datasets as outlined above using the AQMM approach for comparison with the proposed method. The variance-covariance matrix of the random effects was specified as a general positive-definite matrix with no additional structure, using `pdSymm` in the `covariance` argument in the `aqmm` function. This implies allowing for correlation among the random effects. However, in our results, I only present the standard deviations of random effects. The standard errors (SE) for each fixed effect were calculated using the bootstrap method with 50 replications.

Table 5.10: Estimates, standard error (SE) for both the fixed (FEs) and random (REs) effects of the **0.10th quantile** model under **AQMM** with two different error variances

		$\sigma_\epsilon^2 = 1$				$\sigma_\epsilon^2 = 9$			
	True	**Estimate**	**SE**	**t-value**	**Pr**$(>\lvert t\rvert)$	**Estimate**	**SE**	**t-value**	**Pr**$(>\lvert t\rvert)$
FEs									
β_4	0.50	0.571	0.095	6.024	0.000	0.524	0.344	1.523	0.134
β_5	0.30	0.343	0.080	4.274	0.000	0.321	0.231	1.391	0.171
β_6	0	-0.081	0.081	-0.998	0.323	*-0.165*	0.208	-0.793	0.432
β_7	0	*-0.120*	0.089	-1.351	0.183	*-0.257*	0.282	-0.914	0.365
β_8	0	-0.059	0.070	-0.847	0.401	*-0.292*	0.241	-1.210	0.232
β_9	0	0.088	0.087	1.017	0.314	*0.323*	0.262	1.232	0.224
β_{10}	2.00	1.647	0.463	3.554	0.001	-1.148	1.293	-0.888	0.379
β_{11}	1.00	0.775	0.412	1.880	0.066	-0.132	0.884	-0.149	0.882
β_{15}	0	*-1.271*	0.513	-2.478	0.017	*-1.537*	1.348	-1.140	0.260
β_{16}	0	*-0.671*	0.325	-2.068	0.044	*-0.785*	0.891	-0.881	0.383
β_{20}	0	*-1.589*	0.494	-3.218	0.002	*-3.362*	1.187	-2.834	0.007
β_{21}	0	*-0.934*	0.369	-2.528	0.015	*-1.793*	0.982	-1.827	0.074
β_{22}	0	0.005	0.098	0.051	0.959	*-0.401*	0.306	-1.310	0.196
β_{23}	0	0.026	0.098	0.271	0.787	*0.199*	0.265	0.751	0.456
β_{24}	0	-0.048	0.086	-0.553	0.583	*-0.266*	0.276	-0.965	0.339
β_{25}	0	0.027	0.081	0.330	0.743	*0.198*	0.246	0.806	0.424
β_{26}	0	0.088	0.076	1.152	0.255	*0.197*	0.212	0.928	0.358
β_{27}	0	*-0.120*	0.083	-1.451	0.153	*-0.411*	0.259	-1.588	0.119
$S(X_1)$	-	36.634	7.131	5.138	0.000	43.170	13.507	3.196	0.002
$S(X_5)$	-	3.041	2.002	1.519	0.135	2.304	5.316	0.434	0.667
$S(X_7)$	-	6.533	1.973	3.311	0.002	10.738	6.704	1.602	0.116
REs									
σ_{11}	0.90	0.537	-	-	-	0.759	-	-	-
σ_{22}	0.50	0.535	-	-	-	0.304	-	-	-
σ_{33}	0.20	0.204	-	-	-	0.172	-	-	-
σ_{44}	0	0	-	-	-	0.003	-	-	-

In Table 5.10, the results show that AQMM yielded some estimates not close to zero (highlighted in red italics) for the inactive coefficients. This pattern was observed across two different error variances, particularly in the case where $\sigma_\epsilon^2 = 9$. Notably, several inactive coefficients were selected (highlighted in red) for the 0.10th quantile model, based on the p-values from the bootstrap tests. Additionally, some active coefficients were not selected (highlighted in blue), following the same criterion. Regarding the random effects, the estimated standard deviations were relatively close to the true values in two scenarios of error variance.

In the 0.50th quantile model, Table 5.11 shows that AQMM yielded estimates comparable to those of the 0.10th quantile model, though the discrepancies were less severe. In terms of the variable selection, one active coefficient and one inactive coefficient were incorrectly selected, as indicated by the p-values from the bootstrap tests, in the scenario where

$\sigma^2_\epsilon = 1$. Meanwhile, in the case where $\sigma^2_\epsilon = 9$, three active coefficients were not selected. For the random effects, AQMM estimated standard deviations closely aligned with the true values, particularly for the last three random effects.

Table 5.11: Estimates, standard error (SE) for both the fixed (FEs) and random (REs) effects of the **0.50th quantile** model under **AQMM** with two different error variances

		$\sigma^2_\epsilon = 1$				$\sigma^2_\epsilon = 9$			
	True	**Estimate**	**SE**	**t-value**	**Pr**$(>\lvert t\rvert)$	**Estimate**	**SE**	**t-value**	**Pr**$(>\lvert t\rvert)$
FEs									
β_4	0.50	0.536	0.075	7.126	0.000	0.752	0.228	3.292	0.002
β_5	0.30	0.286	0.064	4.458	0.000	0.237	0.231	1.029	0.308
β_6	0	-0.075	0.063	-1.205	0.234	*-0.295*	0.207	-1.426	0.160
β_7	0	-0.022	0.060	-0.374	0.710	-0.017	0.174	-0.100	0.921
β_8	0	-0.017	0.077	-0.220	0.827	-0.019	0.199	-0.095	0.925
β_9	0	-0.003	0.067	-0.046	0.963	0.028	0.217	0.131	0.896
β_{10}	2.00	1.809	0.569	3.180	0.003	2.053	1.275	1.611	0.114
β_{11}	1.00	0.676	0.461	1.464	0.149	0.962	1.002	0.961	0.342
β_{15}	0	*0.310*	0.540	0.574	0.569	*-0.165*	1.222	-0.135	0.893
β_{16}	0	*0.221*	0.292	0.756	0.453	-0.021	0.634	-0.034	0.973
β_{20}	0	*-0.808*	0.427	-1.891	0.065	*-0.422*	1.068	-0.396	0.694
β_{21}	0	*-0.807*	0.347	-2.326	0.024	*-1.025*	0.725	-1.413	0.164
β_{22}	0	-0.032	0.080	-0.403	0.689	0.003	0.224	0.014	0.989
β_{23}	0	0.018	0.063	0.291	0.772	0.086	0.204	0.423	0.674
β_{24}	0	-0.029	0.075	-0.388	0.700	*0.172*	0.215	0.801	0.427
β_{25}	0	-0.057	0.079	-0.721	0.475	*-0.258*	0.194	-1.332	0.189
β_{26}	0	0.058	0.072	0.817	0.418	0.092	0.218	0.420	0.677
β_{27}	0	*-0.126*	0.066	-1.926	0.060	*-0.313*	0.218	-1.434	0.158
$S(X_1)$	-	34.537	13.693	2.522	0.015	27.858	12.617	2.208	0.032
$S(X_5)$	-	-1.690	1.607	-1.052	0.298	-0.325	4.393	-0.074	0.941
$S(X_7)$	-	2.833	1.890	1.499	0.140	-1.067	5.236	-0.204	0.839
REs									
σ_{11}	0.90	0.659	-	-	-	1.163	-	-	-
σ_{22}	0.50	0.465	-	-	-	0.486	-	-	-
σ_{33}	0.20	0.208	-	-	-	0.266	-	-	-
σ_{44}	0	0.000	-	-	-	0.003	-	-	-

Analogous to the two previous quantiles models, in the 0.90th quantile model, some estimates for inactive coefficients were not close to zero, as presented in Table 5.12. The AQMM approach appeared to perform well in this quantile model. Only one inactive coefficient was selected as an (active) significant predictor in the case where $\sigma^2_\epsilon = 1$, while one active coefficient was selected as an (inactive) non-significant predictor in the case where $\sigma^2_\epsilon = 9$, based on the p-values from the bootstrap tests. Concerning the random effects, it was observed that the standard deviations estimated by AQMM were relatively smaller than the true values.

Table 5.12: Estimates, standard error (SE) for both the fixed (FEs) and random (REs) effects of the **0.90th quantile** model under **AQMM** with two different error variances

		$\sigma_\epsilon^2 = 1$				$\sigma_\epsilon^2 = 9$			
	True	**Estimate**	**SE**	**t-value**	**Pr**$(>\lvert t\rvert)$	**Estimate**	**SE**	**t-value**	**Pr**$(>\lvert t\rvert)$
FEs									
β_4	0.50	0.494	0.074	6.715	0.000	0.509	0.249	2.046	0.046
β_5	0.30	0.288	0.074	3.893	0.000	0.386	0.196	1.967	0.055
β_6	0	0.012	0.075	0.158	0.875	*-0.174*	0.249	-0.699	0.488
β_7	0	-0.016	0.068	-0.242	0.810	-0.036	0.222	-0.162	0.872
β_8	0	-0.032	0.069	-0.465	0.644	*-0.169*	0.203	-0.835	0.408
β_9	0	-0.030	0.074	-0.402	0.689	-0.007	0.241	-0.030	0.976
β_{10}	2.00	2.416	0.547	4.415	0.000	3.153	1.214	2.598	0.012
β_{11}	1.00	1.075	0.366	2.937	0.005	2.287	0.881	2.597	0.012
β_{15}	0	*0.351*	0.636	0.552	0.584	*2.029*	1.295	1.567	0.124
β_{16}	0	*0.218*	0.343	0.634	0.529	*1.340*	0.754	1.777	0.082
β_{20}	0	*0.194*	0.501	0.387	0.700	*1.335*	0.914	1.461	0.150
β_{21}	0	0.076	0.278	0.275	0.784	*0.589*	0.652	0.903	0.371
β_{22}	0	*0.148*	0.068	2.158	0.036	*0.364*	0.252	1.442	0.156
β_{23}	0	0.013	0.066	0.196	0.846	0.001	0.223	0.005	0.996
β_{24}	0	0.083	0.078	1.063	0.293	*0.301*	0.265	1.134	0.262
β_{25}	0	-0.095	0.076	-1.258	0.214	*-0.182*	0.259	-0.702	0.486
β_{26}	0	0.064	0.078	0.828	0.412	*0.153*	0.239	0.637	0.527
β_{27}	0	0.015	0.064	0.241	0.811	-0.009	0.242	-0.035	0.972
$S(X_1)$	-	32.685	10.227	3.196	0.002	21.879	7.595	2.881	0.006
$S(X_5)$	-	-2.165	2.529	-0.856	0.396	-7.094	6.879	-1.031	0.308
$S(X_7)$	-	-0.638	2.450	-0.260	0.796	-10.402	5.464	-1.904	0.063
REs									
σ_{11}	0.90	0.448	-	-	-	0.788	-	-	-
σ_{22}	0.50	0.336	-	-	-	0.308	-	-	-
σ_{33}	0.20	0.135	-	-	-	0.174	-	-	-
σ_{44}	0	0	-	-	-	0.003	-	-	-

5.13.1 Summary

The BSGSSMQR approach excelled in fixed effect selection compared to the AQMM approach when applied to simulated datasets, particularly in scenarios where the error variance was assumed to be small. However, it showed relative sensitivity to large error variances, resulting in a slight decrease in performance in this respect and a significant impact on the selection of random effects. Notably, while the AQMM approach struggled to estimate inactive coefficients close to their true values (zero) and sometimes incorrectly selected variables, it performed relatively well in estimation of random effects. Nonetheless, AQMM lacks a specific test to quantify uncertainty in random effects.

5.14 Chapter summary

In this chapter, Bayesian variable selection methods within quantile models and quantile mixed models were developed. Initially, the proposed methodologies for the quantile models combined three key techniques: the Bayesian LASSO-type methods (Bayesian group LASSO and Bayesian sparse group LASSO), a likelihood function based on the scale mixture representation of the asymmetric Laplace (AL) distribution, and spike and slab priors for regression coefficients. The first technique was employed to select the group structures of predictors, such as a multi-level categorical predictor and a group of basis functions, to incorporate nonlinear relationships. The second was utilised as the working likelihood of data in the Bayesian framework, and the third was adopted to yield sparse estimators. Subsequently, the proposed methodology in the quantile models was extended by incorporating the use of mixed models, based on a decomposition for the covariance matrix of random effects, to enable the simultaneous selection of both fixed and random effects in the quantile mixed model framework.

In the context of quantile models, the simulation studies demonstrate that the proposed methods (BGLSSQR and BSGSSQR) generally yield biased estimators, as expected, due to the regularisation method. In addition, they generally perform well in terms of predictive performance compared to existing frequentist and Bayesian QR methods. Notably, in terms of variable selection performance, the proposed methods outperform other methods. This finding reveals that incorporating Bayesian group and sparse group LASSO with spike and slab priors on quantile regression coefficients significantly improves this aspect. Owing to the advantages of BSGSSQR over BGLSSQR, BSGSSQR is more suitable for extension to the context of quantile mixed models.

In the quantile mixed models, the proposed method (BSGSSMQR) demonstrates superior performance in simultaneous selection. Although it provides slightly lower predictive performance compared to AQMM, the results remain within acceptable values. Sensitivity analysis reveals that the proposed method stabilises in model selection accuracy when varying relevant hyperparameters. When implemented with simulated datasets, the results indicate that the proposed method performed well across three quantile models. Furthermore, it exhibits sensitivity to higher error variance, particularly in both estimation and selection of random effects. Given the advantages of the proposed method as described earlier, I will apply it to real longitudinal child growth data in Chapter 6.

Chapter 6

Application to longitudinal child growth data in Scotland

In Chapters 4 and 5, several experimental studies were conducted to evaluate the performance of AQMM and BSGSSMQR within the context of longitudinal child growth data (LCGD). The findings in Chapter 4 suggest that AQMM is sufficiently suitable for modelling this type of data, although it lacks a method for selecting appropriate random effects to capture individual-specific variations. AQMM is capable of constructing reference child growth charts and identifying risk factors (fixed effects) that affect child physical growth measurements, utilising bootstrapping. The novel approach in Chapter 5, BSGSSMQR, offers capabilities beyond AQMM, as it enables the simultaneous selection of fixed and random effects. Therefore, in this chapter, these two models are applied to LCGD in Scotland, providing comprehensive insights into the child physical growth and development of Scottish children.

6.1 Modelling LCGD in Scotland using additive quantile mixed model

In this section, the AQMM approach was applied to the real LCGD obtained from the Growing up in Scotland (GUS) study. Detailed information about the study and data, including collection procedures, variables, and exploratory analysis, is provided in Section 2.6. The analysis aims to achieve two primary objectives: constructing reference growth charts and identifying risk factors associated with child physical growth measurements. The latter specifically focuses on two critical points in the distribution of child physical growth measurements, namely the 0.10th and 0.90th quantiles. These quantiles represent children in the lowest 10% and highest 90% of the distribution, respectively, highlighting those physical growth measurements that fall below or above the typical range. In particular, the upper quantile in this context will represent children at risk of obesity or overweight, a considerable concern in Scotland.

6.1.1 Growth measurements

In this application, the primary focus was on analysing the growth measurements of interest, which included both raw and standardised growth metrics. Specifically, I considered raw measurements of weight and height, alongside weight-for-age z-scores (WAZ) and height-for-age z-scores (HAZ). These standardised measurements provide valuable information about the relative growth status of children in comparison to reference populations, as described in Section 2.3.

6.1.2 Covariates

The analysis focuses on covariates that are likely to have a potential impact and serve as risk factors associated with child growth development, adopting the framework presented in "Multiple risk factors in young children's development" by Sabates and Dex (2012), as described in Section 2.6.3. The list of potential covariates investigated in this analysis is presented in Table 2.5. Note that when identifying risk factors associated with growth measurements, birth weight is not included. With an average birth weight of 3,425.48 grams in the GUS data (3,499.82 grams for males and 3,349.91 grams for females), the observed differences are minimal and unlikely to have a clinically meaningful impact on growth outcomes. Therefore, including birth weight as a risk factor would introduce unnecessary complexity without considerably enhancing the model's explanatory power.

6.1.3 Fitting the models

According to the GUS data, the dataset includes children aged from 10 months to 14 years, encompassing both child and adolescent growth. However, physical growth measurements, such as weight, were only taken twice during the child growth phase, at 10 months and 4 years of age. Therefore, I decided to approach this in two ways: 1) modelling the entire dataset to explore the complete growth trajectory captured in the GUS dataset and 2) modelling only the dataset of children approximately aged 4 to 14 years (Sweeps 4 to 10) to focus specifically on the period including adolescent growth. The latter covers the growth of school-age children and young people in primary or secondary education in Scotland.

Quantile models for constructing the reference growth charts

To construct the reference growth charts, the quantile models with τ values of 0.004, 0.02, 0.09, 0.25, 0.50, 0.75, 0.91, 0.98 and 0.996 were fitted separately by sex with age, as there are distinct differences in growth patterns and rates between boys and girls. These quantiles correspond to the UK-WHO growth charts as presented by Royal College of Paediatrics and Child Health (2013). In this analytical context, I fitted two distinct datasets (i.e. the entire dataset and children approximately aged 4 to 14 years) using the AQMM with cubic P-splines and the squared second-order ($m = 2$) difference to model the smooth effect (non-parametric term) of age in years. Following this spline approach, the number of knots (K) was calculated as $K = \min\{40, \text{the number of unique } x/4\}$, in accordance with the guidance of Ruppert et al. (2003). The equidistant positions of these knots were utilised, as required by P-splines. Note that in the AQMM, the penalisation method will eliminate bases that are not supported by the data (i.e. those with missing age values). Initially, I established a design matrix of random effects, $\mathbf{Z} = [\mathbf{1}, X_1]$, where X_1 represents the variable of age in years. The positive-definite matrix for these random effects was specified as the general positive–definite matrix, with no additional structure (`pdSymm`). This setting serves as an instrument to account for the inherent potential heterogeneity among children by incorporating random intercepts and slopes. However, the fitted models did not capture the GUS data well, as the percentages of observations above each percentiles did not correspond to the expected quantile levels. For example, for the 25th percentile, the percentage of observations above this percentile should be close to 75% (see Tables 6.1 to 6.2 and Figures D.1 to D.6 in Appendix D). Therefore, I refitted these models by revising the design matrix of random effects to include only random intercepts, $\mathbf{Z} = [\mathbf{1}]$. As a result, the refitted models provided quantile curves that fit two GUS datasets well, particularly raw weight and height measurements.

Quantile models for identifying risk factors

In terms of identifying risk factors associated with growth measurements, only the dataset of children aged 4 - 14 years, which includes the adolescent growth period, was considered, as it constitutes the majority in the GUS data. Additionally, fitting the model using this dataset will benefit its implementation for school-age children and young people in primary or secondary education in Scotland. Therefore, the quantile models were fitted at two distinct locations—the lower (the 0.10th quantile) and upper (the 0.90th quantile) locations—without stratification by sex. The former represents the weight threshold for a small proportion of the population, while the latter represents the weight threshold for a large proportion of the population. In this analysis, the sex variable was considered as an interaction term. Within this analytical context, the standard errors for each fixed effect were calculated using the bootstrap method with 50 replications. For convenience and to facilitate comparison with BSGSSMQR in Section 6.2, the response variables (i.e. raw weight, raw height, WAZ, and HAZ) were centred around their means. Additionally, to mitigate the scaling effects of covariates, each covariate was scaled by its mean and standard deviation. Age was exclusively incorporated as a fixed smooth effect within each quantile model. This was achieved using the cubic P-spline method with squared second-order ($m = 2$) difference to establish the bases for this smooth term. The determination of the number of knots (K) adhered to the guidance of Ruppert et al. (2003) with $K = \min\{40, \text{the number of unique } x/4\}$. These knots were positioned equidistantly along the variable x, adhering to the mandatory requirement of P-splines. Initially, all other covariates were assumed to be linear (fixed effect) terms and were included in each quantile model. Subsequently, I considered removing non-significant covariates (p-values < 0.05) from all quantile models, with exception of the smooth term for age. Following this, I refitted these three quantile models to identify significant covariates capable of explaining the variability in child physical growth measurements.

6.1.4 Results - Reference growth patterns

Figure 6.1 shows the nine estimated quantile curves for raw weight measurements in the reference group, representing child weight growth patterns, as fitted using the entire GUS dataset. Each quantile curve appears to fit this dataset well (Table 6.1). The (raw) Weight-for-Age patterns between males (Figure 6.1 (a)) and females (Figure 6.1 (b)) exhibited slight differences, but the smooth quantile curves of both sexes tended to parallel each other across ages. In general, these curves demonstrate a rapid progression from ages 10 months to 4 years (early childhood), stabilised between ages 4 to 10 (middle childhood), and then exhibited a marked increase from age 10 onwards (adolescence) across the nine curves. Furthermore, Figure 6.2 shows the same nine quantile curves for raw weight measurement, but fitted using the GUS dataset for children aged 4 to 14 years, highlighting growth patterns during adolescence. Both charts (Figure 6.2 (a) and Figure 6.2 (b)) demonstrate that children generally experience a steady increase in raw weight from ages 4 to 14.

Figure 6.3 presents additional child growth patterns based on WAZ, fitted using the entire GUS dataset. Each quantile curve appears to fit the dataset much better than the initial fitted models (random intercepts and slopes) (see Table 6.1). However, the percentages of observations above each percentile were not close to the expected values. This suggests that the model specifications may still need adjustment. For example, the model structure, including the random effects, may not align with the characteristics of the WAZ data, such as changes in the within-individual error over time. While the random effects in AQMM account for differences between individuals in their baseline levels and rates of change, they do not accommodate changes in within-individual error over time. Regarding to WAZ growth patterns for males and females, both charts show that WAZ patterns reveal rapid growth from ages 1 to 4 years, a stabilisation from ages 4 to 10 compared to the

Table 6.1: Percentages of observations above each quantile curve in raw weight and WAZ for two AQMM models (random intercepts and slopes v.s. only random intercepts.)

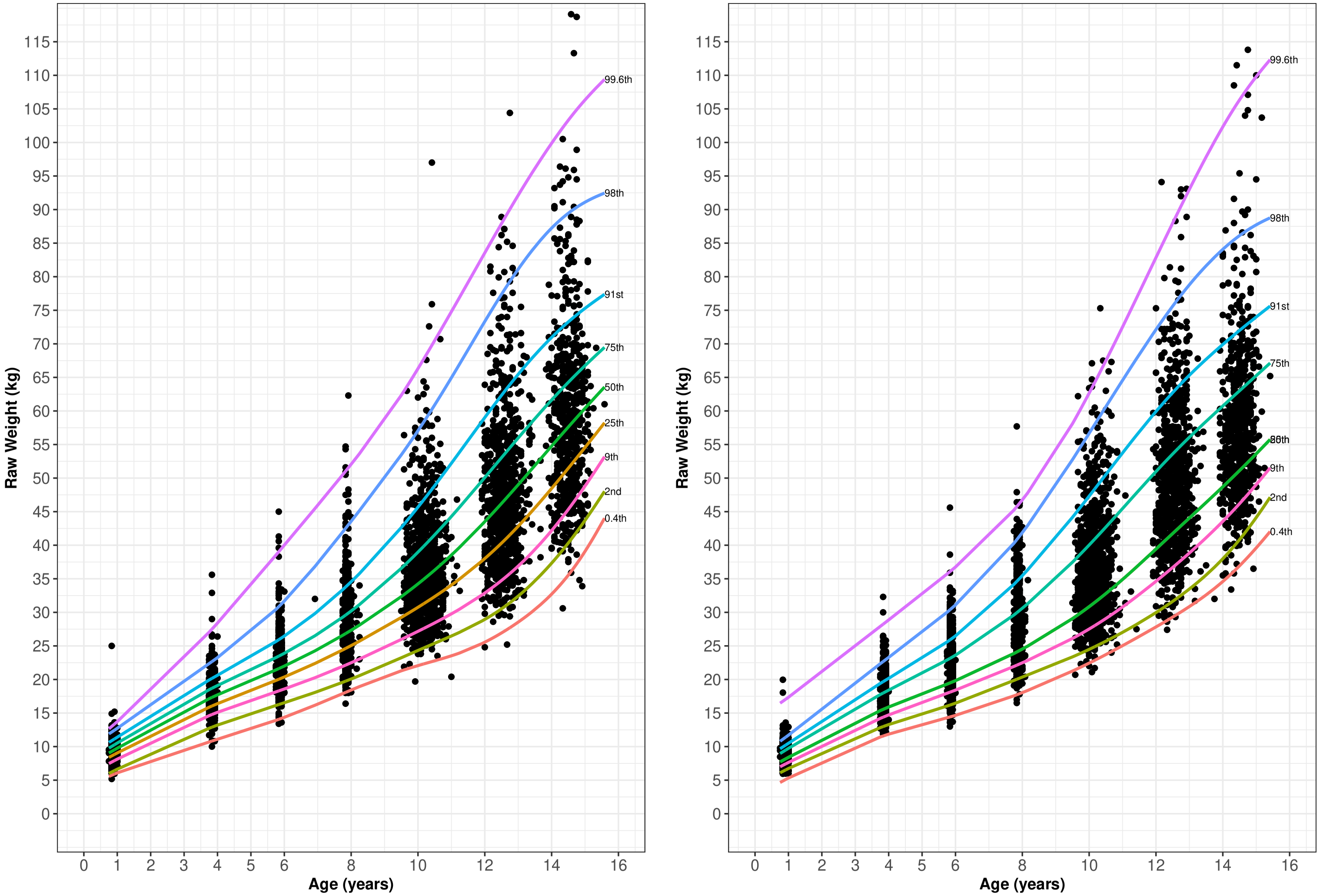
Raw weight

Percentile curves	**Male**		**Female**	
	Intercepts and Slopes	**Intercepts**	**Intercepts and Slopes**	**Intercepts**
Entire GUS dataset				
0.4th	92.82	99.75	93.25	99.71
2nd	84.63	98.68	83.45	97.92
9th	72.90	91.87	71.05	89.56
25th	59.30	73.57	57.15	72.78
50th	44.11	48.54	44.17	49.58
75th	30.84	25.23	30.71	26.14
91st	19.81	9.34	20.40	10.38
98th	11.93	1.93	12.33	1.85
99.6th	7.26	0.41	6.98	0.36
Children aged 4 - 14 years dataset				
0.4th	88.25	99.38	86.31	99.14
2nd	77.64	97.22	73.90	96.00
9th	63.69	89.22	61.33	86.60
25th	52.62	72.90	51.35	70.49
50th	43.89	48.76	42.60	48.79
75th	35.10	25.70	34.78	26.83
91st	26.81	9.21	26.64	11.03
98th	17.79	2.27	17.85	2.91
99.6th	12.59	0.49	10.49	0.52

WAZ

Percentile curves	**Male**		**Female**	
	Intercepts and Slopes	**Intercepts**	**Intercepts and Slopes**	**Intercepts**
Entire GUS dataset				
0.4th	88.11	96.88	85.53	94.73
2nd	78.87	89.98	75.91	86.29
9th	68.16	76.08	65.55	72.32
25th	59.29	63.75	56.82	60.58
50th	50.75	50.88	49.07	48.94
75th	42.45	38.25	40.91	37.27
91st	33.06	24.47	32.30	26.10
98th	22.00	11.57	22.25	14.51
99.6th	13.59	4.35	13.75	5.44
Children aged 4 - 14 years dataset				
0.4th	80.98	89.40	79.23	88.83
2nd	70.52	79.81	69.26	78.19
9th	61.79	69.51	61.52	67.26
25th	55.31	59.08	54.03	56.86
50th	49.62	49.71	48.39	48.09
75th	43.87	40.92	41.85	39.07
91st	37.63	30.78	35.23	29.67
98th	29.21	20.70	26.73	19.34
99.6th	20.35	12.17	19.96	13.40

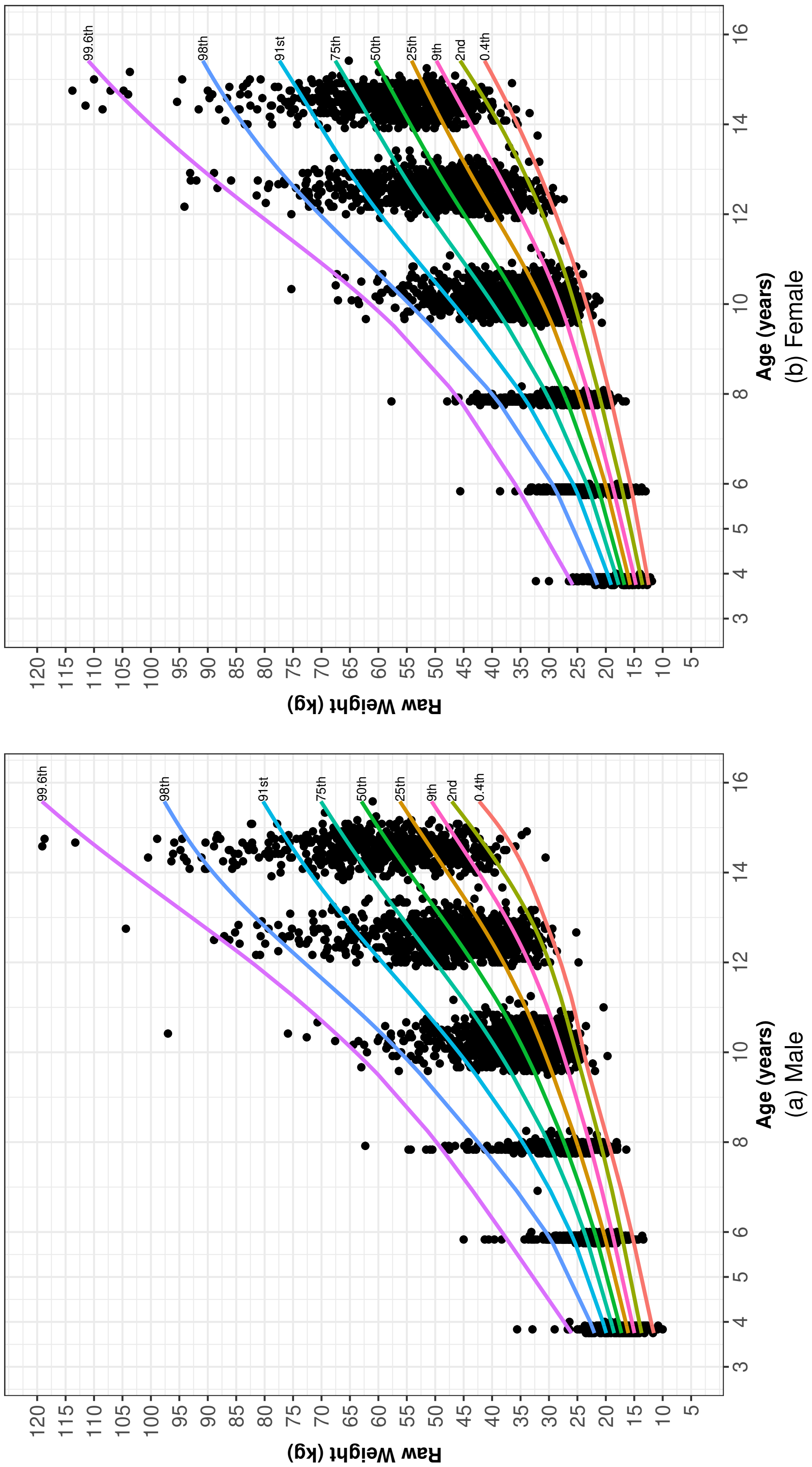
earlier years, and a gradual uptick from ages 10 to 14 years. Notably, at the extreme quantiles, such as the 0.04th and 99.6th, the curves show some fluctuations, particularly at older ages.



(a) Male

(b) Female

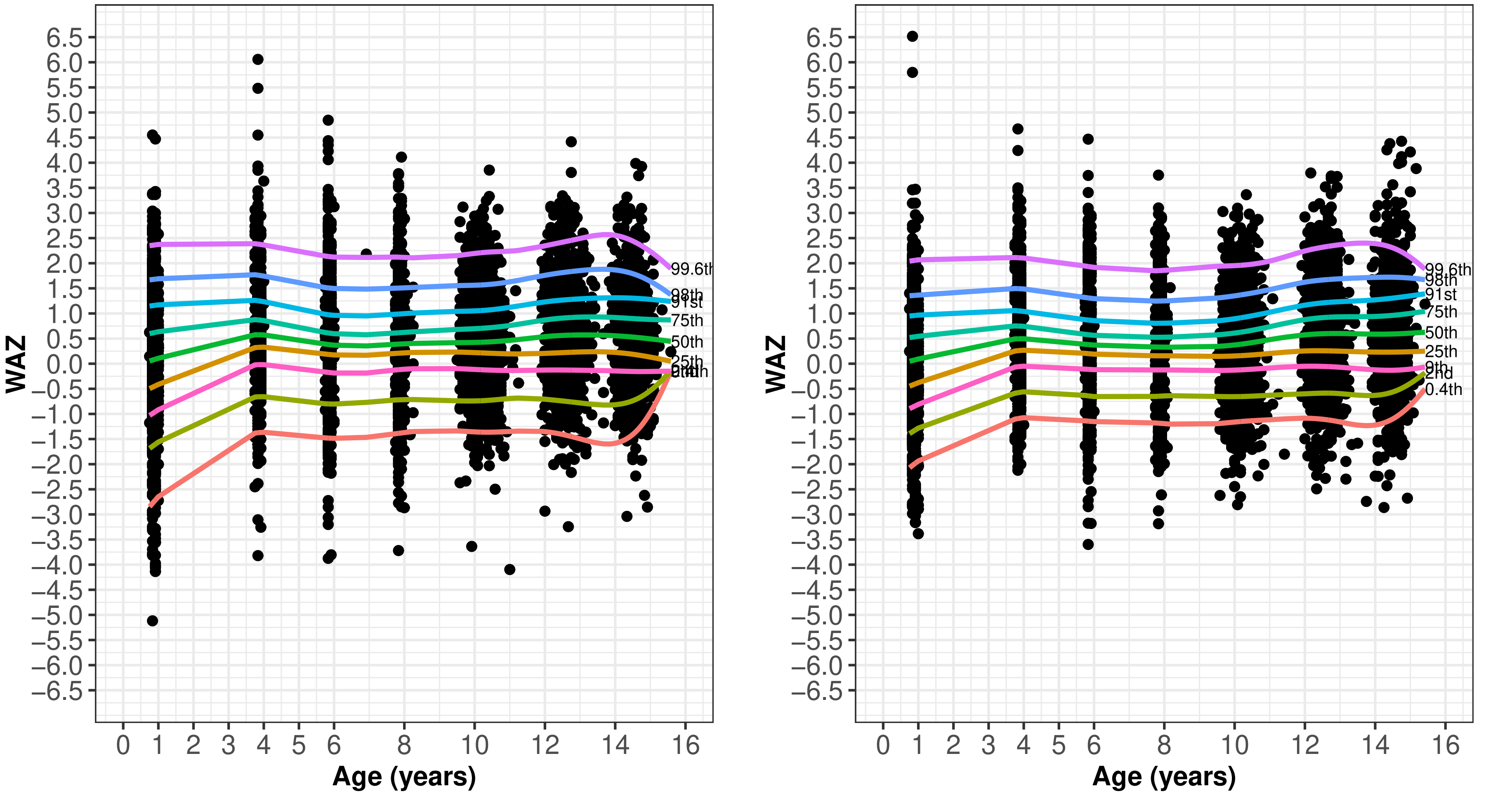
Figure 6.1: Conditional quantile curves for raw weight, estimated using the AQMM with cubic P-splines and fitted with random intercepts of age, for a child in the reference group across nine quantile levels using the entire GUS dataset.



(a) Male

(b) Female

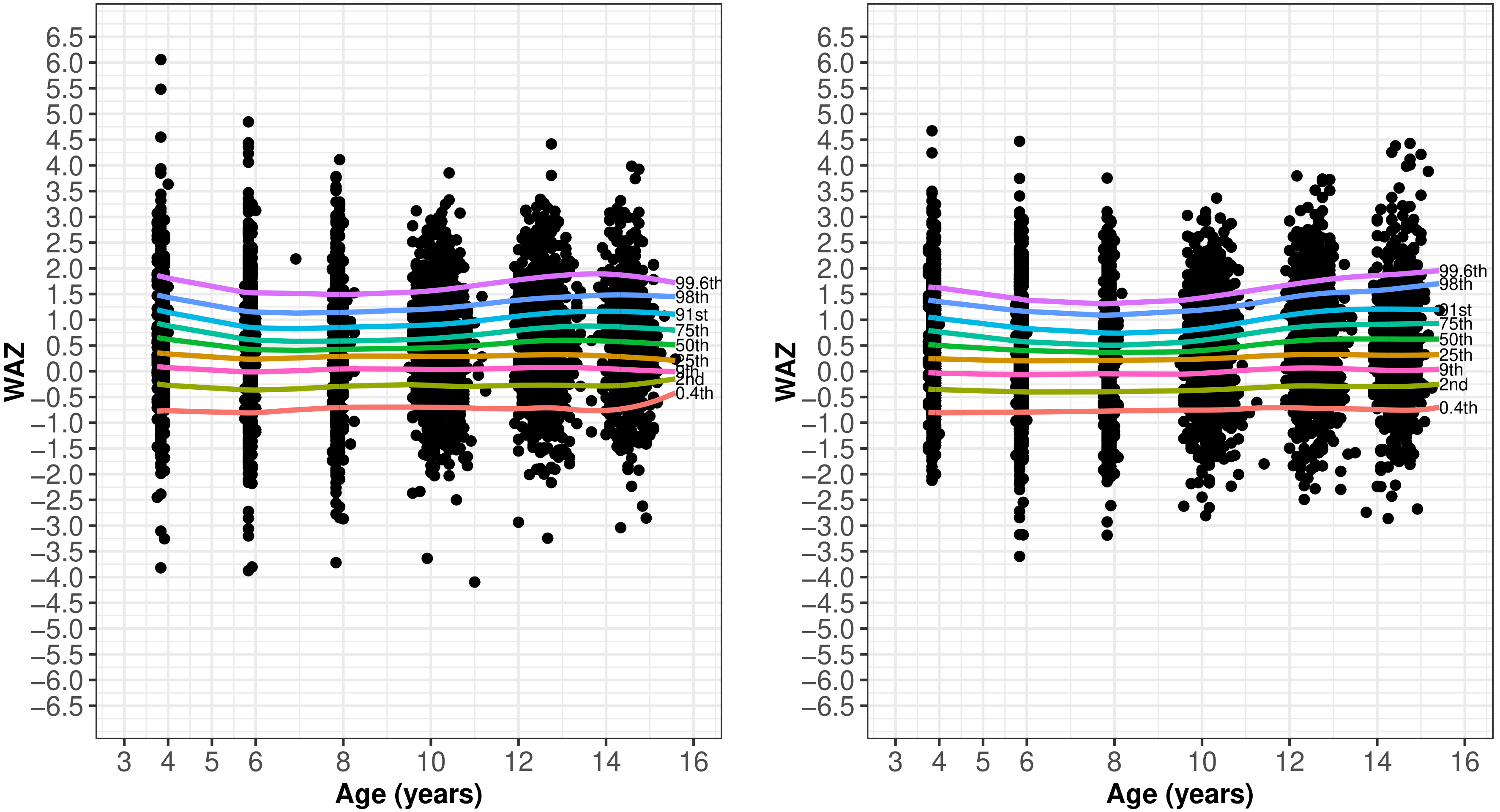
Figure 6.2: Conditional quantile curves for raw weight, estimated using the AQMM with cubic P-splines and fitted with randon intercepts of age, for a child in the reference group across nine quantile levels using the children aged 4 - 14 years GUS dataset.



(a) Male

(b) Female

Figure 6.3: Conditional quantile curves for WAZ, estimated using the AQMM with cubic P-splines and fitted with random intercepts, for a child in the reference group across nine quantile levels using from the entire GUS dataset.



(a) Male

(b) Female

Figure 6.4: Conditional quantile curves for WAZ, estimated using the AQMM with cubic P-splines and fitted with random intercepts, for a child in the reference group across nine quantile levels using from the children aged 4 - 14 years GUS dataset.

Regarding WAZ growth patterns for children aged 4 to 14 years, Figure 6.4 provides analogous evidence that each quantile curve did not fit the dataset well. When considering WAZ growth patterns, both male and female growth patterns generally show a steady progression from ages 4 to 14. While both genders exhibit similar trends, males tend to have higher WAZ scores at the upper quantiles compared to females, indicating slightly different growth trajectories.

Table 6.2: Percentages of observations above each quantile curve in raw height and HAZ for two AQMM models (random intercepts and slopes v.s. only random intercepts.)

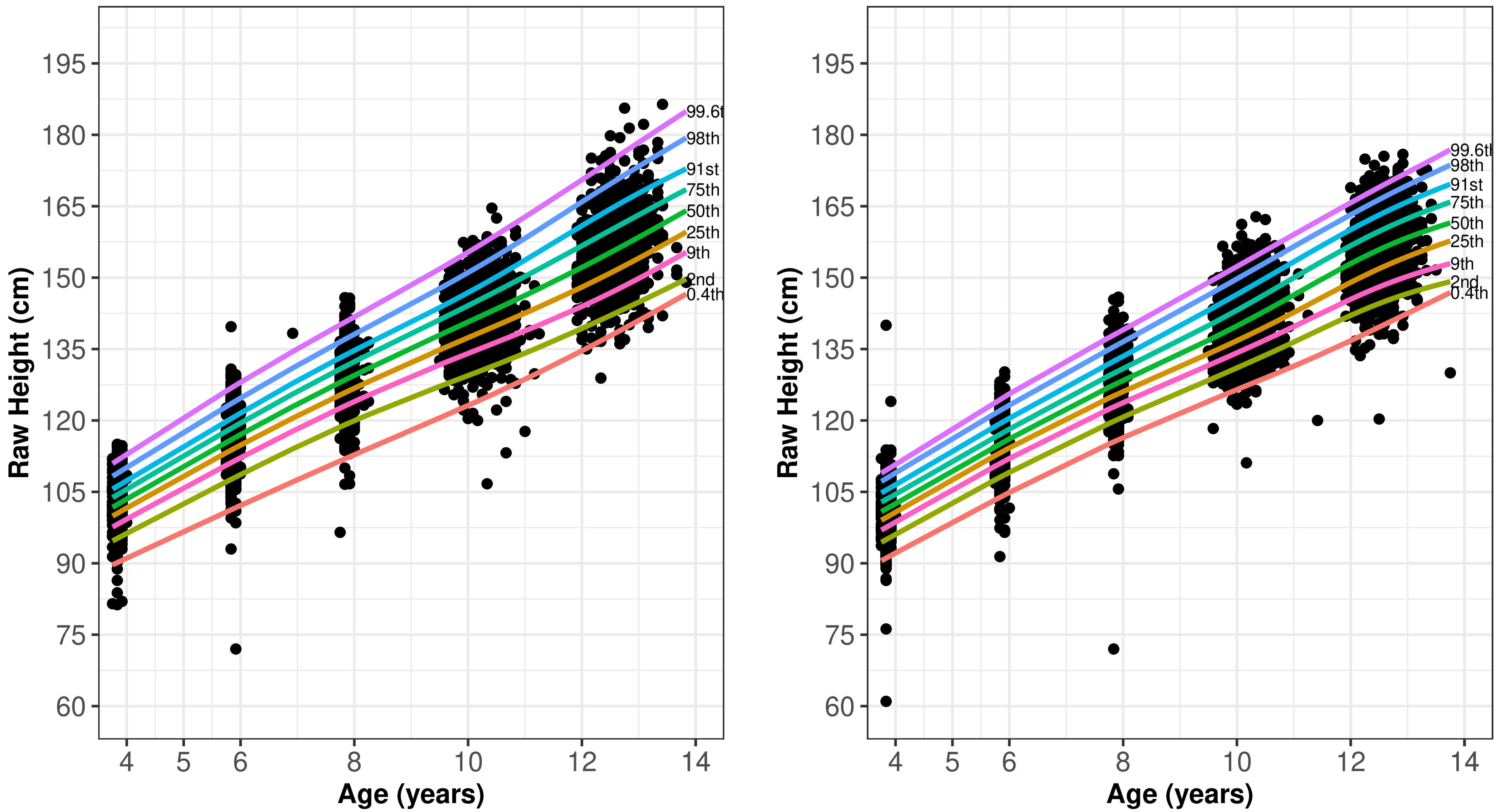
Raw height

Quantiles curves	**Male**		**Female**	
	Intercepts and Slopes	**Intercepts**	**Intercepts and Slopes**	**Intercepts**
0.4th	86.04	99.28	85.09	98.60
2nd	72.14	95.87	67.20	92.77
9th	63.80	84.97	60.27	81.58
25th	56.49	69.61	54.68	67.02
50th	51.33	50.74	50.30	50.22
75th	45.95	31.33	45.04	32.25
91st	39.94	16.00	40.13	17.75
98th	31.29	4.60	32.07	7.37
99.6th	16.34	1.33	20.87	3.02

HAZ

Quantiles curves	**Male**		**Female**	
	Intercepts and Slopes	**Intercepts**	**Intercepts and Slopes**	**Intercepts**
0.4th	85.75	89.12	83.16	88.84
2nd	71.07	76.53	66.73	73.82
9th	63.02	66.26	60.05	63.94
25th	56.51	58.03	54.82	56.81
50th	51.33	51.31	50.57	50.28
75th	46.51	45.22	45.22	43.44
91st	40.39	36.94	40.11	35.66
98th	30.57	25.32	31.91	25.67
99.6th	15.38	10.12	20.95	16.05

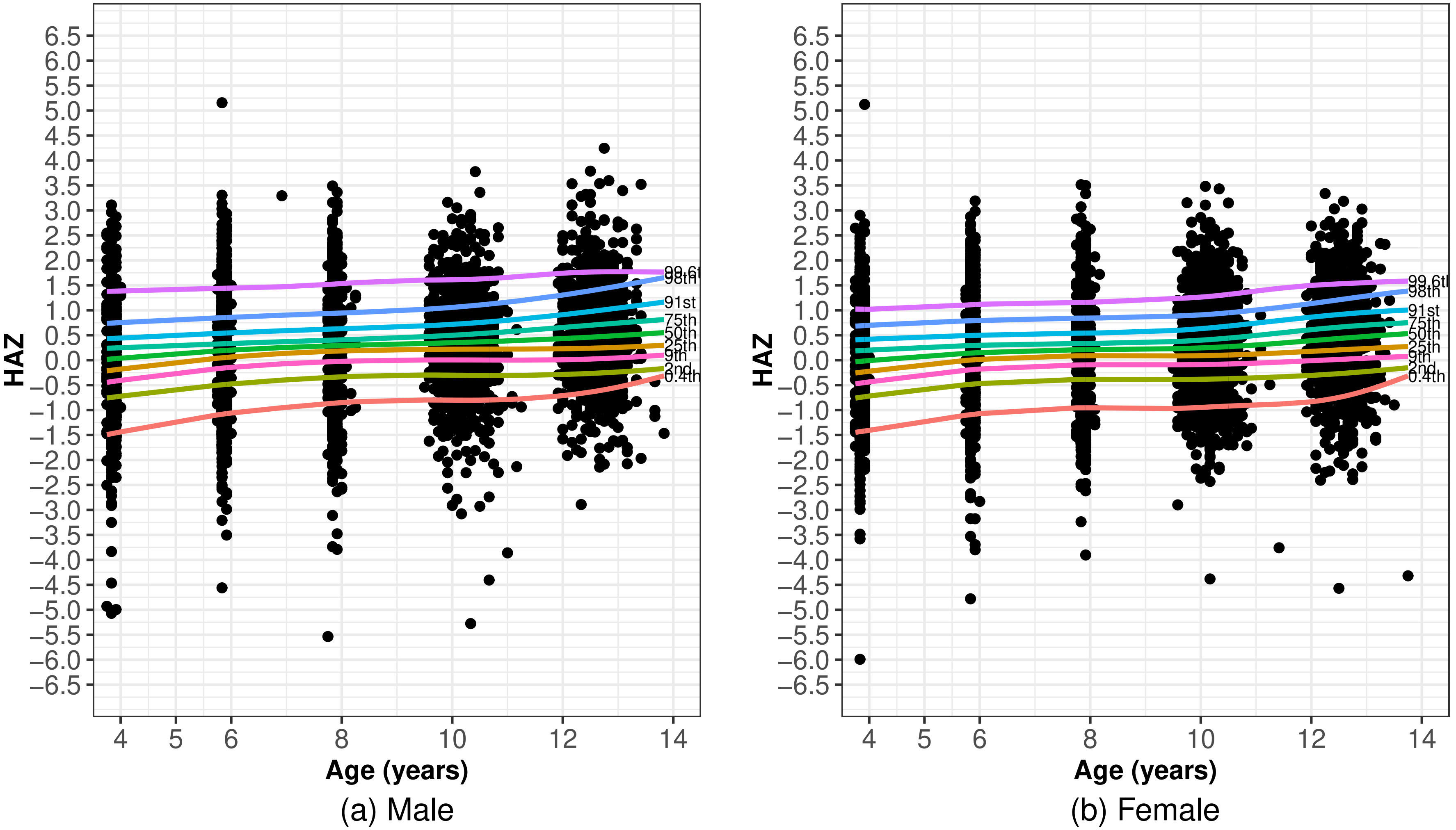
For the raw height-for-age growth patterns, each quantile curve fitted the GUS dataset well, as the percentages of observations above each percentile were not close to the expected value (see Table 6.2). Figure 6.5 shows that both males and females experienced an increase in height from ages 4 to 14. The growth patterns were similar between sexes, with slightly difference in the upper quantile. In males, height increased steadily across all quantiles, indicating consistent growth during childhood and early adolescence. The upper quantiles (91st, 98th, and 99.6th) show a more pronounced growth spurt compared to the lower quantiles. Female growth patterns mirrored those of males, with a steady increase in height across all quantiles. The differences between the lower and upper quantiles were less pronounced compared to males, suggesting a more uniform growth trajectory among females.



(a) Male

(b) Female

Figure 6.5: Conditional quantile curves for raw height, estimated using the AQMM with cubic P-splines and fitted with random intercepts of age, for a child in the reference group across nine quantile levels using the GUS dataset.



(a) Male

(b) Female

Figure 6.6: Conditional quantile curves for HAZ, estimated using the AQMM with cubic P-splines and fitted with random intercepts of age, for a child in the reference group across nine quantile levels using the GUS dataset.

Additionally, for the height-for-age Z-score (HAZ), none of the quantile curves fit the GUS dataset as anticipated (with the exception of the median curve), similar to the observations for WAZ cases. Regarding to HAZ growth patterns for males and females, Figure 6.6 shows that children generally grow at a consistent rate when their height was compared to a reference population of children of the same age and sex.

6.1.5 Results - Risk factors associated with child physical growth measurements

Table 6.3 present a summary of significant risk factors from the initial fitted quantile models for each physical growth measurement. Generally, the results indicate that significant risk factors at the two quantiles vary across physical growth measurement. The estimates from each initial model, using the AQMM with cubic P-splines and corresponding to this summary tables, are presented in Tables D.1 to D.4 in Appendix D. These significant risk factors will be included in fitting the two refined quantile models ($\tau = 0.10$ and $\tau = 0.90$).

Table 6.3: A summary of factors associated with four child growth measurements: initial fitted quantile models using the AQMM with cubic P-splines

	Raw weight		**WAZ**		**Raw height**		**HAZ**	
	$\tau = 0.10$	$\tau = 0.90$	$\tau = 0.10$	$\tau = 0.90$	$\tau = 0.10$	$\tau = 0.90$	$\tau = 0.10$	$\tau = 0.90$
Sex					✓	✓		
Low birth weight	✓	✓	✓	✓	✓	✓	✓	✓
Ethnicity of a child	✓				✓		✓	
Child's health in general		✓				✓	✓	
Number of accidents or injuries of child								
Child's birth order					✓	✓		
Mother's marital status	✓		✓	✓		✓	✓	✓
Urban-rural classification	✓	✓				✓	✓	
Household size					✓		✓	
Mother's age at first child's birth					✓	✓	✓	✓
Respondent's alcoholic drinks		✓	✓	✓		✓		
Respondent's current health			✓	✓			✓	✓
Smoking cigarettes while pregnant			✓	✓	✓			
drinking alcohol while pregnant	✓	✓				✓		
Respondent's health problem(s) in a year						✓		
Respondent's current job						✓	✓	
Deprivation quintile	✓	✓	✓	✓	✓	✓	✓	✓
Equivalised income	✓		✓	✓	✓	✓	✓	✓
Linear basis term of age								

Raw weight

Table 6.4 presents the estimates for both fixed and random effects derived from the AQMM with cubic P-splines for the centred raw weight in children aged 4 - 14 years from the GUS dataset, refitting only significant effects from the initially fitted quantile models (see Table 6.3). The findings indicate that only ***low birth weight*** had a significant impact on (centred) raw weight across both quantiles, with a negative effect whose magnitude varied between them. The coefficient can be interpreted as follows:

Table 6.4: Estimates from the AQMM with cubic P-splines for the **centred raw weight** growth measurement in **children aged 4 - 14 years** from the GUS dataset (standard errors in brackets)†

	$\tau = 0.10$	$\tau = 0.90$
Fixed effects		
(Intercept)	-1.7103 (0.9521)	**2.4926^a (0.8171)**
Low birth weight (Yes)	**-1.2382^a (0.2724)**	**-1.7226^a (0.2660)**
Ethnicity of a child (White)	-0.3465 (0.3974)	-
Child's health in general (Good)	-	0.0868 (0.1354)
Child's health in general (Fair, Bad, Very Bad)	-	-0.0679 (0.3316)
Mother's marital status (Single)	0.0387 (0.0830)	-
Mother's marital status (Other)	0.1251 (0.1516)	-
Urban-rural classification (Other urban)	-0.0067 (0.1354)	-0.2060 (0.1583)
Urban-rural classification (Small, accessible towns)	-0.0129 (0.1687)	-0.1363 (0.1803)
Urban-rural classification (Small, remote towns)	0.0405 (0.2641)	-0.4602 (0.2655)
Urban-rural classification (Accessible rural)	0.0536 (0.1590)	-0.0946 (0.2103)
Urban-rural classification (Remote rural)	0.1565 (0.1469)	-0.4119 (0.2288)
Respondent's alcoholic drinks (Every day)	-	-0.1944 (0.4666)
Respondent's alcoholic drinks (4 - 6 times a week)	-	-0.0420 (0.4212)
Respondent's alcoholic drinks (2 - 3 times a week)	-	0.1181 (0.3813)
Respondent's alcoholic drinks (Once a week)	-	-0.0870 (0.4244)
Respondent's alcoholic drinks (2 - 3 times a month)	-	0.2376 (0.3930)
Respondent's alcoholic drinks (Once a month or less)	-	-0.0450 (0.7048)
Respondent's alcoholic drinks (Not in the last year)	-	0.3265 (0.5041)
Drinking alcohol while pregnant ($\geq$ 3 - 4 times a week)	-0.1776 (0.9699)	-0.5860 (0.6408)
Drinking alcohol while pregnant (1 - 2 times a week)	-0.2852 (0.8578)	-0.1779 (0.7277)
Drinking alcohol while pregnant (2 - 3 times a month)	0.0787 (0.8661)	-0.1383 (0.6006)
Drinking alcohol while pregnant (\textless once a month)	0.0418 (0.8472)	-0.1084 (0.6152)
Deprivation quintile (2)	0.1144 (0.1408)	0.1776 (0.1588)
Deprivation quintile (3)	0.0387 (0.1855)	0.0748 (0.1660)
Deprivation quintile (4)	0.2580 (0.2183)	0.3414 (0.1977)
Deprivation quintile (5)	-0.1682 (0.2249)	0.2368 (0.2051)
Equivalised income	**0.2098^a (0.0478)**	-
Linear basis term of Age (in year)	-73.3176 (72.0365)	-80.1245 (78.2812)
Random effects		
$\hat{\sigma}_0$ (SD of intercepts Age in year)	1.9779	1.9317
$\hat{\sigma}_1$ (SD of slopes of the Age in year)	0.6281	0.6134
$\hat{\rho}_{01}$ (Correlation of intercepts and slopes)	-0.9709	-0.9669

^{a} p < 0.001, ^{b} p < 0.005, ^{c} p < 0.05

† The reference categories are given in Tables 2.8 to 2.11.

For children in the lowest 10% of the centred raw weight distribution, having a low birth weight is associated with a centred raw weight that is lower by 1.24 kg compared to those without low birth weight. In contrast, for children in the highest 90% of the centred raw weight distribution, having a low birth weight is associated with centred raw weight lower by 1.72 kg compared to those without low birth weight.

Notably, ***equivalised income*** was a significant factor only at the 0.10th quantile, showing a positive effect. The coefficient indicates that, for children at the lower 10% of the centred raw weight distribution, a one standard deviation increase in equivalised income (£13097.16) is associated with an expected increase in centred raw weight of 0.21 kg.

Regarding random effects, the estimated standard deviations for the intercepts and temporal (age) slope effects were similar across the two quantiles. Additionally, the correlations between these random effects remained consistently strong and negative across both quantiles. This suggests that children with a higher baseline centred raw weight (approximately 4 years of age) tend to grow at a slower rate over time.

WAZ

Table 6.5: Estimates from the AQMM with cubic P-splines for the **centred WAZ** growth measurement in **children aged 4 - 14 years** the GUS dataset (standard errors in brackets)†

	$\tau = 0.10$	$\tau = 0.90$
Fixed effects		
(Intercept)	-0.3880 (0.2385)	0.2885 (0.1805)
Low birth weight (Yes)	**-0.6220^a (0.1201)**	**-0.5519^a (0.1037)**
Mother's marital status (Single)	-0.0015 (0.0130)	-0.0068 (0.0164)
Mother's marital status (Other)	0.0479 (0.0290)	0.0329 (0.0255)
Respondent's alcoholic drinks (Every day)	-0.1014 (0.2233)	-0.2491 (0.1690)
Respondent's alcoholic drinks (4 - 6 times a week)	-0.0301 (0.1926)	-0.0669 (0.1393)
Respondent's alcoholic drinks (2 - 3 times a week)	0.0223 (0.2043)	-0.0824 (0.1474)
Respondent's alcoholic drinks (Once a week)	0.0088 (0.2115)	-0.0738 (0.1591)
Respondent's alcoholic drinks (2 - 3 times a month)	0.0594 (0.2033)	-0.0191 (0.1557)
Respondent's alcoholic drinks (Once a month or less)	-0.0015 (0.2295)	-0.0394 (0.1980)
Respondent's alcoholic drinks (Not in the last year)	0.1764 (0.2039)	0.1685 (0.1461)
Respondent's current health (Very good)	**0.0347 (0.0242)**	**0.0348^c (0.0160)**
Respondent's current health (Good)	0.0484 (0.0254)	**0.0536^b (0.0179)**
Respondent's current health (Fair, Poor)	0.0765 (0.0401)	0.0358 (0.0292)
Smoking cigarettes while pregnant (Yes)	**0.1365^c (0.0540)**	**0.1298^c (0.0532)**
Deprivation quintile (2)	0.0457 (0.0265)	**0.0481^c (0.0222)**
Deprivation quintile (3)	0.0609 (0.0360)	0.0478 (0.0273)
Deprivation quintile (4)	**0.1265^b (0.0459)**	**0.1177^b (0.0364)**
Deprivation quintile (5)	0.0493 (0.0531)	0.0387 (0.0456)
Equivalised income	**0.0266^b (0.0080)**	**0.0169^c (0.0071)**
Linear basis term of Age (in year)	-0.1317 (0.1999)	-0.3207 (0.3350)
Random effects		
$\hat{\sigma}_0$ (SD of intercepts Age in year)	0.2250	0.2190
$\hat{\sigma}_1$ (SD of slopes of the Age in year)	0.0150	0.0147
$\hat{\rho}_{01}$ (Correlation of intercepts and slopes)	-0.3500	-0.3504

^{a} p < 0.001, ^{b} p < 0.005, ^{c} p < 0.05
† The reference categories are given in Tables 2.8 to 2.11.

Table 6.5 presents the estimates for both fixed and random effects derived from the AQMM with cubic P-splines, as applied to the centred WAZ in children aged 4 - 14 years from the GUS dataset. In this child growth measurement, ***low birth weight***, ***respondent's current health***, ***smoking cigarettes while pregnant***, ***deprivation quintile***, and ***equivalised income*** were found to be significantly associated with the centred WAZ across two quantiles. Among these factors, only ***low birth weight*** exhibited a negative effect, with a slight difference in magnitude across the two quantiles. The coefficients can be interpreted as follows:

For children in the lowest 10% of the centred WAZ distribution, having a low birth weight is associated with a decrease of 0.62 in centred WAZ compared to those without low birth weight. Living with a respondent in very good health is associated with an increase of 0.03 in centred WAZ compared to living with an respondent in excellent health. Smoking cigarettes during pregnancy is associated with an increase of 0.14 in centred WAZ compared to not smoking cigarettes during pregnant. Children in deprivation quintile 4 have a centred WAZ that is 0.13 units higher than children in deprivation quintile 1. Additionally, a one standard deviation increase in equivalised income (£13097.16) is expected to result in an increase of 0.03 in centred WAZ.

For children in the highest 90% of the centred WAZ distribution, having a low birth weight is associated with a decrease of 0.55 in centred WAZ compared to those without low birth weight. Living with a respondent in very good and good health is associated with increases of 0.03 and 0.05 in centred WAZ, respectively, compared to living with an respondent in excellent health. Smoking cigarettes during pregnancy is associated with an increase of 0.13 in centred WAZ compared to not smoking cigarettes during pregnant. Living in deprivation quintiles 2 and 4 is associated with increases of 0.05 and 0.13 in centred WAZ, respectively, compared to living in deprivation quintile 1. Additionally, a one standard deviation increase in equivalised income (£13097.16) is expected to result in an increase of 0.02 in centred WAZ.

For the random effects, the estimated standard deviations for the intercepts and temporal (age) slope effects were similar across two extreme quantiles, with the temporal slope effects showing relatively low variability. The correlations were negative and exhibited medium strength across the different quantiles, indicating that children with a higher baseline centered WAZ tend to grow at a slower rate over time.

Raw height

Estimates of the fixed effects and standard deviations of random effects from the cubic AQMM with P-splines for the centred raw height are presented in Table 6.6. The results

Table 6.6: Estimates from the AQMM with cubic P-splines for the **(centred) raw height** growth measurement in **children aged 4 - 14 years** the GUS dataset (standard errors in brackets)†

	$\tau = 0.10$	$\tau = 0.90$
Fixed effects		
(Intercept)	**-1.3512^{c} (0.5566)**	-0.0612 (1.3345)
Sex (Male)	**1.1000^{a} (0.1720)**	**1.1638^{a} (0.1933)**
Low birth weight (Yes)	**-2.6014^{a} (0.4306)**	**-2.1979^{a} (0.4047)**
Ethnicity of a child (White)	-0.6731 (0.6254)	-
Child's health in general (Good)	-	**0.3541^{a} (0.0877)**
Child's health in general (Fair, Bad, Very Bad)	-	-0.1301 (0.2533)
Child's birth order	**-0.2101^{c} (0.0969)**	**-0.2330^{b} (0.0815)**
Mother's marital status (Single)	-	0.0384 (0.0761)
Mother's marital status (Other)	-	**0.3840^{c} (0.1700)**
Urban-rural classification (Other urban)	-	-0.0959 (0.1968)
Urban-rural classification (Small, accessible towns)	-	-0.2405 (0.3070)
Urban-rural classification (Small, remote towns)	-	0.1963 (0.4028)
Urban-rural classification (Accessible rural)	-	-0.2488 (0.2199)
Urban-rural classification (Remote rural)	-	-0.3100 (0.2970)
Household size	0.0679 (0.0677)	-
Mother's age at first child's birth (<20 years old)	0.7368 (0.3959)	0.5159 (0.4121)
Mother's age at first child's birth ($\geq$ 30 years old)	0.1135 (0.4222)	0.0540 (0.4181)
Respondent's alcoholic drinks (Every day)	-	0.9584 (0.8032)
Respondent's alcoholic drinks (4 - 6 times a week)	-	0.7671 (0.8798)
Respondent's alcoholic drinks (2 - 3 times a week)	-	0.7227 (0.8362)
Respondent's alcoholic drinks (Once a week)	-	0.8145 (0.8340)
Respondent's alcoholic drinks (2 - 3 times a month)	-	0.8045 (0.7806)
Respondent's alcoholic drinks (Once a month or less)	-	0.5335 (0.7598)
Respondent's alcoholic drinks (Not in the last year)	-	0.3313 (0.8877)
Smoking cigarettes while pregnant (Yes)	**-0.5073^{c} (0.2132)**	-
Drinking alcohol while pregnant ($\geq$ 3 - 4 times a week)	-	**6.8726^{a} (1.0222)**
Drinking alcohol while pregnant (1 - 2 times a week)	-	0.6403 (0.9733)
Drinking alcohol while pregnant (2 - 3 times a month)	-	0.0990 (0.9753)
Drinking alcohol while pregnant (<once a month)	-	0.1164 (0.9519)
Respondent's health problem(s) in a year (Yes)	-	**0.2965^{b} (0.1084)**
Respondent's current job (No)	-	0.0607 (0.2155)
Deprivation quintile (2)	0.0837 (0.1208)	0.2051 (0.1360)
Deprivation quintile (3)	0.1057 (0.1411)	0.1975 (0.1424)
Deprivation quintile (4)	0.2482 (0.1784)	**0.3291^{c} (0.1501)**
Deprivation quintile (5)	-0.2253 (0.2205)	0.0014 (0.2082)
Equivalised income	0.0714 (0.0432)	0.0691 (0.0431)
Linear basis term of Age (in year)	-90.3125 (79.7049)	-92.7692 (81.5810)
Random effects		
$\hat{\sigma}_0$ (SD of intercepts Age in year)	1.5561	1.7356
$\hat{\sigma}_1$ (SD of slopes of the Age in year)	0.2556	0.2653
$\hat{\rho}_{01}$ (Correlation of intercepts and slopes)	0.3343	0.3070

^{a} $p < 0.001$, ^{b} $p < 0.005$, ^{c} $p < 0.05$

† The reference categories are given in Tables 2.8 to 2.11.

show that three factors were consistently associated with (centred) raw height across the two quantiles, exhibiting consistent effect magnitudes: ***sex***, ***low birth weight***, and ***child's***

birth order. At the 0.10th quantile, ***smoking cigarettes while pregnant*** emerged as additional significant factor associated with centred raw height. Meanwhile, at the 0.90th quantile, factors such as the ***child's health in general***, ***the mother's marital status***, the ***drinking alcohol while pregnant***, and ***deprivation quintile*** were also associated with centred raw height.

For children in the lowest 10% of the centred raw height distribution, males have a centred raw height that is 1.10 cm higher than females. Low birth weight is associated with centred height lower by 2.60 cm compared to those with normal birth weight. Smoking cigarettes during pregnancy is associated with a reduction of 0.51 cm in centred raw height compared to not smoking. Additionally, a one standard deviation increase in the child's birth order (0.80) is expected to result in a decrease of 0.21 cm in centred raw height.

For children in the highest 90% of the centred raw height, males have a centred raw height that is 1.16 cm higher than females. Low birth weight is associated with a decrease of 2.20 cm compared to those with normal birth weight. Children in good health have a centred raw height that is 0.35 cm higher than those in very good health. Children living with mother of an other marital status have a centred raw height that is 0.38 cm higher than those living with a married mother. Drinking alcohol more than 3 to 4 times a week during pregnancy is associated with an increase of 6.87 cm in centred raw height compared to not drinking. Children living with a respondent who has a health problem have a centred raw height that is 0.30 cm higher compared to those living with a respondent with no health problems. Children in deprivation quintile 4 have a centred raw height that is 0.32 cm higher than children in deprivation quintile 1. Additionally, a one standard deviation increase in the child's birth order (1.00) is expected to result in a decrease of 0.23 cm in centred raw height.

When focusing on the random effects, the estimated standard deviations reveal greater variability in individual intercepts and less variability in individual slopes across the two quantiles. Notably, the deviations at the 0.10th quantile were slightly smaller than those at the 0.90th quantile. The correlations between these random effects were medium and positive across the two quantiles, indicating that children with higher initial heights tend to have a steeper or more pronounced growth trajectory.

HAZ

Table 6.7 presents estimates of the fixed effects and standard deviations of random effects from the AQMM with cubic P-splines for the centred HAZ. The analysis shows that factors such as ***low birth weight*** and ***equivalised income*** were associated with the centred HAZ across two quantiles. The effect of ***low birth weight*** was generally consistent, while the effect of ***equivalised income*** varied between quantiles. Additionally, the ***ethnicity of a child***, ***mother's marital status***, and ***respondent's current health*** were significant factors associated with the centred HAZ at the 0.10th quantile only. Each coefficient can be interpreted as follows:

Table 6.7: Estimates from the AQMM with cubic P-splines for the **centred HAZ** growth measurement in **children aged 4 - 14 years** the GUS dataset (standard errors in brackets)†

	$\tau = 0.10$	$\tau = 0.90$
Fixed effects		
(Intercept)	-0.1925 (0.2253)	**0.2547^c (0.1141)**
Low birth weight (Yes)	**-0.4211^c (0.1883)**	**-0.4329^a (0.0904)**
Ethnicity of a child (White)	**-0.4811^c (0.1817)**	-
Child's health in general (Good)	0.0254 (0.0283)	-
Child's health in general (Fair, Bad, Very Bad)	0.1219 (0.1067)	-
Mother's marital status (Single)	0.0260 (0.0365)	-0.0188 (0.0113)
Mother's marital status (Other)	**0.1242^c (0.0541)**	0.0267 (0.0249)
Urban-rural classification (Other urban)	0.0318 (0.0719)	-
Urban-rural classification (Small, accessible towns)	0.1216 (0.1389)	-
Urban-rural classification (Small, remote towns)	0.2030 (0.1678)	-
Urban-rural classification (Accessible rural)	0.0517 (0.1004)	-
Urban-rural classification (Remote rural)	0.1841 (0.1215)	-
Household size	0.0279 (0.0202)	-
Mother's age at first child's birth (<20 years old)	0.2246 (0.1463)	0.0791 (0.0728)
Mother's age at first child's birth ($\geq$ 30 years old)	0.0630 (0.1444)	0.0232 (0.0736)
Respondent's current health (Very good)	0.1106 (0.0596)	0.0020 (0.0207)
Respondent's current health (Good)	**0.1216^c (0.0575)**	0.0227 (0.0237)
Respondent's current health (Fair, Poor)	0.1543 (0.0912)	0.0249 (0.0303)
Respondent's current job (No)	0.1641 (0.1068)	-
Deprivation quintile (2)	0.1393 (0.0853)	0.0039 (0.0256)
Deprivation quintile (3)	0.1438 (0.0902)	-0.0101 (0.0287)
Deprivation quintile (4)	0.2344 (0.1203)	0.0193 (0.0291)
Deprivation quintile (5)	0.1815 (0.1363)	-0.0155 (0.0368)
Equivalised income	**0.0945^c (0.0388)**	**0.0167^c (0.0065)**
Linear basis term of Age (in year)	-0.6118 (0.7109)	-0.9088 (0.8120)
Random effects		
$\hat{\sigma}_0$ (SD of intercepts Age in year)	0.2670	0.2298
$\hat{\sigma}_1$ (SD of slopes of the Age in year)	0.0117	0.0100
$\hat{\rho}_{01}$ (Correlation of intercepts and slopes)	-0.1126	-0.2645

^{a} $p < 0.001$, ^{b} $p < 0.005$, ^{c} $p < 0.05$

† The reference categories are given in Tables 2.8 to 2.11.

For children in the lowest 10% of the centred HAZ distribution, low birth weight is associated with a decrease of 0.42 units compared to those with normal birth weight. White children have a centred HAZ that is 0.48 units lower than that of children from other ethnicities. Living with a mother of an other marital status have a centred HAZ that is 0.12 units higher than those living with a married mother. Living with a respondent in good health have a centred HAZ that is 0.12 units higher compared to those living with a respondent in excellent health. Additionally, a one standard deviation increase in equivalised income (£13097.16) is expected to result in an increase of 0.09 in centred WAZ. For children in the hightest 90% of the centred HAZ distribution, low birth weight is associated with a decrease of 0.43 units compared to those with normal birth weight. Additionally, a one standard deviation increase in equivalised income (£13097.16) is expected to result in an increase of 0.02 in centred WAZ.

For the random effects, it was observed that the variability in individual intercepts and temporal slopes at each quantile was relatively small and very small, respectively, as indicated by the low values of the estimated standard deviations. Additionally, the correlation between these random effects were relatively small and negative across the two quantiles, indicating that children with a higher baseline centered HAZ tend to grow at a slower rate over time.

6.1.6 Summary

It can generally be concluded that reference growth patterns vary between the sexes and different child growth measurements when analysing the GUS data in both datasets – the entire GUS dataset (children aged 10 months to 14 years) and the dataset for children aged 4 to 14 using the AQMM method. This underscores the importance of considering sex-specific approaches in child growth pattern analysis, recognising that males and females may exhibit distinct growth patterns. In terms of raw weight, each smooth quantile curve appears to parallel the others across different ages for both males and females. For the entire GUS dataset, the quantile curves indicate that children experience rapid growth in the first 4 years, then stabilise up to the age of 10 years, followed by another increase. While this trend remains consistent when considering the WAZ, the quantile curves show differences due to the varying scales used. For the dataset of children aged 4 to 14 years, each quantile curve shows a steady increase in raw weight ovrer this age range. For both raw height and HAZ, each quantile curve is nearly linear for both males and females, indicating consistent height growth in children.

Regarding the identification of risk factors, the significant factors varied across the lower (0.10th) and upper (0.90th) quantiles of the physical growth measurements. Additionally, variations in child growth measurements, whether assessed through raw or z-score scales, emphasise the importance of considering the unique characteristics of each measurement approach. Each approach captures different aspects of child growth and is influenced by its own set of risk factors. Notably, different quantiles of child growth measurements are associated with distinct risk factors, reflecting the diverse groups of children they represent.

6.2 Modelling LCGD in Scotland using Bayesian sparse group LASSO-mixed quantile regression model

In this section, the Bayesian sparse group LASSO-mixed quantile regression model (BSGSSMQR), as outlined in Section 5.8, was applied to the GUS data described in Section 2.6. This analysis focused exclusively on identifying risk factors associated with growth measurements at two distinct locations (the 0.10th and 0.90th quantiles) during the child growth period, including the adolescent period (children aged 4 to 14 years), similar to the approach in Section 6.1. Moreover, it also aimed to determine the appropriate variability of individual linear trends (intercepts and age slopes) through the selection of random effects. My focus was narrowed to raw weight and WAZ as child growth measurements, due to the particular concern of childhood obesity in Scotland. This choice was motivated by the Scottish Heath Survey Report highlighted an increase in the prevalence of childhood obesity in 2021 compared to the 2017 figures (Birtwistle et al., 2022). Covariates were considered in a manner similar to that described in Section 6.1.2.

6.2.1 Fitting the model

The quantile models with $\tau = 0.10$ and 0.90 were fitted. Within this analytical framework, age was exclusively incorporated as a fixed smooth component in each quantile model. This was achieved by using the cubic B-spline method to establish the bases for this smooth term. The determination of the number of knots (K) adhered to the guidelines of Ruppert et al. (2003), setting K as $\min\{40, \text{the number of unique } x/4\}$. These knots were strategically placed at quantiles of the variable age, offering insights into how the effects of age vary across different segments of the growth measurement distribution. Note that this strategy differs from the knot positioning used in Section 6.1 due to differences in spline methods used, as P-splines require equidistant knots, whereas B-splines offer flexibility in this regard. Moreover, evidence suggests that quantile-based knots are more suitable than equidistant knots for fitting data with B-splines, as they adapt better to the distribution of the data and can improve the accuracy of the fit, particularly in cases of data with explicit peaks (Harrell, 2015; Maturana-Russel & Meyer, 2021). All other independent variables in each model were assumed to be linear. Additionally, both raw weight and WAZ were centred by their mean, and each covariate, including bases of age, and dummy variables, was standardised. Furthermore, a design matrix of random effects, $\mathbf{Z} = [\mathbf{1}, X_1]$, wherein X_1 represents the variable of age, was established. This configuration serve as an instrument to account for the inherent potential heterogeneity among children by incorporating random intercepts and slopes.

Regarding the selected prior specifications, I chose $a_1 = a_2 = c_1, = c_2 = 1$ for the Beta priors corresponding to both π_0 and π_1, respectively. Similarly, the inverse Gamma priors employed for σ were configured with $g_1 = g_2 = 0.1$. Additionally, a fixed value of $p_{l'0} = 0.5$ was assigned to $d_{l'}$. The prior mean and variance for $\boldsymbol{a}$ were set to $\mathbf{0}$ and $0.5\mathbf{I}$, respectively, denoted as $\mathbf{A}_0 = 0.5\mathbf{I}$. The hyperparameter t of s^2 was estimated through a Monte Carlo EM algorithm with 100 updates and iterations. To ensure adequate convergence and accuracy, I ran the Gibbs sampler for 15,000 iterations, with a burn-in period of 5000 iterations. Subsequently, I used 10,000 samplings of the posteriors to summarise the model parameters effectively.

6.2.2 Results

The style of the tables in this section, particularly in presenting the results of selecting linear fixed effects, varies from the previous section to enable a comparison between the posterior mean and posterior median for each quantile model. Subsequently, two quantile models for each growth measurement are presented in individual tables.

Raw weight

Table 6.8 presents the posterior mean and median estimates of the basis functions of age related to (centred) raw weight in males across two quantile models ($\tau = 0.10$ and $\tau = 0.90$). The results indicate that the posterior median estimator produced zero estimates for the basis function 22 (**S(Age22)**) of age only across all two quantile models. This suggests that this basis function is not requisite for capturing the non-linear relationship between age and (centred) raw weight. Another interesting result is that the two quantile models appeared to have similar estimates. Figure 6.7 shows the impact of age on centred raw weight values for two quantiles: the 0.10th quantile (solid red line) and the 0.90th quantile (dashed blue line). Both lines indicate that as children age from 4 to 14 years, the centred raw weight exhibits a non-linear growth pattern. This non-linear aged effect appears to fluctuate throughout this period, with a slight upward trend as age increases, suggesting a positive relationship between age and centred raw weight. Such fluctuation may be attributed to factors such as gender differences, data heterogeneity, or interaction effects. For instance, the differences in growth trajectories between males and females, are more clearly captured when analysed separately.

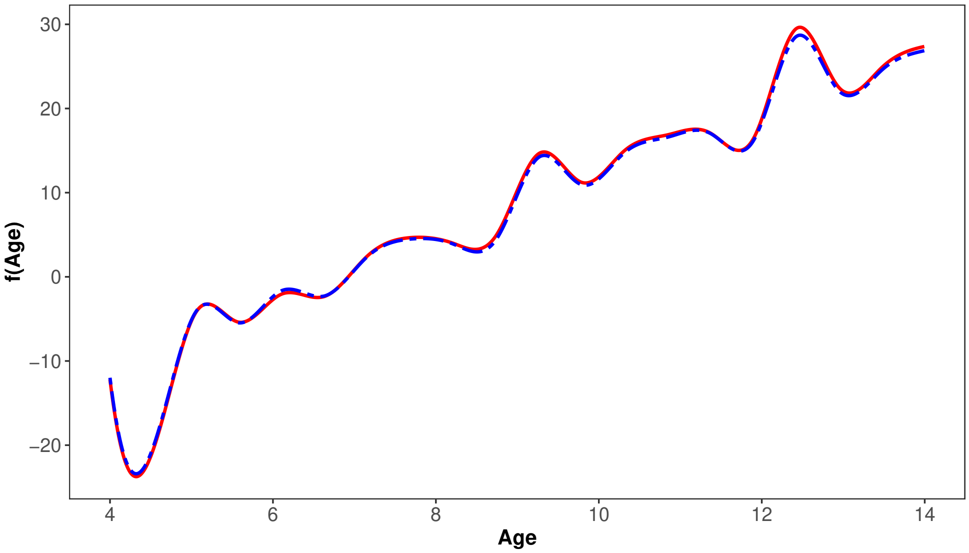


Figure 6.7: Smoothed effect of age for the (centred) raw weight (*solid* red *line*: $\tau = 0.10$, and *two dashed* blue *line*: $\tau = 0.90$)

Table 6.8: Posterior mean, standard deviation (SD), and posterior median for fixed (basis terms) effects across two quantile models for the centred **raw weight**

Fixed effects	$\tau = 0.10$			$\tau = 0.90$		
(basis terms)	**Mean**	**SD**	**Median**	**Mean**	**SD**	**Median**
S(Age1)	-3.28	0.62	-3.23	-3.25	0.63	-3.20
S(Age2)	3.42	0.45	3.38	3.43	0.48	3.38
S(Age3)	0.83	0.34	0.84	0.69	0.38	0.72
S(Age4)	1.71	0.25	1.71	1.78	0.27	1.77
S(Age5)	0.48	0.35	0.47	0.46	0.35	0.45
S(Age6)	3.14	0.23	3.13	3.10	0.24	3.08
S(Age7)	2.63	0.25	2.66	2.55	0.26	2.58
S(Age8)	1.82	0.37	1.82	1.81	0.37	1.81
S(Age9)	0.58	0.38	0.66	0.55	0.38	0.62
S(Age10)	0.49	0.23	0.51	0.48	0.24	0.50
S(Age11)	1.26	0.47	1.39	1.25	0.47	1.36
S(Age12)	4.58	0.24	4.57	4.50	0.24	4.51
S(Age13)	4.15	0.25	4.16	4.06	0.25	4.07
S(Age14)	3.43	0.36	3.45	3.39	0.36	3.43
S(Age15)	0.99	0.51	1.17	0.97	0.51	1.15
S(Age16)	1.13	0.23	1.14	1.10	0.24	1.10
S(Age17)	2.07	0.50	2.17	2.05	0.49	2.14
S(Age18)	4.74	0.36	4.74	4.68	0.36	4.68
S(Age19)	5.54	0.23	5.55	5.42	0.24	5.43
S(Age20)	4.24	0.34	4.26	4.20	0.34	4.22
S(Age21)	1.23	0.48	1.33	1.20	0.48	1.31
S(Age22)	0.16	0.20	-	0.16	0.20	-
S(Age23)	0.50	0.17	0.50	0.50	0.17	0.49

Table 6.9: Posterior mean, standard deviation (SD), and posterior median for both fixed (linear predictors) and random effects of the **0.10**th quantile model for the **raw weight**†

Fixed effects (linear predictors)	**Mean**	**SD**	**Median**
Sex (Male)	0.01	0.04	-
Low birth weight (Yes)	-0.35	0.12	**-0.36**
Ethnicity of a child (White)	-0.01	0.04	-
Child's health in general (Good)	0.00	0.03	-
Child's health in general (Fair, Bad, Very Bad)	-0.03	0.04	-
Number of accidents or injuries of child	0.02	0.04	-
Child's birth order	0.00	0.05	-
Mother's marital status (Single)	-0.03	0.04	-
Mother's marital status (Other)	0.04	0.05	-
Urban-rural classification (Other urban)	-0.02	0.05	-
Urban-rural classification (Small, accessible towns)	-0.01	0.04	-
Urban-rural classification (Small, remote towns)	-0.02	0.05	-
Urban-rural classification (Accessible rural)	0.02	0.05	-
Urban-rural classification (Remote rural)	0.01	0.04	-
Household size	-0.01	0.05	-
Mother's age at first child's birth ($<$ 20 years old)	0.02	0.05	-
Mother's age at first child's birth ($\geq$ 30 years old)	-0.01	0.05	-
Respondent's alcoholic drinks (Every day)	-0.01	0.04	-
Respondent's alcoholic drinks (4 - 6 times a week)	-0.02	0.05	-
Respondent's alcoholic drinks (2 - 3 times a week)	0.00	0.04	-
Respondent's alcoholic drinks (Once a week)	-0.02	0.05	-
Respondent's alcoholic drinks (2 -3 times a month)	0.01	0.04	-
Respondent's alcoholic drinks (Once a month or less)	-0.01	0.05	-
Respondent's alcoholic drinks (Not in the last year)	0.04	0.06	-
Respondent's current health (Very good)	-0.02	0.04	-
Respondent's current health (Good)	0.03	0.05	-
Respondent's current health (Fair, Poor)	0.01	0.04	-
Smoking cigarettes while pregnant (Yes)	0.05	0.08	-
Drinking alcohol while pregnant ($\geq$ 3 - 4 times a week)	-0.01	0.04	-
Drinking alcohol while pregnant (1 - 2 times a week)	0.00	0.04	-
Drinking alcohol while pregnant (2 - 3 times a month)	-0.01	0.04	-
Drinking alcohol while pregnant ($<$ once a month)	-0.02	0.05	-
Respondent's health problem(s) in a year (Yes)	0.01	0.04	-
Respondent's current job (No)	-0.01	0.03	-
Deprivation quintile (2)	0.01	0.04	-
Deprivation quintile (3)	0.00	0.04	-
Deprivation quintile (4)	0.13	0.10	**0.15**
Deprivation quintile (5)	0.00	0.04	-
Equivalised income	0.10	0.09	**0.10**
Standard deviation (Random effects)			
$\hat{\sigma}_0$ (Intercept Age)	2.79	0.25	**2.79**
$\hat{\sigma}_1$ (Slope of Age)	1.91	0.14	**1.91**

† The reference categories are given in Tables 2.9 to 2.11.

Table 6.10: Posterior mean, standard deviation (SD), and posterior median for both fixed (linear predictors) and random effects of the **0.90**th quantile model for the **raw weight**†

Fixed effects (linear predictors)	**Mean**	**SD**	**Median**
Sex (Male)	0.01	0.04	-
Low birth weight (Yes)	-0.34	0.12	**-0.35**
Ethnicity of a child (White)	-0.01	0.04	-
Child's health in general (Good)	0.00	0.03	-
Child's health in general (Fair, Bad, Very Bad)	-0.02	0.04	-
Number of accidents or injuries of child	0.02	0.04	-
Child's birth order	0.00	0.05	-
Mother's marital status (Single)	-0.03	0.04	-
Mother's marital status (Other)	0.03	0.05	-
Urban-rural classification (Other urban)	-0.02	0.05	-
Urban-rural classification (Small, accessible towns)	-0.01	0.04	-
Urban-rural classification (Small, remote towns)	-0.02	0.05	-
Urban-rural classification (Accessible rural)	0.02	0.05	-
Urban-rural classification (Remote rural)	0.01	0.04	-
Household size	-0.01	0.05	-
Mother's age at first child's birth ($<$ 20 years old)	0.02	0.05	-
Mother's age at first child's birth ($\geq$ 30 years old)	-0.01	0.04	-
Respondent's alcoholic drinks (Every day)	-0.01	0.04	-
Respondent's alcoholic drinks (4 - 6 times a week)	-0.02	0.05	-
Respondent's alcoholic drinks (2 - 3 times a week)	0.00	0.04	-
Respondent's alcoholic drinks (Once a week)	-0.02	0.05	-
Respondent's alcoholic drinks (2 -3 times a month)	0.01	0.04	-
Respondent's alcoholic drinks (Once a month or less)	-0.01	0.05	-
Respondent's alcoholic drinks (Not in the last year)	0.03	0.06	-
Respondent's current health (Very good)	-0.02	0.04	-
Respondent's current health (Good)	0.03	0.05	-
Respondent's current health (Fair, Poor)	0.01	0.04	-
Smoking cigarettes while pregnant (Yes)	0.05	0.08	-
Drinking alcohol while pregnant ($\geq$ 3 - 4 times a week)	-0.01	0.04	-
Drinking alcohol while pregnant (1 - 2 times a week)	0.00	0.04	-
Drinking alcohol while pregnant (2 - 3 times a month)	-0.01	0.04	-
Drinking alcohol while pregnant ($<$ once a month)	-0.02	0.05	-
Respondent's health problem(s) in a year (Yes)	0.01	0.04	-
Respondent's current job (No)	-0.01	0.03	-
Deprivation quintile (2)	0.01	0.04	-
Deprivation quintile (3)	0.00	0.04	-
Deprivation quintile (4)	0.13	0.09	**0.14**
Deprivation quintile (5)	0.00	0.04	-
Equivalised income	0.09	0.09	**0.10**
Standard deviation (Random effects)			
$\hat{\sigma}_0$ (Intercept Age)	2.79	0.25	**2.79**
$\hat{\sigma}_1$ (Slope of Age)	1.92	0.14	**1.92**

† The reference categories are given in Tables 2.9 to 2.11.

Tables 6.9 to 6.10 present the posterior mean and median estimates of the fixed effects, as well as the standard deviations of the random effects, for (centred) raw weight in males across two quantile models. The fixed effects, including ***low birth weight***, ***deprivation quintile***, and ***equivalised income***, were selected for both the 0.10th and 0.90th quantile models. Regarding the random effects, both the random intercept age and the random slope of age were selected in all two quantile models, with the former exhibiting higher variability than the later. This indicates that there was variability in individual linear trends (intercepts and temporal slopes) across the two quantiles.

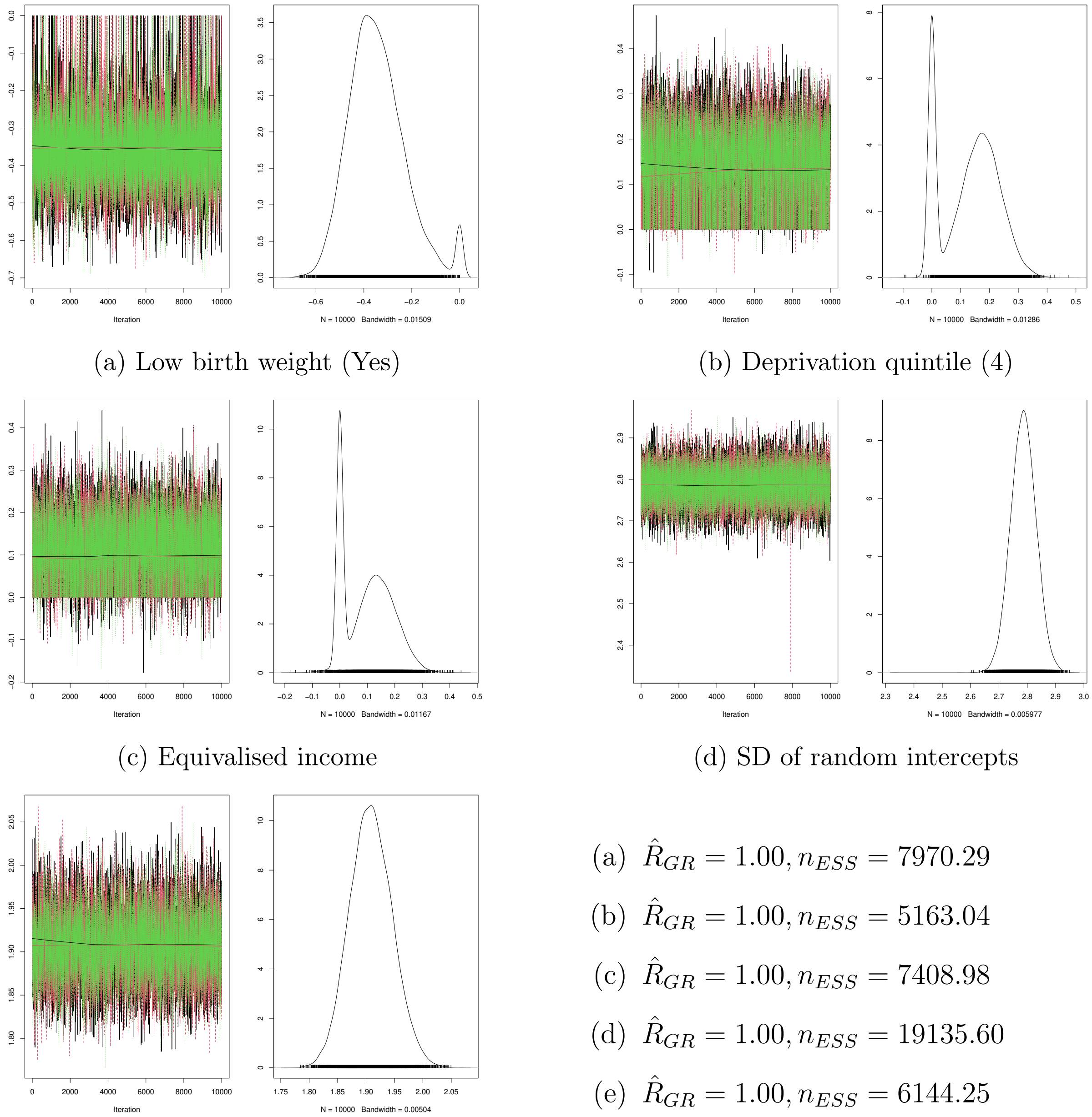
Considering only the continuous variable, ***equivalised income*** exhibited a positive effect with similar magnitudes across the two quantile models. This implies that a one standard deviation increase in equivalised income (£13097.16) is expected to result in an increase of 0.10 in centred raw weight for children in the lowest 10% and highest 90% of the centred raw weight distribution.

In relation to the categorical variables "***low birth weight***", this factor was included in both quantile models and demonstrated a negative effect. This suggests that children with a low birth weight were likely to show a decrease of 0.36 and 0.35 in centred raw weight values at the 0.10th and 0.90th quantiles, respectively, of the centred raw weight distribution, compared to those without low birth weight (the reference group). In the context of the ***deprivation quintile*** factor, only the "***quintile (4)***" category was selected across both quantile models, showing a positive effect. This implies that children in this specific deprivation category were expected to exhibit higher centred raw weight values at these two specific quantiles compared to those in the "***quintile (1) - least deprived***" category (the reference group). Specifically, children in deprivation quintile 4 have a centred raw weight that is around 0.14 - 0.15 kg higher than children in deprivation quintile 1. In contrast, other deprivation categories were not selected.

Convergence diagnostic

To assess the convergence of the samples generated by the MCMC simulation to target posterior distribution, I ran three independent MCMC chains and employed trace plot assessments alongside the Gelman-Rubin diagnostic ($\hat{R}_{GR}$) (Gelman & Rubin, 1992). Additionally, the effective sample size (ESS, n_{ESS}) was used to estimate the information loss due to autocorrelation in the chains and to determine the number of independent samples the MCMC chain is equivalent to (Gelman, 2014). The analysis here focused on selected parameters, particularly those related to linear fixed effects and random effects.

Figures 6.8 and 6.9 present the trace plots and density plots for selected parameters from



(a) Low birth weight (Yes)

(b) Deprivation quintile (4)

(c) Equivalised income

(d) SD of random intercepts

(e) SD of random slopes

(a) $\hat{R}_{GR} = 1.00, n_{ESS} = 7970.29$

(b) $\hat{R}_{GR} = 1.00, n_{ESS} = 5163.04$

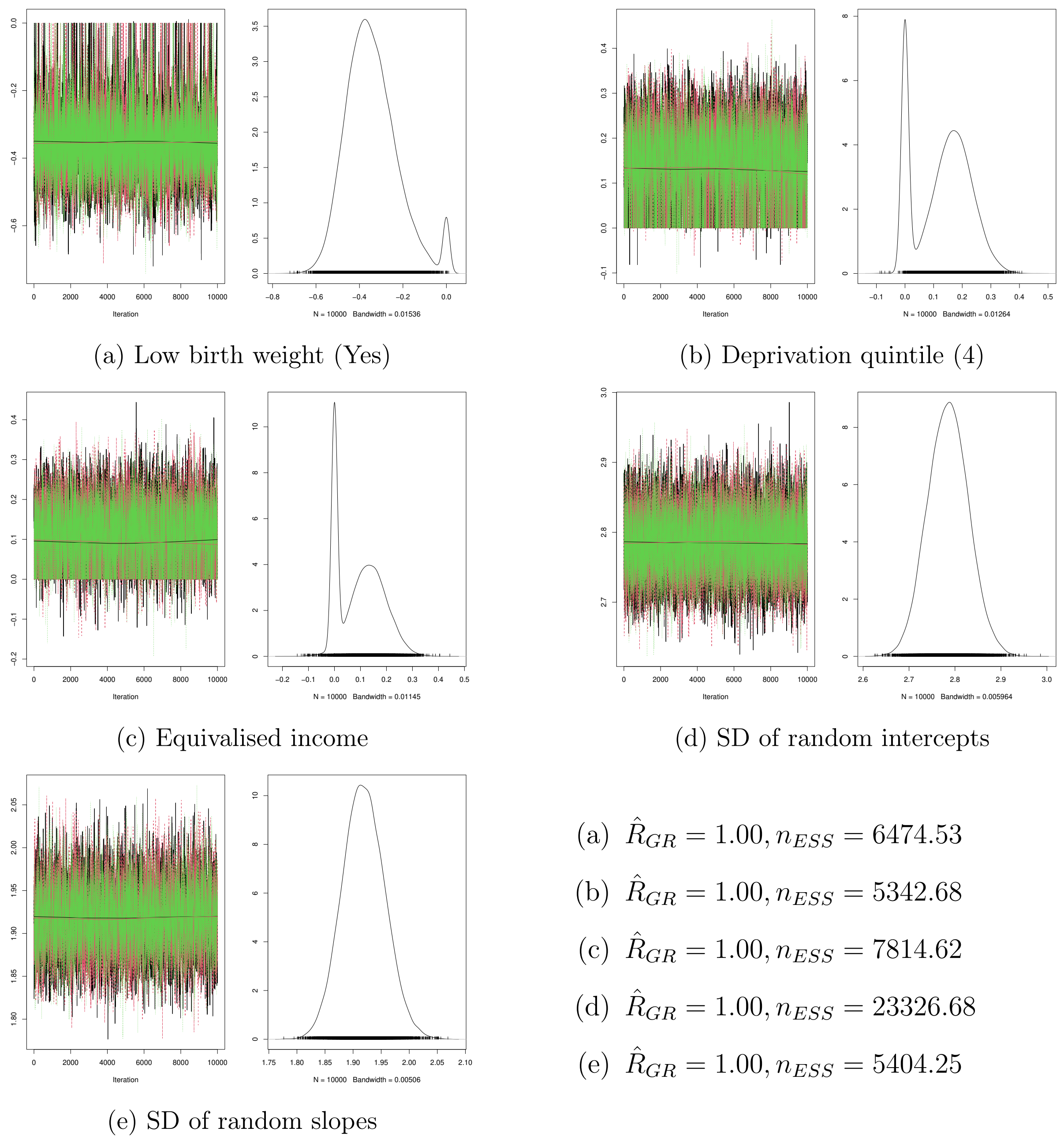
(c) $\hat{R}_{GR} = 1.00, n_{ESS} = 7408.98$

(d) $\hat{R}_{GR} = 1.00, n_{ESS} = 19135.60$

(e) $\hat{R}_{GR} = 1.00, n_{ESS} = 6144.25$

Figure 6.8: Trace plots of MCMC samples for each selected linear predictor and random effect from the 0.10th quantile model for the centred raw weight

the 0.10th and 0.90th quantile models for centred raw weight. The trace plots demonstrate convergence, with all chains fluctuating around a stable mean and showing no discernible trends over time. Specially, the plots of three linear fixed predictors (risk factors) reveal slight pseudo floor or ceiling effects approaching zero, indicating that the slab component of the spike-and-slab prior has predominantly influenced each selected fixed variable. Similarly, the trace plots for the standard deviations of the two random effects show samples fluctuating around a stable mean, with no discernible trend or drift. This suggests that these linear predictors and random effects are included in the final model. Additionally,



(a) Low birth weight (Yes)

(b) Deprivation quintile (4)

(c) Equivalised income

(d) SD of random intercepts

(e) SD of random slopes

(a) $\hat{R}_{GR} = 1.00, n_{ESS} = 6474.53$

(b) $\hat{R}_{GR} = 1.00, n_{ESS} = 5342.68$

(c) $\hat{R}_{GR} = 1.00, n_{ESS} = 7814.62$

(d) $\hat{R}_{GR} = 1.00, n_{ESS} = 23326.68$

(e) $\hat{R}_{GR} = 1.00, n_{ESS} = 5404.25$

Figure 6.9: Trace plots of MCMC samples for each selected linear predictor and random effect from the 0.90th quantile model for the centred raw weight

each $\hat{R}_{GR}$ value is equal to 1, indicating that the chains have likely converged to the same distribution and that the chains are well-mixed. Moreover, each n_{ESS} value exceeds 1,000, which is generally a positive sign. This suggests that the MCMC chains have generated a substantial number of effectively independent samples, implying that the sampling process is robust and the resulting estimates are likely to be reliable.

For the basis functions of age in both quantile models, trace plots are presented in Figures D.7 (the 0.10th quantile model) to D.8 (the 0.90th quantile model) in Appendix D, along with the Gelman-Rubin diagnostic and the effective sample size in Tables D.5 and D.6. Generally, the behaviour of the selected basis functions is similar to that the selected linear predictors, as mentioned previously. Meanwhile, the only basis function, **S(Age22)**, exhibits a substantial pseudo floor effect approaching zero in trace plots, indicating that this term is a non-selected predictor.

WAZ

Results for the posterior mean and median estimates of the basis functions related to age are presented in Table 6.11, where the centred WAZ was fitted. The results indicate that three basis functions were selected for two quantile models (the 0.10th and 0.90th quantiles), suggesting that these bases were important for predicting the centred WAZ. However, the selected basis functions varied between two quantiles. For the 0.10the quantile, **S(Age1)**, **S(Age3)** and **S(Age4)**, were selected, whereas **S(Age2)**, **S(Age3)** and **S(Age6)**, were chosen for the 0.90th quantile. This variation suggests that the smoothed age effects differ between these two quantiles, as illustrated in Figure 6.10. The figure shows a U-shaped in the early age range (ages 4 to 6 years), indicating that the centred WAZ initially decreases with age, reaches a minimum point, and then begins to increase with further ageing.

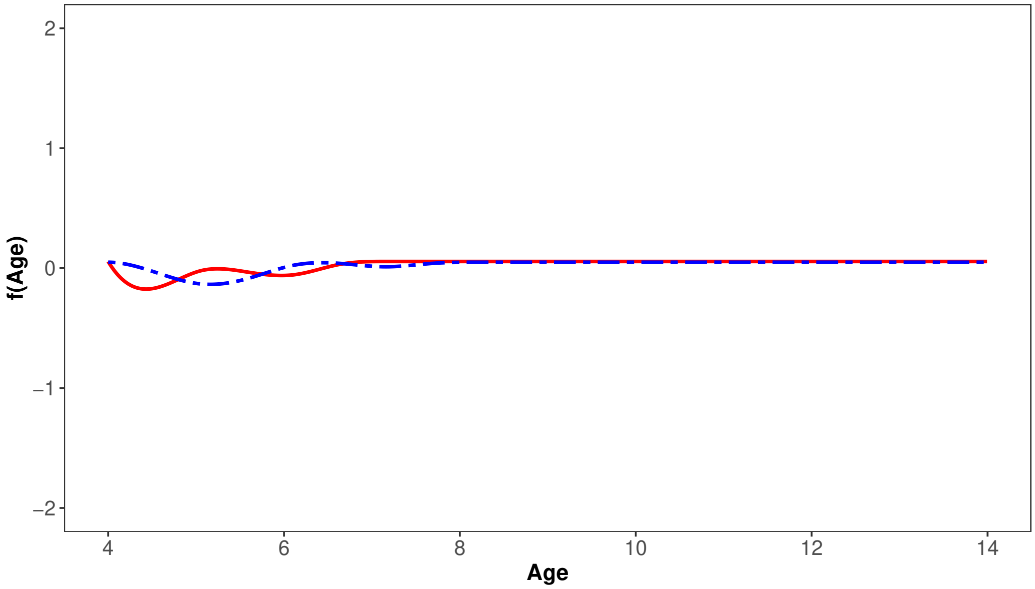


Figure 6.10: Smoothed age effect for the (centred) WAZ (*solid red line*: $\tau = 0.10$, and *two dashed blue line*: $\tau = 0.90$)

Regarding fixed effects (linear predictors), the ***low birth weight*** factor emerged as the only linear fixed effect selected for the 0.10th quantile model, with a small negative magnitude. This suggests that children with a low birth weight were likely to have a decrease of 0.01 in centred WAZ values at the 0.10th quantile of the centred WAZ distribution, compared to those without low birth weight (the reference group). In contrast, no linear fixed effects were selected for the 0.90th quantile model.

Both random intercepts and slopes were selected in these two specified quantile models, with the former exhibiting greater variability than the later. The standard deviations of these effects were relatively similar across the two quantile models. This implies that there was variability in individual intercepts and slopes in the centred WAZ across the two quantiles. The corresponding results are presented in Tables 6.12 to 6.13.

Table 6.11: Posterior mean, standard deviation (SD), and posterior median for fixed (basis terms) effects across two quantile models for the **WAZ**

Fixed effects	$\tau = 0.10$			$\tau = 0.90$		
(basis terms)	**Mean**	**SD**	**Median**	**Mean**	**SD**	**Median**
S(Age1)	-0.06	0.09	-0.05	0.01	0.05	-
S(Age2)	-0.01	0.06	-	-0.05	0.05	-0.04
S(Age3)	-0.03	0.06	-0.02	-0.04	0.05	-0.04
S(Age4)	-0.03	0.06	-0.02	-0.02	0.05	-
S(Age5)	0.00	0.01	-	0.00	0.01	-
S(Age6)	-0.02	0.02	-	-0.02	0.02	-0.01
S(Age7)	-0.01	0.02	-	-0.01	0.02	-
S(Age8)	-0.01	0.02	-	-0.01	0.02	-
S(Age9)	-0.01	0.01	-	-0.01	0.01	-
S(Age10)	0.00	0.01	-	0.00	0.01	-
S(Age11)	0.01	0.01	-	0.00	0.01	-
S(Age12)	0.02	0.02	-	0.00	0.01	-
S(Age13)	0.02	0.02	-	0.00	0.01	-
S(Age14)	0.01	0.02	-	0.00	0.01	-
S(Age15)	0.00	0.01	-	0.00	0.01	-
S(Age16)	0.00	0.01	-	0.00	0.01	-
S(Age17)	0.01	0.02	-	0.00	0.01	-
S(Age18)	0.01	0.03	-	0.00	0.02	-
S(Age19)	0.01	0.03	-	0.00	0.02	-
S(Age20)	0.01	0.03	-	0.00	0.02	-
S(Age21)	0.00	0.01	-	0.00	0.01	-
S(Age22)	0.00	0.01	-	0.00	0.01	-
S(Age23)	0.00	0.02	-	0.00	0.01	-

Table 6.12: Posterior mean, standard deviation (SD), and posterior median for both fixed (linear predictors) and random effects of the **0.10**th quantile model for the **WAZ**†

Fixed effects (linear predictors)	**Mean**	**SD**	**Median**
Sex (Male)	0.00	0.02	-
Low birth weight (Yes)	-0.04	0.07	**-0.01**
Ethnicity of a child (White)	0.00	0.02	-
Child's health in general (Good)	0.00	0.02	-
Child's health in general (Fair, Bad, Very Bad)	-0.01	0.02	-
Number of accidents or injuries of child	0.00	0.01	-
Child's birth order	0.00	0.09	-
Mother's marital status (Single)	0.00	0.01	-
Mother's marital status (Other)	0.00	0.01	-
Urban-rural classification (Other urban)	0.00	0.07	-
Urban-rural classification (Small, accessible towns)	0.00	0.02	-
Urban-rural classification (Small, remote towns)	0.00	0.02	-
Urban-rural classification (Accessible rural)	0.00	0.04	-
Urban-rural classification (Remote rural)	0.00	0.02	-
Household size	-0.01	0.09	-
Mother's age at first child's birth ($<$ 20 years old)	0.00	0.05	-
Mother's age at first child's birth ($\geq$ 30 years old)	0.00	0.09	-
Respondent's alcoholic drinks (Every day)	0.00	0.02	-
Respondent's alcoholic drinks (4 - 6 times a week)	0.00	0.04	-
Respondent's alcoholic drinks (2 - 3 times a week)	0.00	0.04	-
Respondent's alcoholic drinks (Once a week)	0.00	0.03	-
Respondent's alcoholic drinks (2 -3 times a month)	0.00	0.02	-
Respondent's alcoholic drinks (Once a month or less)	0.00	0.05	-
Respondent's alcoholic drinks (Not in the last year)	0.00	0.03	-
Respondent's current health (Very good)	0.00	0.01	-
Respondent's current health (Good)	0.00	0.02	-
Respondent's current health (Fair, Poor)	0.00	0.01	-
Smoking cigarettes while pregnant (Yes)	0.01	0.09	-
Drinking alcohol while pregnant ($\geq$ 3 - 4 times a week)	0.00	0.01	-
Drinking alcohol while pregnant (1 - 2 times a week)	0.00	0.02	-
Drinking alcohol while pregnant (2 - 3 times a month)	0.00	0.02	-
Drinking alcohol while pregnant ($<$ once a month)	0.00	0.02	-
Respondent's health problem(s) in a year (Yes)	0.00	0.02	-
Respondent's current job (No)	0.00	0.01	-
Deprivation quintile (2)	0.00	0.02	-
Deprivation quintile (3)	0.00	0.01	-
Deprivation quintile (4)	0.01	0.02	-
Deprivation quintile (5)	0.00	0.03	-
Equivalised income	0.00	0.09	-
Standard deviation (Random effects)			
$\hat{\sigma}_0$ (Intercept Age)	0.85	0.05	**0.85**
$\hat{\sigma}_1$ (Slope of Age)	0.20	0.01	**0.20**

† The reference categories are given in Tables 2.9 to 2.11.

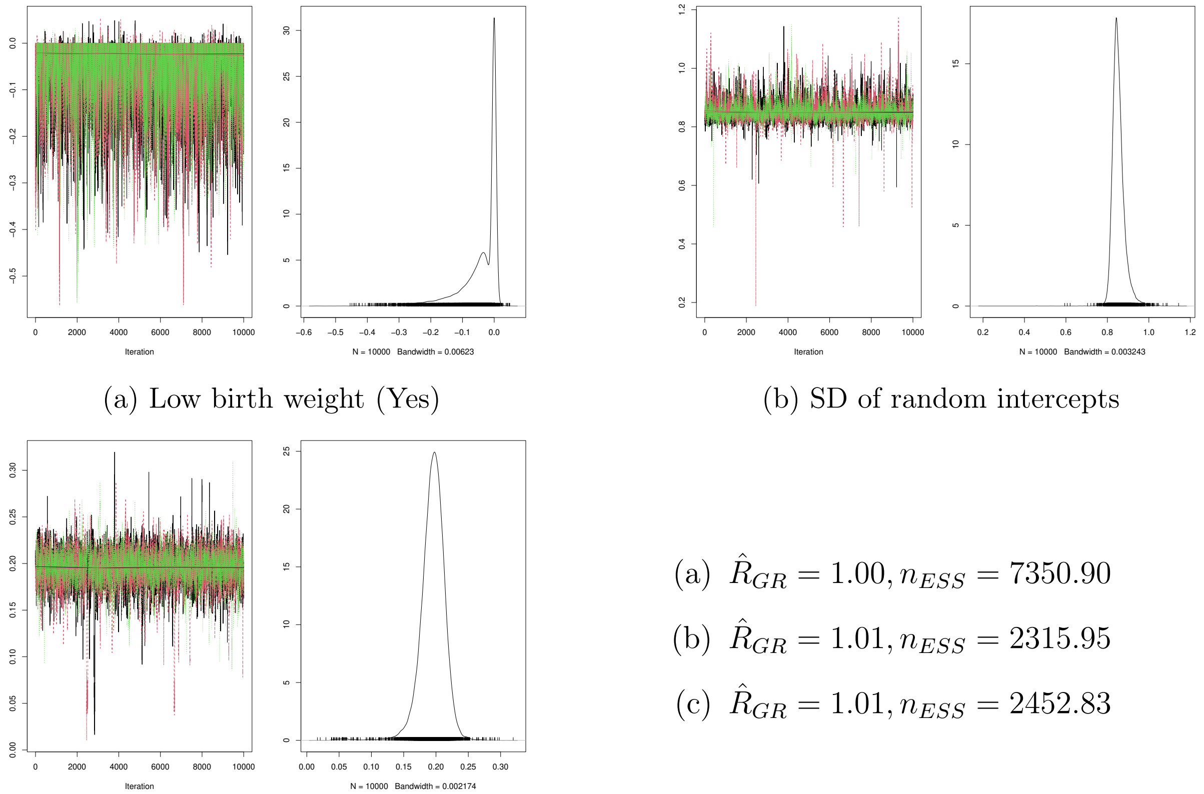
Table 6.13: Posterior mean, standard deviation (SD), and posterior median for both fixed (linear predictors) and random effects of the **0.90**th quantile model for the **WAZ**†

Fixed effects (linear predictors)	**Mean**	**SD**	**Median**
Sex (Male)	0.00	0.02	-
Low birth weight (Yes)	-0.03	0.05	-
Ethnicity of a child (White)	0.00	0.02	-
Child's health in general (Good)	0.00	0.02	-
Child's health in general (Fair, Bad, Very Bad)	-0.01	0.02	-
Number of accidents or injuries of child	0.00	0.01	-
Child's birth order	0.00	0.11	-
Mother's marital status (Single)	0.00	0.00	-
Mother's marital status (Other)	0.00	0.01	-
Urban-rural classification (Other urban)	0.00	0.06	-
Urban-rural classification (Small, accessible towns)	0.00	0.01	-
Urban-rural classification (Small, remote towns)	0.00	0.02	-
Urban-rural classification (Accessible rural)	0.00	0.04	-
Urban-rural classification (Remote rural)	0.00	0.02	-
Household size	0.00	0.09	-
Mother's age at first child's birth ($<$ 20 years old)	0.01	0.05	-
Mother's age at first child's birth ($\geq$ 30 years old)	-0.01	0.10	-
Respondent's alcoholic drinks (Every day)	0.00	0.02	-
Respondent's alcoholic drinks (4 - 6 times a week)	0.00	0.03	-
Respondent's alcoholic drinks (2 - 3 times a week)	0.00	0.05	-
Respondent's alcoholic drinks (Once a week)	0.00	0.03	-
Respondent's alcoholic drinks (2 -3 times a month)	0.00	0.02	-
Respondent's alcoholic drinks (Once a month or less)	0.00	0.04	-
Respondent's alcoholic drinks (Not in the last year)	0.01	0.03	-
Respondent's current health (Very good)	0.00	0.01	-
Respondent's current health (Good)	0.00	0.01	-
Respondent's current health (Fair, Poor)	0.00	0.02	-
Smoking cigarettes while pregnant (Yes)	0.01	0.08	-
Drinking alcohol while pregnant ($\geq$ 3 - 4 times a week)	0.00	0.01	-
Drinking alcohol while pregnant (1 - 2 times a week)	0.00	0.02	-
Drinking alcohol while pregnant (2 - 3 times a month)	0.00	0.02	-
Drinking alcohol while pregnant ($<$ once a month)	0.00	0.02	-
Respondent's health problem(s) in a year (Yes)	0.00	0.02	-
Respondent's current job (No)	0.00	0.01	-
Deprivation quintile (2)	0.00	0.01	-
Deprivation quintile (3)	0.00	0.01	-
Deprivation quintile (4)	0.01	0.03	-
Deprivation quintile (5)	0.00	0.02	-
Equivalised income	0.00	0.09	-
Standard deviation (Random effects)			
$\hat{\sigma}_0$ (Intercept Age)	0.86	0.06	**0.85**
$\hat{\sigma}_1$ (Slope of Age)	0.20	0.01	**0.20**

† The reference categories are given in Tables 2.9 to 2.11.

Convergence diagnostic

Similar to the convergence diagnostics for the raw weight case, three independent MCMC chains were run for the WAZ case. Subsequently, trace plot assessments and the Gelman-Rubin diagnostic ($\hat{R}_{GR}$) were employed to assess convergence in the MCMC simulation towards the target posterior distribution. Morevover, to determine how many independent samples the MCMC chain is equivalent to, the effective sample size or ESS (n_{ESS}), was calculated. Here, I present convergence diagnostics only for selected parameters, including both linear fixed effects (predictors) and random effects. For the basis functions of age, the relevant diagnostics were presented in Appendix D.



(a) Low birth weight (Yes)

(b) SD of random intercepts

(c) SD of random slopes

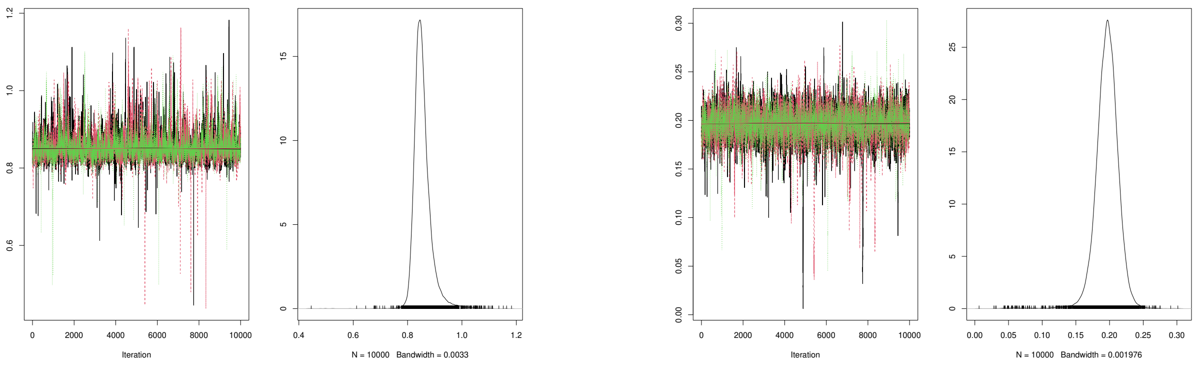
(a) $\hat{R}_{GR} = 1.00, n_{ESS} = 7350.90$

(b) $\hat{R}_{GR} = 1.01, n_{ESS} = 2315.95$

(c) $\hat{R}_{GR} = 1.01, n_{ESS} = 2452.83$

Figure 6.11: Trace plots of MCMC samples for each selected linear predictor and random effect from the 0.10th quantile model for the centred WAZ

Figures 6.11 and 6.12 present trace plots of selected parameters for linear fixed effect and random effects, along with density plots, the Gelman-Rubin diagnostic ($\hat{R}_{GR}$), and the effective sample size (n_{ESS}), for the 0.10th and 0.90th quantile models, respectively. Each trace plot shows that no discernible trends over time are observed, and the samples for all chains fluctuated around a stable mean, generally indicating convergence. Additionally, each $\hat{R}_{GR}$ is less than 1.1, confirming that the MCMC chains have reached a stationary distribution and are well-mixed. However, the plot of a selected linear fixed predictor, ***low***



(a) SD of random intercepts
$\hat{R}_{GR} = 1.00, n_{ESS} = 2163.08$

(b) SD of random slopes
$\hat{R}_{GR} = 1.01, n_{ESS} = 3139.01$

Figure 6.12: Trace plots of MCMC samples for each selected linear predictor and random effect from the 0.90th quantile model for the centred WAZ

birth weight, for the 0.10th quantile model shows a pseudo ceiling effect approaching zero. This is characterised by a high peak in the spike component and a wide spread in the slab component in of the density plot. Notably, the slab component of the spike-and-slab prior has predominantly influenced this predictor, suggesting that more coefficients were retained (not shrunk to zero) compared to those that were. Furthermore, the chains have a relatively large number of independent samples, as each n_{ESS} exceeds 1,000, indicating that the generated samples are sufficient for accurate and reliable estimates.

6.2.3 Summary

When analysing the dataset of children ages 4 to 14 years for both raw weight and WAZ using the BSGSSMQR method, the age effect exhibited a clear, non-linear pattern. For raw weight, the patterns at the 0.10th and 0.90th quantiles were relatively similar as age increased with fluctuations. In contrast, for WAZ, the patterns differed at younger ages but became more similar as age progressed for both quantiles. Despite the variations in results between raw weight and WAZ, low birth weight was a significant factor influencing both child growth measurements at the 0.10th and 0.90th quantiles, although it was not as influential at the 0.90th quantile for the WAZ. When focusing exclusively on raw weight, other risk factors such as deprivation quintile and equivalised income were significantly impacting child growth, regardless of whether children fell into the upper (the 0.90th quantiles) or lower (the 0.10th quantile) distributions of raw weight. Additionally, the fitted method revealed strong variations in both the initial baseline (intercepts) and the rate of growth (slopes) among children. For WAZ, low birth weight was only associated with the 0.10th quantile of this measurement. Furthermore, the fitted method indicated substantial variations in the initial baseline (intercepts), but relatively small variations in the rate of growth (slopes) for both quantiles.

6.3 Comparison findings between AQMM and BSGSSMQR

In this section, the results from both AQMM and BSGSSMQR are presented in two parts. The first part focuses on demonstrating the three fitted growth curves at the 0.10, 0.05th, and 0.90th quantiles for both males and females, as estimated by these two approaches. To ensure a fair comparison, the fitting processes for both approaches were made similar in terms of random effects specifications. This approach was implemented because BSGSSMQR requires the inclusion of all possible random effects and automatically identifies these effects through a Bayesian variable selection method. It was observed that BSGSSMQR identified both random intercepts and random slopes in the models. Consequently, both random intercepts and random slopes were also included in AQMM. Therefore, in this part, the growth quantile curves are presented for comparison rather than to identify the best-fitted growth curves. The second part summarises the significant or selected risk factors associated with two weight measurements (i.e. raw weight and WAZ), as identified by both the bootstrap method used in AQMM (Section 6.1) and BSGSSMQR (Section 6.2), in Tables 6.14 to 6.15. To ensure a fair comparison, some standardised coefficients from BSGSSMQR, such as ***low birth weight***, will be converted to their original scale for comparison purposes using the follow formula:

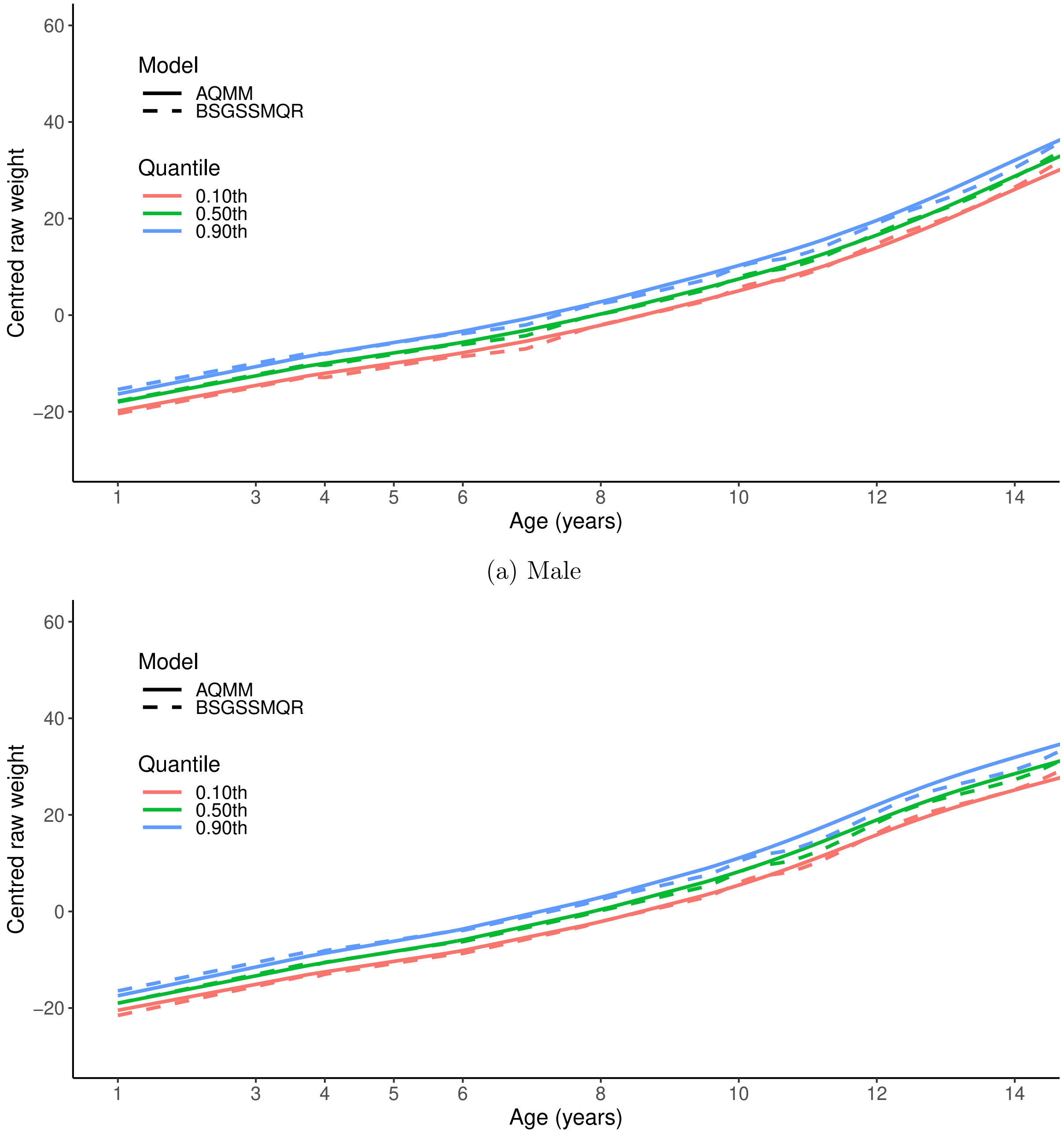
$$\beta_p = \frac{\beta'_p}{\sigma_p},$$

where β_p is the original scale coefficient, β'_p is the standardised scale coefficient, and σ_p is the standard deviation of the p-th predictor.

6.3.1 Quantile growth curves

Figures 6.13 and 6.14 present three quantile growth curves across ages ranging from 1 to 14 years for centred raw weight and centred WAZ, estimated by AQMM and BSGSSMQR using the entire GUS dataset, with both approaches incorporating random intercepts and random slopes. Generally, the growth curves produced by both methods exhibit similar trends. However, the curves produced by AQMM appeared to be smoother compared to those from BSGSSMQR.

Regarding the centred raw weight, growth curves from both models are relatively close from ages 1 to 4 year. After this period, modest departures are observed, particularly during the period of puberty (ages 8 and 14 for females and 9 and 14 for males). Additionally, BSGSSMQR provided slightly fluctuating growth curves, while AQMM showed strong, smooth trends throughout the ages. In terms of the centred WAZ, modest departures were

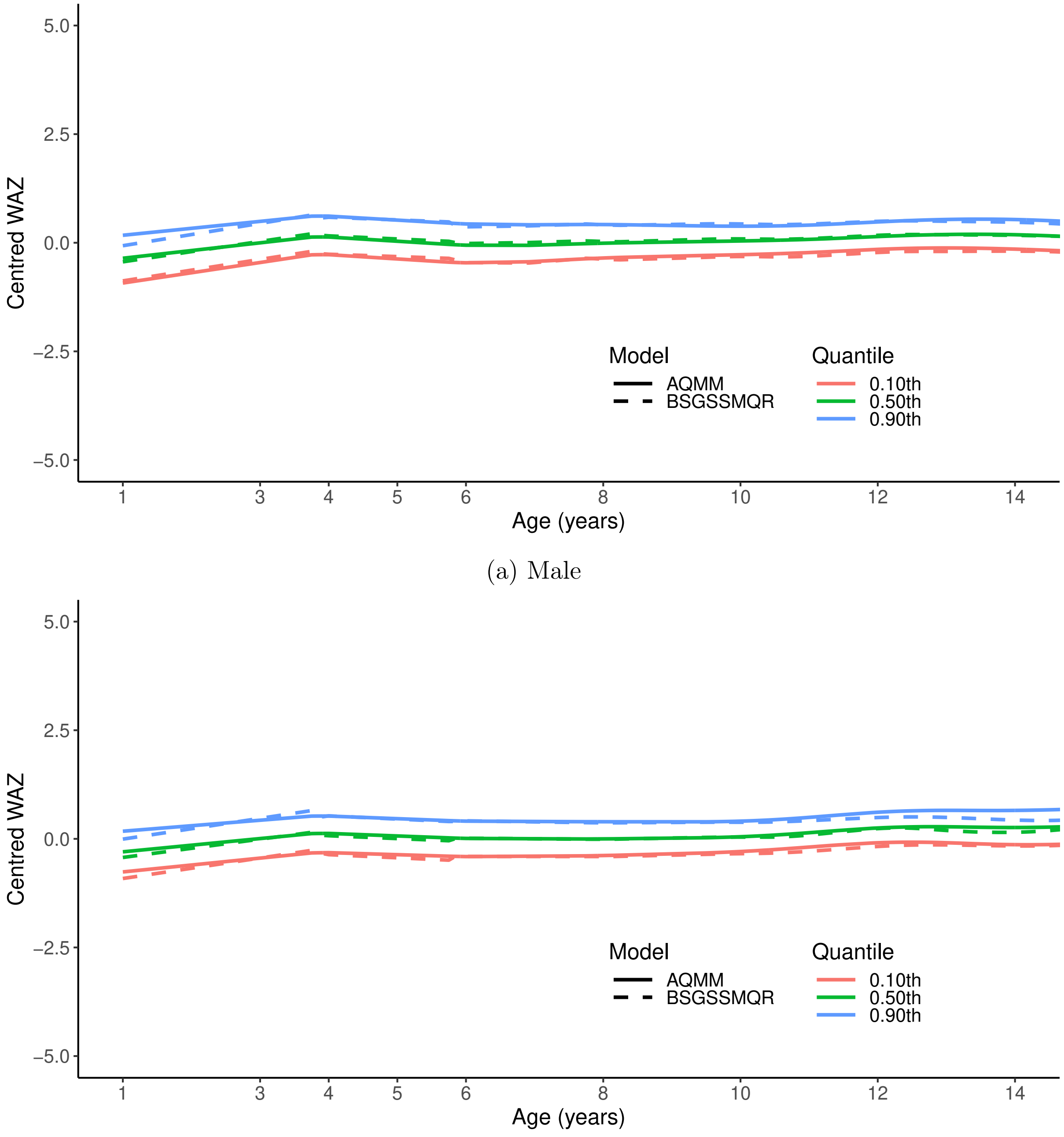


(a) Male

(b) Female

Figure 6.13: Comparison of AQMM and BSGSSMQR growth curves for the centred raw weight

observed, particularly at the beginning of the age rage. Growth curves from approximately ages 4 to 10 years are relatively close. However, for ages 10 to 14 years (older children), modest departures were noted, especially in BSGSSMQR for females. The departures in quantile growth curves for both growth measurements between the two methods may be attributed to the varying splines used in each approach. Additionally, it was observed that both methods similarly provided growth curves that did not appear to capture the GUS data well when including both random intercepts and slopes, as seen in Section 6.1.



(a) Male

(b) Female

Figure 6.14: Comparison of AQMM and BSGSSMQR growth curves for the centred WAZ

As the quantile growth curves from both approaches were relatively comparable, I validated this by calculating the percentage of points above the 0.10th, 0.50th, and 0.90th quantiles, finding that the results were similar to those of AQMM, as the percentages of observations above each percentile did not align closely with the expected values. This suggests that BSGSSMQR exhibits similar behaviour to AQMM when both random intercepts and random slopes were selected, indicating that neither model fits the GUS data well. In context of BSGSSMQR, this could be attributed to the method used (see the last term of equation 5.14 in Section 5.8.2), which the regularisation parameter (λ_3) is restricted to 1, allowing the data to guide the modelling process and avoiding the need for additional hyperparameter tuning. While this approach has yielded good results in simulation studies, it does not guarantee optimal outcomes and may ultimately lead to model misspecification in the context of random effect selection when applied to the real data. In the feature work, this parameter should be allowed to vary or be estimated rather than remaining fixed at 1.

6.3.2 Identification of risk factors

Table 6.14 presents a summary of significant or selected risk factors associated with centred raw weight at two different quantiles ($\tau = 0.10$ and $\tau = 0.90$), as fitted from the dataset of children aged 4 to 14 years. The comparison highlights the differences in how the AQMM and BSGSSMQR approaches identify and estimate these risk factors. It can be observed that the bootstrap method implemented in AQMM identified significant risk factors that varied across the two quantiles. For example, ***low birth weight*** has different point estimates at both quantiles. Additionally, some factors, such as ***equivalised income*** are signficant at $\tau = 0.10$ but not at $\tau = 0.90$. In contrast, BSGSSMQR identified the same risk factors, which remained consistent across these specified quantiles in terms of both magnitudes and directions.

Table 6.14: Significant (selected) risk factors associated with the centred raw weight along with their point estimates: comparing AQMM and BSGSSMQR

Factor	$\tau = 0.10$		$\tau = 0.90$	
	AQMM	**BSGSSMQR**	**AQMM**	**BSGSSMQR**
Low birth weight (Yes)	-1.24	-1.64	-1.72	-1.59
Deprivation quintile (4)		0.15		0.14
Equivalised income	0.21	0.10		0.10

Regarding WAZ (see Table 6.15), significant risk factors identified by AQMM varied noticeably across the two quartiles. In contrast, only one risk factor, ***low birth weight***, was selected by BSGSSMQR, and this was only at the 0.10th quantile. It can be observed that AQMM provided a consistent magnitude and direction of significant risk factors that appeared at both quantiles, except for ***low birth weight***

Table 6.15: Significant (selected) risk factors associated with the centred WAZ along with their point estimates : comparing AQMM and BSGSSMQR

Factor	$\tau = 0.10$		$\tau = 0.90$	
	AQMM	**BSGSSMQR**	**AQMM**	**BSGSSMQR**
Low birth weight (Yes)	-0.62	-0.05	-0.55	
Respondent's current health (Very good)	0.03		0.03	
Respondent's current health (Good)			0.05	
Smoking cigarettes while pregnant (Yes)	0.14		0.13	
Deprivation quintile (2)			0.05	
Deprivation quintile (4)	0.13		0.12	
Equivalised income	0.03		0.02	

6.4 Chapter summary

In this chapter, two different framework methods, such as frequentist and Bayesian approaches, were applied to the GUS data, which consists of longitudinal child growth data (LCGD) from Scotland. The former, a method known as AQMM, facilitates the analysis of this data within a mixed-effects model framework. This method enables the easy fitting of non-linear trajectories through the mixed effects representation of smoothing splines. Additionally, it involves identifying risk factors by incorporating potential variables as fixed effects (non-linear predictors and linear predictors) in the quantile mixed model and identifying these through the bootstrap method. I applied the AQMM approach with cubic P-splines in two key aspects: constructing reference growth charts and identifying risk factors associated with child growth measurements. In the first aspect, it was observed that the reference growth curves fitted by this approach, using only random intercepts, were well-aligned with both GUS datasets (the entire dataset and the dataset of children aged 4 to 14 years) and varied by sex and growth measurement scale. It is evident that raw weight and WAZ exhibited non-linear curves, while raw height and HAZ displayed curves that are more linear. Some extreme quantile curves, such as 0.04th and 0.90th quantiles, exhibited non-parallel trends with other quantiles in WAZ, raw height, and HAZ. In the second aspect, the risk factors were evaluated in relation to child physical growth measurements using the dataset of children aged 4 to 14 years, representing the school-age children and young people in primary or secondary school in Scotland. The analysis indicates that significant risk factors varied across the two quantiles and different child growth measurements.

The latter approach, the Bayesian sparse group LASSO-mixed quantile regression model (BSGSSMQR), was also applied to analyse the same dataset as the previous method, with a focus on the simultaneous selection of both fixed and random effects. This method facilitates modelling of non-linear relationships between child growth measurements and predictors using cubic B-splines with quantile knots. By utilising a Bayesian LASSO-type approach with spike and slab priors on quantile regression coefficients, it enables users to select relatively large B-spline bases for constructing the smoothing function. Additionally, it allows for the automatic removal of some bases that lack data support by shrinking the coefficients of these bases towards zero. This methodology is also employed for the selection of linear predictors. Furthermore, a Bayesian LASSO-type approach is used in the selection of random effects, shrinking their standard deviations towards zero. The analysis reveals that age exhibited a non-linear effect on raw weight across the lower and upper quantiles. Selected risk factors (fixed effects) slightly differ across two weight measurement scales, but exhibit consistency in risk factor selection across the three specified quantiles, as mentioned previously. Moreover, the results indicate that there was variability in individual linear trends, such as intercepts and slopes, which were selected in the three specified quantile models. However, it was observed that BSGSSMQR, with selected random intercepts and random slopes, did not produced quantile growth curves that fit to the GUS data well. This may be due to the penalty term used for random effect selection, which leads to the misspecification of random effects. Therefore, this issue should be addressed in future work.

In comparison, BSGSSMQR exhibits similar behaviour to AQMM in terms of fitting quantile growth curves using the GUS data. Furthermore, BSGSSMQR demonstrates greater parsimony in variable selection than the bootstrap method implemented in AQMM. While AQMM identified risk factors associated with child growth measurements that vary across quantiles, BSGSSMQR resulted in more consistent variable selection across quantiles.

Chapter 7

Conclusions

The primary focus of this thesis is to thoroughly examine longitudinal child growth data using statistical models, with a particular emphasis on quantile regression. The aim is to provide comprehensive insights into physical growth and development of children. This comprehensive understanding can subsequently be used to inform policies focused on prevention, promotion, and support, ensuring children experience healthy growth. The scope of the research extended beyond mere data description; it encompassed capturing the characteristics of the data to accurately reflect child growth and also evaluating potential risk factors affecting growth across diverse children's groups.

As discussed in Chapter 2, longitudinal child growth data (LCGD) offer unique advantages over cross-sectional data, notably the ability to track trends and changes in individual child's growth over time. Tools such as child growth charts or models, derived from this type of data, addresses fundamental scientific questions in child growth studies. However, these advantages come with complexities, including varying non-linear growth patterns, heterogeneous variability, and autocorrelation within individuals. To account for the latter two features, mixed-effects models have been a popular statistical approach, while spline methods typically model the first feature. This integration forms what is known as flexible mixed-effect models, as discussed in Chapter 3. However, these models primarily focus on modelling the central location of child growth measurements, such as the mean, and may not adequately describe growth changes across the diversity of children's groups.

To address this limitation, this thesis outlines a quantile regression based on a likelihood-based method via the asymmetric Laplace (AL) distribution, combined with random effects and splines. This method estimates the conditional quantiles of the dependent variable, representing several locations of child growth measurements. However, this approach has crucial limitations in inference using the bootstrap method, potentially leading to inappropriate results in fixed effects selection. Moreover, it lacks a specific method for selecting random effects. Consequently, this study aimed to overcome these challenges by developing Bayesian variable selection methods within the quantile mixed-effects models.

7.1 Modelling the longitudinal child growth data with flexible quantile models

In Chapter 4, two flexible quantile models were discussed for modelling longitudinal child growth data. The first is an existing method known as the quantile specific autoregressive model (QSAM), which is based on the regression spline approach. This model utilises three components to describe data: a non-parametric function, a first-order autoregressive model, and a linear predictor function. It has been applied to longitudinal child growth data for constructing reference growth curves. The second model is the additive quantile mixed model (AQMM), which incorporates elements of both the additive quantile model and the mixed-effect model. The former is employed to capture the non-linear relationship, while the latter accounts for the dependency between observations taken from the same individual at different times. In this chapter, several simulation studies were conducted to explore the performances of these two methods when implemented with the context of longitudinal child growth data. In terms of the modelling process, AQMM allows users to employ any spline method to model the non-linear relationship between child growth measurements and age, accommodating a relatively large number of bases. In contrast, QSAM requires optimisation in terms of the number of bases used. According to the simulation results, AQMM proved to be the superior model for estimating the quantiles of child growth measurements, particularly when the simulated data exhibited autocorrelated growth outcomes, in scenarios of homogeneous and heterogeneous variability, and with both balanced and unbalanced data. Moreover, when simulated data exhibited variability in individual linear trends (intercepts and temporal slopes), AQMM maintained its predictive performance and acceptable computational efficiency. This observation still occurred in scenarios that simulated data had additional features such as the between-individual differences in intra-individual variation and autocorrelation. However, it should be noted that these variations should not be large. This is because, in AQMM, the error variance and the variance of the random effects are treated as homogeneous across individual groups or levels of covariates. Overall, AQMM is a suitable choice for implementation with longitudinal child growth data and is therefore used as the model to fit real data in Chapter 6.

7.2 Variable selection methods in quantile regression

Chapter 5 focused on variable selection methods in the context of quantile regression, aiming to address the limitations of the AQMM approach in this area. Two main aspects were considered: variable selection in fixed effects within the quantile model and variable selection including both fixed and random effects within the quantile mixed model. The first aspect specifically aimed to identify a more effective method capable of selecting variables in quantile models containing only fixed effects. This method would then be extended for variable selection in the context of the quantile mixed model in the second aspect. In the first aspect, two novel Bayesian variable selection methods were proposed. These methods were developed by integrating Bayesian LASSO-type methods with spike-and-slab priors on quantile regression coefficients, similar to the work of Xu and Ghosh (2015). The first method, named as BGLSSQR, is based on the Bayesian group LASSO, which facilitates the selection of variables exhibiting group structures. The second method, BSGSSQR, extends the first, allowing for the selection of a group of predictors and individual within a group using Bayesian sparse group LASSO. The performance of each method was assessed through simulation studies and compared with existing methods. The results indicated that these two proposed methods generally yielded biased estimators, similar to other regularisation methods, whether frequentist or Bayesian. Notably, they were superior to both existing frequentist and Bayesian approaches and maintained their predictive performance and model selection accuracy in both simple and complex scenarios. This included situations with high sparsity, either independence or correlation among predictors, and a variety of variable groups and predictors. However, BSGSSQR offers an advantage over BGLSSQR as it also provides the ability to select individual levels within a group variable.

Therefore, BSGSSQR was exclusively chosen for incorporation into the quantile mixed model as the proposed method for the second aspect mentioned earlier. This proposed method, BSGSSMQR, incorporates elements of BSGSSQR and introduces a decomposition for the covariance matrix of random effects, enabling the simultaneous selection of both fixed and random effects. Unlike AQMM, BSGSSMQR estimates the smooth function through spline coefficients estimated in a penalised form. In AQMM, however, the smooth function is estimated by treating a set of spline coefficients as random variables, which are then predicted in a linear mixed model. In this aspect, a simulation study was also conducted to assess the proposed method in comparison to AQMM. Regarding the fixed effect selection, the proposed method demonstrates good performance compared with the bootstrap method implemented in AQMM. Moreover, it facilitates the simultaneous selection of random effects.

7.3 Analysis of longitudinal child growth data in Scotland

In Chapter 6, the analysis of real data is presented. The longitudinal child growth data from the Growing up in Scotland (GUS) study were employed for this purpose. Initially, the AQMM approach was applied with two objectives: constructing the reference growth charts and identifying risk factors associated with child growth measurements. This approach was chosen based on results from Chapter 4, which suggested that this approach is suitable and flexible for analysing these data. In the first objective, AQMM demonstrated that reference growth curves exhibited non-linear patterns in both males and females, particularly in measurements relating to weight. In terms of identifying risk factors associated with child growth measurements, the bootstrap method was utilised. The results indicated that significant risk factors varied across the lower (the 0.10th), and upper (the 0.90th) quantiles. Some risk factors were significant at lower quantiles but not at upper ones (or vice versa), while others showed significance across these two quantiles with either the same or differing magnitudes of effect. Furthermore, the results indicate that there was variation in individual linear trends (intercepts and slopes) across the two specified quantiles, especially in both the raw scale of weight and height. However, the variation in slopes on both the z-score scale of weight and height was small.

Secondly, BSGSSMQR was applied to identify risk factors associated with child growth measurements only relating to weight, and to account for and indicate necessary variation at the individual level, as such as in intercept and slopes. For the raw weight, age was non-linearly associated with this growth measurement across two specified quantiles. Its effect appeared consistent throughout the age range. Three risk factors, ***Low birth weight***, ***Deprivation quintile***, and ***Equivalised income***, were selected consistent across two specified quantile models. Regarding accounting for variation at the individual-level, both random intercepts and random slopes were selected in the model. This indicates that there was individual variability in both intercept and growth velocity around the population prediction at the lower, and upper ends of the raw weight distribution. It implies that the growth of individual children deviates from the population by a shift in the intercept, and that they also grow faster or slower than the population across quantiles. Meanwhile, for the z-score scale of weight or WAZ, age was non-linearly associated with this growth measurement at two quantile models. Only ***Low birth weight*** was selected as the risk factor at the 0.10th quantile. There was individual variability in both intercept and growth velocity around the population prediction for this measurement as well, but variability seemed to be small across two quantiles of the WAZ distribution. However, BSGSSMQR appears to select unnecessary individual variability, specifically in the form of random slopes, leading to poor model fit. This may explain why the quantile growth curves did not fit the GUS data well.

In the comparison between AQMM and BSGSSMQR, it was found that the latter is more parsimonious than the former, resulting in a model with the fewest possible variables or parameters. However, most of significant risk factors selected in BSGSSMQR are a subset of those in AQMM. This evidence is akin to the work of Xu and Ghosh (2015), as BSGSSMQR was developed in a similar manner. Another point to note is that while the bootstrap method implemented in AQMM cannot assess the significance of the basis functions for age, this can be evaluated using BSGSSMQR.

7.4 Implications of empirical findings for policy in Scotland

The findings presented in this thesis, particularly those from Chapter 6, offer opportunities to identify actionable steps, enhance current effects and programmes, and address gaps in policy decisions for prioritising interventions in Scotland, with a focus on reducing child obesity in school-age children and young people in primary or secondary education. The strength of my research findings lies in analysing risks factors affecting child growth in school-age children and young people in primary or secondary education in Scotland. The evidence presented here would be beneficial for the Scottish government and public health agencies. For example, three risk factors – ***Low birth weight***, ***Deprivation quintile*** and ***Equivalised income*** – were found to be significantly associated with the 90th percentile of the weight growth distribution, indicating that children in the highest 10% of the weight growth distribution are likely at risk of obesity. However, it is important to note that child obesity is typically defined using specific criteria such as body mass index (BMI) percentiles, rather than directly from weight growth distribution percentiles. Nonetheless, these findings can provide valuable preliminary information in this context.

Firstly, it is found that children at the higher end of the weight distribution (those in the 90th percentile) who were born with low birth weight tend to have lower weights compared to their peers who were not born with low birth weight. This suggests that low birth weight has a particularly strong association with reduced weight at the higher end of the distribution. While these children may still be heavier than others, their weight is less than it would be if they had not been born with low birth weight. In terms of policy-making, particularly for reducing child overweight or obesity, the Scottish government should implement comprehensive policies that address various aspects of this issue. First, nutrition should be prioritised by implementing programmes to educate parents and communities on appropriate meal for low-birth-weight children. Second, this group of children should have their growth monitored and evaluated frequently.

Secondly, a deprivation quintile categorizes a population into five groups based on the level of deprivation they experience, which includes a lack of resources or access to basic needs such as income, employment, education, healthcare, safety, housing, and services. This measure serves as a poverty risk factor associated with child growth, reflecting the impact of childhood poverty. In the context of the GUS study, the first quintile (Deprivation quintile 1) represents the least deprived 20% of the population and the fifth quintile (Deprivation quintile 5) represents the most deprived 20%. This implies that children whose parents fall into higher quintiles of deprivation are more likely to face childhood poverty than those in lower quintiles. In this analysis, it is evident that children living in the 4th quintile of deprivation had a higher 0.90th quantile of weight compared to those in other quintiles. This suggests that living in more deprived areas may contribute to increased weight in children, raising the likelihood of them becoming overweight or obese. To reduce the risk of childhood obesity, the Scotland government or relevant authorities should consider strategies aimed at reducing childhood poverty, such as implementing policies specifically targeting families living in high level of deprivation. Fortunately, in March 2022, the Scottish Government launched the "*Best Start, Bright Futures: tackling child poverty delivery plan 2022 to 2026*" (APS Group Scotland, 2022), which includes relevant policies to reduce in this regard. The evidence from this thesis supports this policy and highlights the importance of focusing on children living in deprived areas, particularly those in deprivation quintile 4.

Thirdly, equivalised income, a measure that adjusts total household income for household size and composition using an equivalence scale, is another poverty risk factor associated with the higher end of the weight distribution. This adjustment accounts for the specific needs of household members, allowing for comparisons of living standards across households of different sizes and compositions, with higher equivalised income generally indicating a higher standard of living. In this thesis, the results reveal that as equivalised income increases, the weight of children at the 90th percentile of the weight distribution also tends to increase. This suggests that children in households with higher equivalised income are likely to be heavier at the upper end of the weight distribution. While higher income is typically associated with better access to healthy food and healthcare, several factors could explain this finding. For instance, children in higher-income families might have greater access to a wide variety of foods, including those that are high in calories and less nutritious. Additionally, these families might lead more sedentary lifestyles. This insight could inform policies aimed at weight management and nutritional education.

However, it is crucial to identify which equivalised income brackets should be targeted to ensure the effectiveness of these interventions. For example, programmes could focus on families within equivalised incomes quintiles where children tend to have higher weights, promoting balanced diets and healthy lifestyles and prevent excessive weight gain.

7.5 Limitations and future work

There are three main limitations to this study: the data, the analytical context and the proposed methods. Regarding the data, there were relatively few repeated data points in the GUS dataset. The maximum number of points was seven for weight measurement and five for height measurement. This limitation arises from the design of the Growing Up in Scotland study, which aimed to collect these growth measurements at specific sweeps, not at every sweep. As a result, this led to broadly spaced age intervals, particularly between 10 months and 4 years, which might impact the adequate representation of critical development stages.

In Chapter 6, three shortcomings are noted in the analytical context. Firstly, the potential risk factors were selected based on alignment with a single framework for studying child growth measurements. This approach may have overlooked other significant risk factors. For future work, other frameworks should be considered to identify additional important risk factors. Additionally, some risk factors, such as equivalised income, may be challenging to modify directly in practice. Secondly, the analysis did not consider interaction effects. Including these kinds of effects may provide a more realistic understanding of child growth, due to the multifaceted nature of growth and development in children, rather than focusing solely on individual effects. Finally, accounting for variation at the individual level was limited to random intercepts and random slopes in linear trend. However, in real-world problems, the situation can be more complex. For example, the trend may follow a high-order polynomial, such as a quadratic trend. Therefore, in the future work, these three aspects should be considered to provide a more realistic child growth model.

In the Bayesian group and sparse group LASSO approaches, there is an underlying assumption that group variables form a non-overlapping structure, such as multiple categorical variables and bases of smoothing functions representing continuous variables. However, in real-world scenarios, risk factors associated with child growth and development are not confined to this structure. They can overlap, meaning that certain factors or indicators are relevant to multiple aspects or dimension of child's growth and development. For example, family income, parental education, and housing conditions might overlap between groups related to economic factors and those groups related to environmental factors. Further work could extend these approaches to accommodate this phenomenon. Another aspect is that the penalty term used with the regularisation parameter is fixed at 1, associated with random effect selection in BSGSSMQR appears to mis-specify the random effects, which may not adequately capture the variability in the GUS data. Therefore, seeking a more appropriate method should be a focus in future work. One possible idea is to extend the Bayesian Group LASSO to incorporate shrinkage of random effects with group structure by using truncated normal priors, thereby addressing a similar problem in the selection of fixed effects. Moreover, a primary challenge in implementing a Bayesian approach with the traditional MCMC methods is its computational cost, which escalates with increasing sample size and the number of random effects. In improving this respect of Bayesian inference, seeking a faster alternative to the traditional methods, such as the Integrated Nested Laplace Approximation (INLA), may be a viable approach. In literature, there is interesting work that has applied the INLA method in quantile regression, for example, Yue and Rue (2011).

Most of approaches in this thesis, including AQMM and the proposed methods, rely on the independent estimation of each conditional quantile by a likelihood-based method via the asymmetric Laplace (AL) distribution. Consequently, they may breach the quantile monotonicity property, as outlined in Section 3.4.4, leading to crossing quantile curves, particularly in the extreme quantiles. This is especially relevant when constructing reference child growth curves. In future work, this limitation should be considered. An interesting workaround in this area is the work of Merhi Bleik (2019). The author employed a Bayesian approach with a Metropolis-Hastings within Gibbs algorithm to estimate simultaneous conditional quantile curves.

In conclusion, this thesis aims to examine longitudinal child growth data, with the objective of providing comprehensive insights into physical growth and development of children. The models used in this thesis are based on quantile mixed models within both frequentist and Bayesian frameworks. These approaches were compared through simulation studies and then applied to the real data, specifically the physical growth measurements (i.e. raw weight, raw height, WAZ, and HAZ) from the Growing Up in Scotland dataset. The novelty of this work lies in the development of a Bayesian variable selection method within quantile mixed models, providing a comprehensive framework to address the challenges associated with the simultaneous selection of both fixed and random effects in these models. This novel method combines four key components: the Bayesian sparse group LASSO method, a likelihood function based on the scale mixture representation of the asymmetric Laplace (AL) distribution, spike and slab priors for quantile regression coefficients, and the utilisation of mixed models based on a decomposition for the covariance matrix of random effects. The first component enables the selection of both group variables and individuals within these group variables. The second is employed as a working likelihood in the Bayesian method and for estimating the conditional quantiles. The third is designed to place a point mass at zero for quantile regression coefficients, ensuring effective identification and selection variables. The final component is adopted to select random effects. This model outperforms other methods in terms of model selection accuracy, especially when the study's objective is to select variables (fixed effects) exhibiting a non-overlapping group structure. Additionally, it has the capability of selecting random effects over others. However, there are some limitations to this work, which have been mentioned previously, along with future work that could overcome these limitations.

Appendix A

Joint posterior distributions

A.1 Joint posterior distribution of $\boldsymbol{\beta}, \boldsymbol{\eta}^2, \boldsymbol{v}, \sigma,\ \pi_0$

Let $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{ld_l})'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G)'$, $\boldsymbol{\eta}^2 = (\eta_1^2, \ldots, \eta_G^2)$, $\mathbf{X}$ is an $n \times p$ design matrix, $\boldsymbol{v} = (v_1, \ldots, v_n)'$, $\mathbf{V} = \mathrm{diag}(\boldsymbol{v}^{-1})$, $\boldsymbol{\Psi}_l$ is a known $d_l \times d_l$ positive definite matrix, and π_0 is the probability of $\boldsymbol{\beta}_l = 0$. The joint posterior distribution of $\boldsymbol{\beta}, \boldsymbol{\eta}^2, \boldsymbol{v}, \sigma, \pi_0$ is

$$
\begin{aligned}
p(\boldsymbol{\beta},&\boldsymbol{\eta}^2, \boldsymbol{v}, \sigma, \pi_0|\mathbf{y}, \mathbf{X})\\
&\propto l(\mathbf{y}|\boldsymbol{\beta}, \sigma, \boldsymbol{v})p(\boldsymbol{\beta}|\sigma, \boldsymbol{v}, \pi_0)p(\boldsymbol{\eta}^2|\lambda^2)p(\boldsymbol{v}|\sigma)p(\sigma)p(\pi_0)\\
&\propto (\sigma)^{-\frac{3n}{2}}\left(\prod_{i=1}^{n} v_i^{-\frac{1}{2}}\right)\\
&\quad\times \exp\left\{-\frac{1}{4\sigma}\left[\Big(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\Big)'\mathbf{V}\Big(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\Big) + \zeta\sum_{i=1}^{n} v_i\right]\right\}\\
&\quad\times \prod_{l=1}^{G}\left[(1-\pi_0)|2\pi\eta_l^2\boldsymbol{\Psi}_l^{-1}|^{-\frac{d_l}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}_l'(\eta^2\boldsymbol{\Psi}_l^{-1})^{-1}\boldsymbol{\beta}_l\right\}I[\boldsymbol{\beta}_l \neq 0] + \pi_0\delta_0(\boldsymbol{\beta}_l)\right]\\
&\quad\times \prod_{l=1}^{G}\left[(\lambda^2)^{-\frac{d_l+1}{2}}(\eta_l^2)^{\frac{d_l+1}{2}-1}\exp\left\{-\frac{\lambda^2}{2}\eta_l^2\right\}\right]\\
&\quad\times (\sigma)^{-g_1-1}\exp\left\{-\frac{g_2}{\sigma}\right\}\\
&\quad\times \pi_0^{a_1-1}(1-\pi_0)^{a_2-1}.
\end{aligned}
$$

A.2 Joint posterior distribution of $\boldsymbol{b}, \boldsymbol{\theta}^2, \boldsymbol{v}, \sigma, \pi_0, \pi_1$

Let $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{X}_l$ is an $n \times d_l$ matrix, $\mathbf{K}_{\tau,l}^{1/2} = \text{diag}\{\theta_{l1}, \ldots, \theta_{ld_l}\}, \theta_{lj} \geq 0, l = 1, \ldots, G; j = 1, \ldots, d_l$, $\boldsymbol{b}_l = (b_{l1}, \ldots, b_{ld_l})'$, $\boldsymbol{b} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_G)'$, $\boldsymbol{\theta}_l^2 = (\theta_{l1}^2, \ldots, \theta_{ld_l}^2)'$, $\boldsymbol{\theta}^2 = (\boldsymbol{\theta}_l^2, \ldots, \boldsymbol{\theta}_G^2)'$, $\boldsymbol{v} = (v_1, \ldots, v_n)'$, $\mathbf{V} = \text{diag}(\boldsymbol{v}^{-1})$, π_0 is the probability of $\boldsymbol{b}_l = 0$, and π_1 is the probability of $\theta_{lj} = 0$. The joint posterior distribution of $\boldsymbol{b}, \boldsymbol{\theta}^2, \boldsymbol{v}, \sigma, \pi_0, \pi_1$ is

$$
\begin{aligned}
&p(\boldsymbol{b},\boldsymbol{\theta}^2, \boldsymbol{v}, \sigma, \pi_0, \pi_1, s^2 | \mathbf{y}, \mathbf{X}) \\
&\quad \propto l(\mathbf{y}|\boldsymbol{\beta}, \sigma, \boldsymbol{v}) p(\boldsymbol{b}_l|\pi_0) p(\boldsymbol{\theta}^2|\pi_1) p(\boldsymbol{v}|\sigma) p(\sigma) p(\pi_0) p(\pi_1) p(s^2) \\
&\quad \propto (\sigma)^{-\frac{3n}{2}} \left(\prod_{i=1}^{n} v_i^{-\frac{1}{2}} \right) \\
&\qquad \times \exp\left\{ -\frac{1}{4\sigma} \left[\Big(\mathbf{y} - \sum_{l=1}^{G} \mathbf{X}_l \mathbf{K}_l^{1/2} \boldsymbol{b}_l - \xi \boldsymbol{v} \Big)' \mathbf{V} \Big(\mathbf{y} - \sum_{l=1}^{G} \mathbf{X}_l \mathbf{K}_l^{1/2} \boldsymbol{b}_l - \xi \boldsymbol{v} \Big) + \zeta \sum_{i=1}^{n} v_i \right] \right\} \\
&\qquad \times \prod_{l=1}^{G} \left[(1-\pi_0)(2\pi)^{-\frac{m_l}{2}} \exp\left\{ -\frac{1}{2} \boldsymbol{b}_l' \boldsymbol{b}_l \right\} I[\boldsymbol{b}_l \neq 0] + \pi_0 \delta_0(\boldsymbol{b}_l) \right] \\
&\qquad \times \prod_{l=1}^{G} \prod_{j=1}^{d_l} \left[(1-\pi_1) \cdot 2(2\pi s^2)^{-\frac{1}{2}} \exp\left\{ -\frac{\theta_{lj}^2}{2s^2} \right\} I[\theta_{lj} \neq 0] + \pi_1 \delta_0(\theta_{lj}) \right] \\
&\qquad \times (\sigma)^{-g_1-1} \exp\left\{ -\frac{g_2}{\sigma} \right\} \\
&\qquad \times \pi_0^{a_1-1} (1-\pi_0)^{a_2-1} \\
&\qquad \times \pi_1^{c_1-1} (1-\pi_1)^{c_2-1} \\
&\qquad \times t (s^2)^{-2} \exp\left\{ -\frac{t}{s^2} \right\}.
\end{aligned}
$$

Appendix B

Additional results from Chapter 4

In this appendix, additional results corresponding to Chapter 4 are presented, specifically from the two simulation studies conducted (Study 4.1 and 4.2).

B.1 Study 4.1

Each figure representing the results from the homogeneous variance scenario in Study 4.1 is listed in the table below.

Figure	**Metric**	**Scenario (refer to Table 4.1)**			
		#	**Data design**	**Variance-covariance types of errors (R_i)**	**Sample sizes**
B.1	MAE	1	Balanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	100
B.1	MAE	2	Balanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	1000
B.1	MAE	5	Unbalanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	100
B.1	MAE	6	Unbalanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	1000
B.2	RS	1	Balanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	100
B.2	RS	2	Balanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	1000
B.2	RS	5	Unbalanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	100
B.2	RS	6	Unbalanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	1000
B.3	PNR	1	Balanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	100
B.3	PNR	2	Balanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	1000
B.3	PNR	5	Unbalanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	100
B.3	PNR	6	Unbalanced data	Homogeneous ($\sigma^2 = 2, \phi = 1.45$)	1000

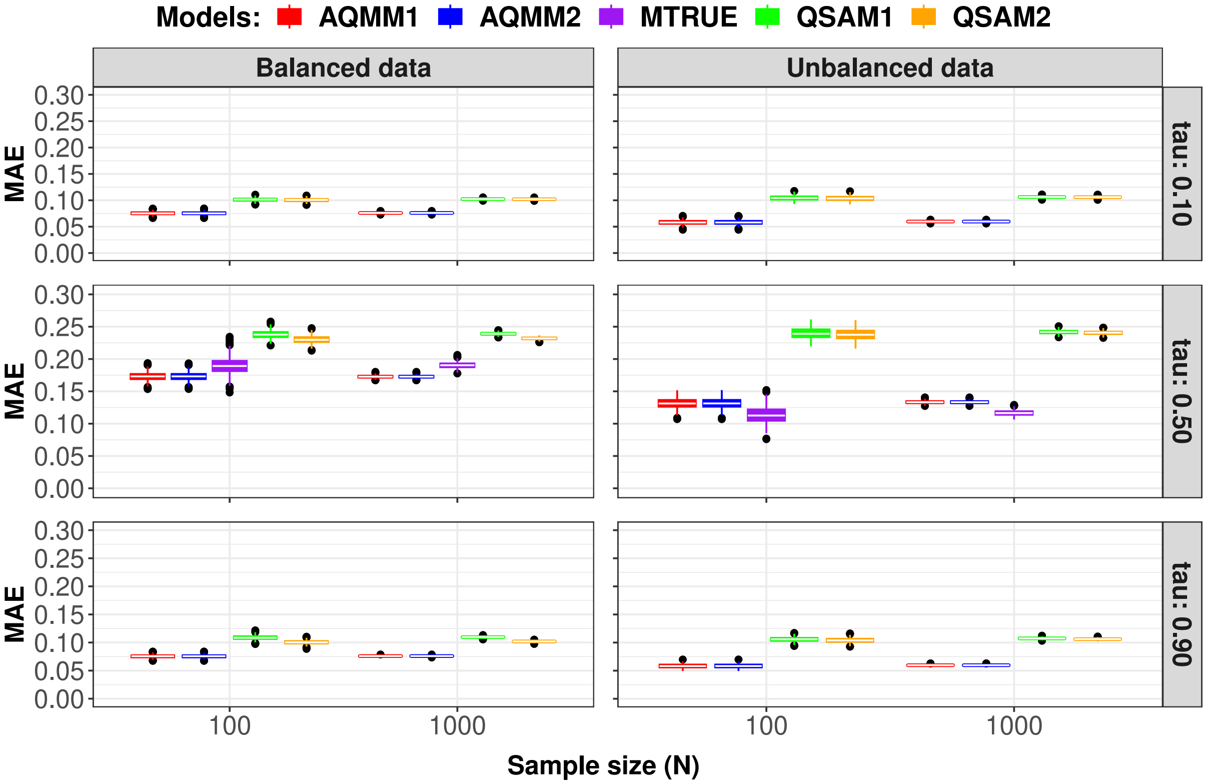


Figure B.1: The MAE of the four models, including the MSE for the true model (MTRUE), in the *homogeneous* scenario of Study 4.1. The left column presents the results for the balanced data scenario, while the right column shows the results for the unbalanced data scenario. The three rows display the results for quantile levels at 0.10, 0.50 and 0.90, respectively.

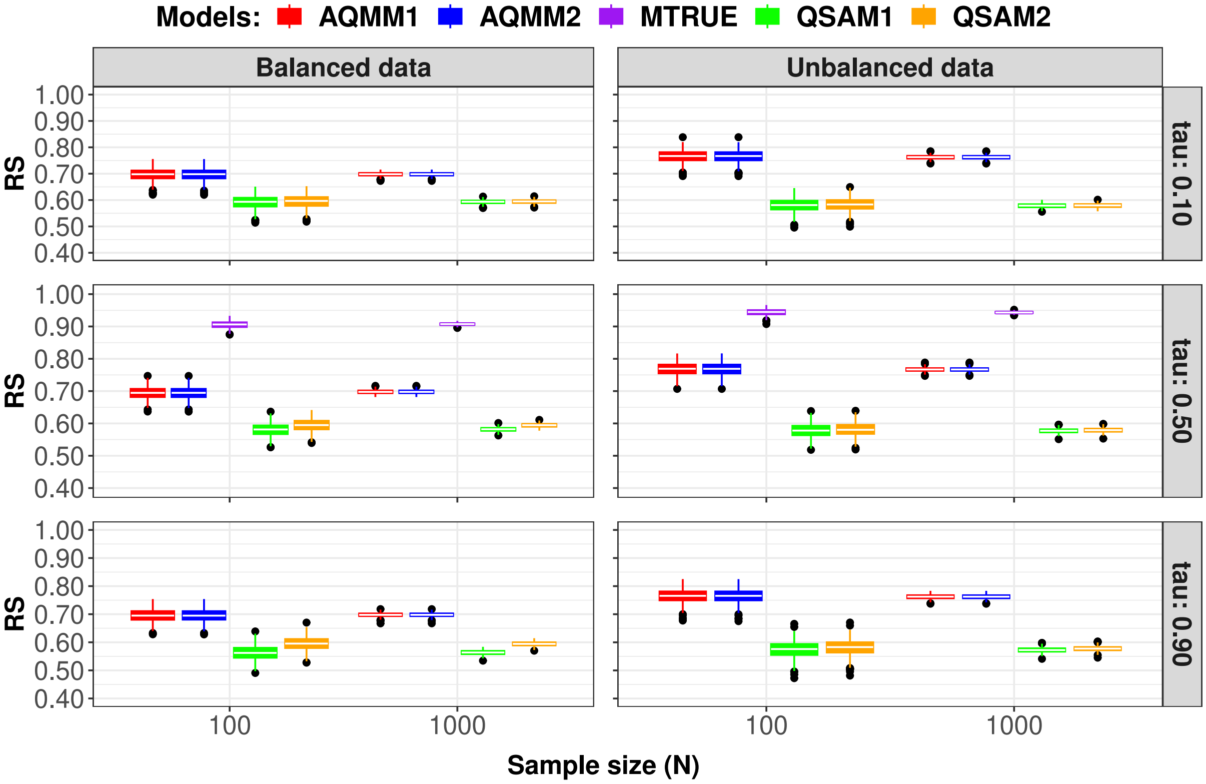


Figure B.2: The RS of the four models, including the true model (MTRUE), in the *homogeneous* scenario of Study 4.1. The left column presents the results for the balanced data scenario, while the right column shows the results for the unbalanced data scenario. The three rows display the results for quantile levels at 0.10, 0.50 and 0.90, respectively.

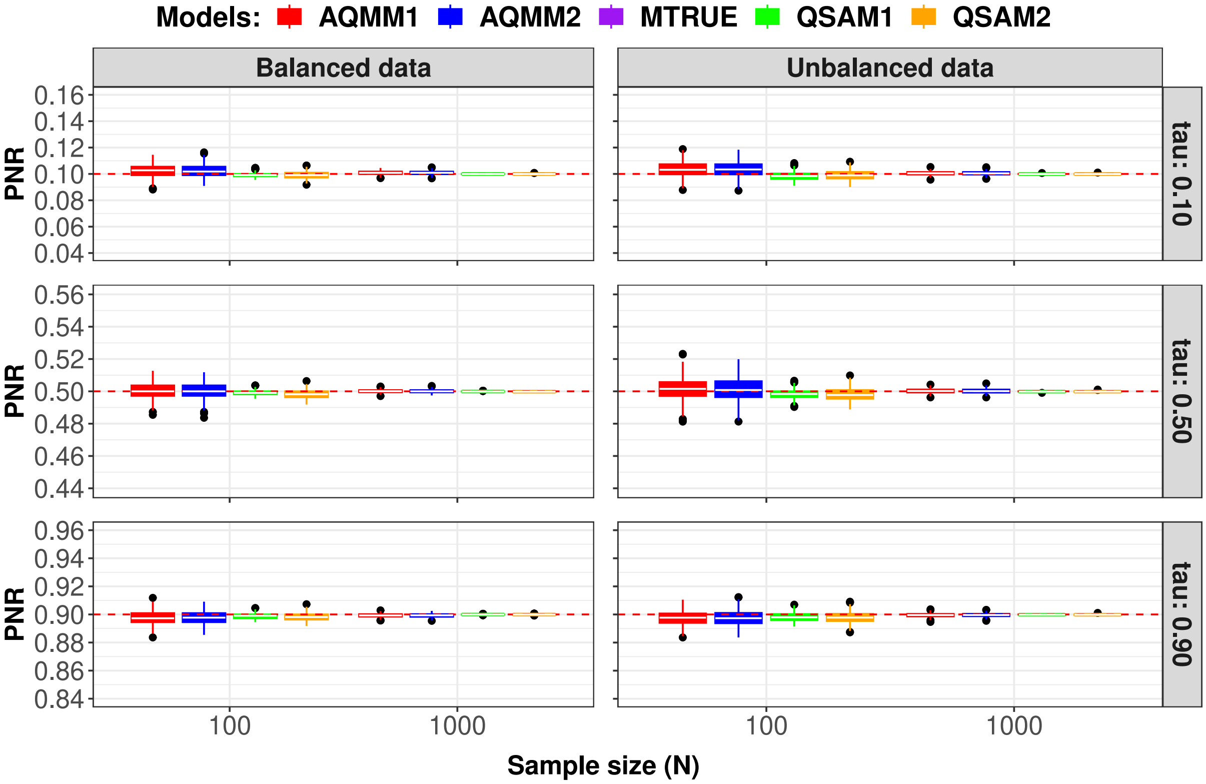


Figure B.3: The PNR of the four models in the *homogeneous* scenario Study 4.1. The left column presents the results for the balanced data scenario, while the right column shows the results for the unbalanced data scenario. The three rows display the results for quantile levels at 0.10, 0.50 and 0.90, respectively. The red dashed lines represent the expected quantile levels, $\tau = 0.10, 0.50$, and 0.90, respectively.

B.2 Study 4.2

Each figure and table representing the additional results from Study 4.2 is listed in the table below.

Figure/Table	Metric	Scenario (refer to Table 4.1)				
		#	Data design	Variance-covariance types of errors (R_i)	Sample sizes	Error distribution
B.4	MAE	1	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Normal
B.4	MAE	2	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Normal
B.4	MAE	5	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Normal
B.4	MAE	6	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Normal
B.4	MAE	1	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Student's t
B.4	MAE	2	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Student's t
B.4	MAE	5	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Student's t
B.4	MAE	6	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Student's t
B.5	RS	1	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Normal
B.5	RS	2	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Normal
B.5	RS	5	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Normal
B.5	RS	6	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Normal
B.5	RS	1	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Student's t
B.5	RS	2	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Student's t
B.5	RS	5	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Student's t
B.5	RS	6	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Student's t
B.6	PNR	1	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Normal
B.6	PNR	2	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Normal
B.6	PNR	5	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Normal
B.6	PNR	6	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Normal
B.6	PNR	1	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Student's t
B.6	PNR	2	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Student's t
B.6	PNR	5	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Student's t
B.6	PNR	6	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Student's t
B.1	Bias, RMSE	1	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Normal
B.1	Bias, RMSE	1	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Student's t
B.2	Bias, RMSE	2	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Normal
B.2	Bias, RMSE	2	Balanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Student's t
B.3	Bias, RMSE	5	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Normal
B.3	Bias, RMSE	5	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	100	Student's t
B.4	Bias, RMSE	6	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Normal
B.4	Bias, RMSE	6	Unbalanced data	Homogeneous ($\sigma^2 = 1, \phi = 1.45$)	1000	Student's t
B.5	Bias, RMSE	3	Balanced data	Heterogeneous ($\sigma^2 = 1, \phi = 1.45, \alpha = -0.50$)	100	Normal
B.5	Bias, RMSE	3	Balanced data	Heterogeneous ($\sigma^2 = 1, \phi = 1.45, \alpha = -0.50$)	100	Student's t
B.6	Bias, RMSE	4	Balanced data	Heterogeneous ($\sigma^2 = 1, \phi = 1.45, \alpha = -0.50$)	1000	Normal
B.6	Bias, RMSE	4	Balanced data	Heterogeneous ($\sigma^2 = 1, \phi = 1.45, \alpha = -0.50$)	1000	Student's t

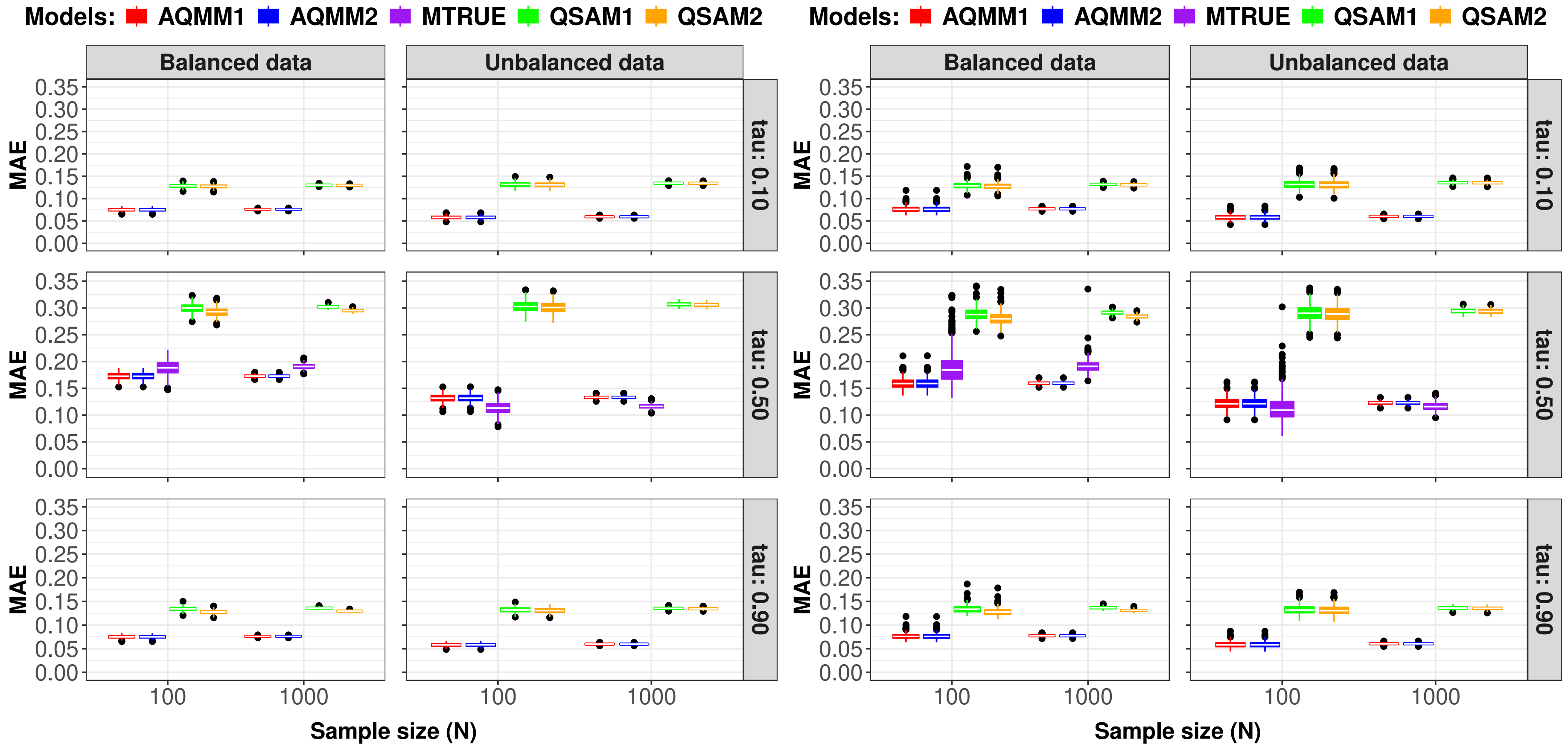


Figure B.4: The MAE of the four models in the *homogeneous* scenario of Study 4.2. The left figure represents the normal error case, while the right figure represents the t ($\mathcal{T}_4$) error case. The three rows in each figure contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively.

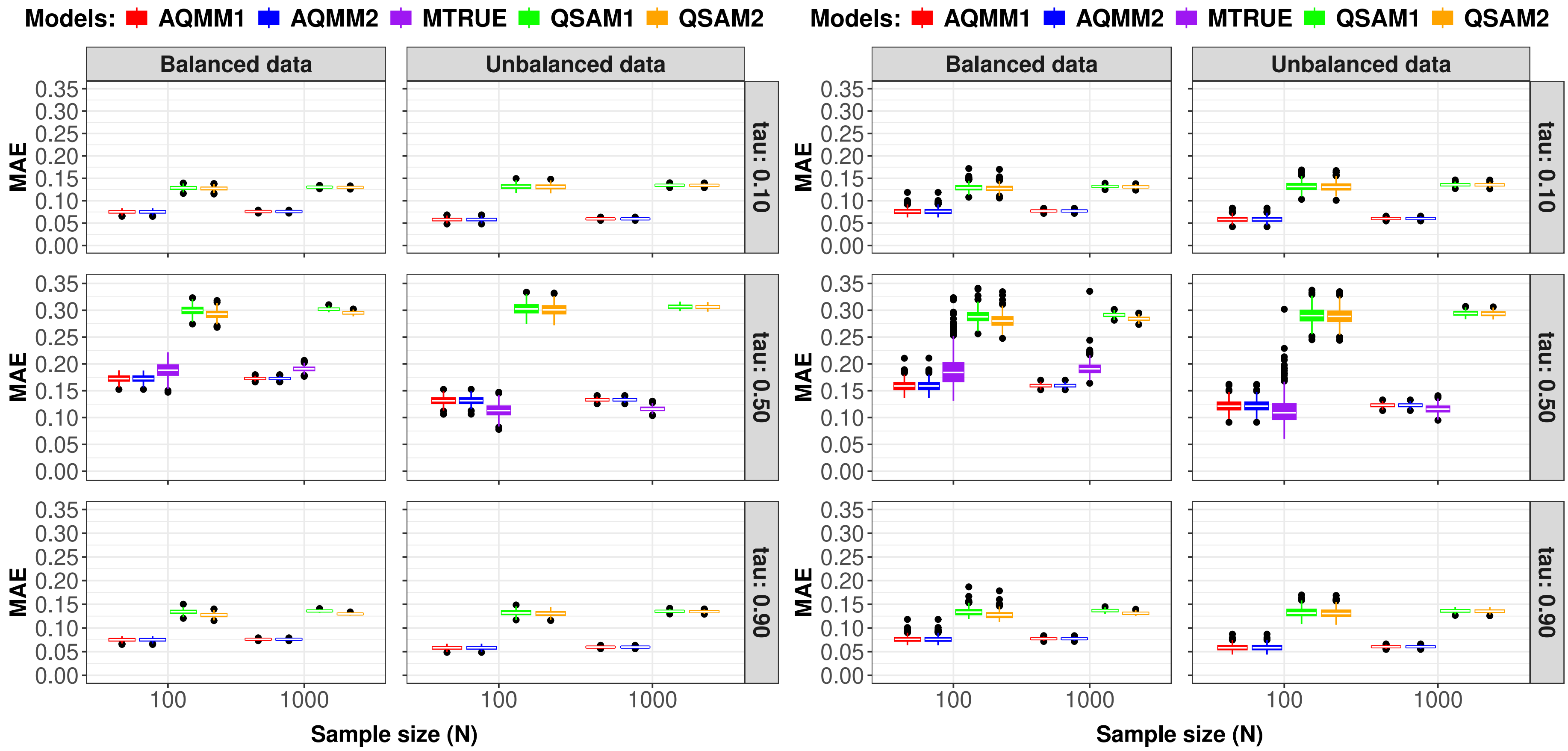


Figure B.5: The RS of the four models in the *homogeneous* scenario of Study 4.2. The left figure represents the normal error case, while the right figure represents the t ($\mathcal{T}_4$) error case. The three rows in each figure contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively.

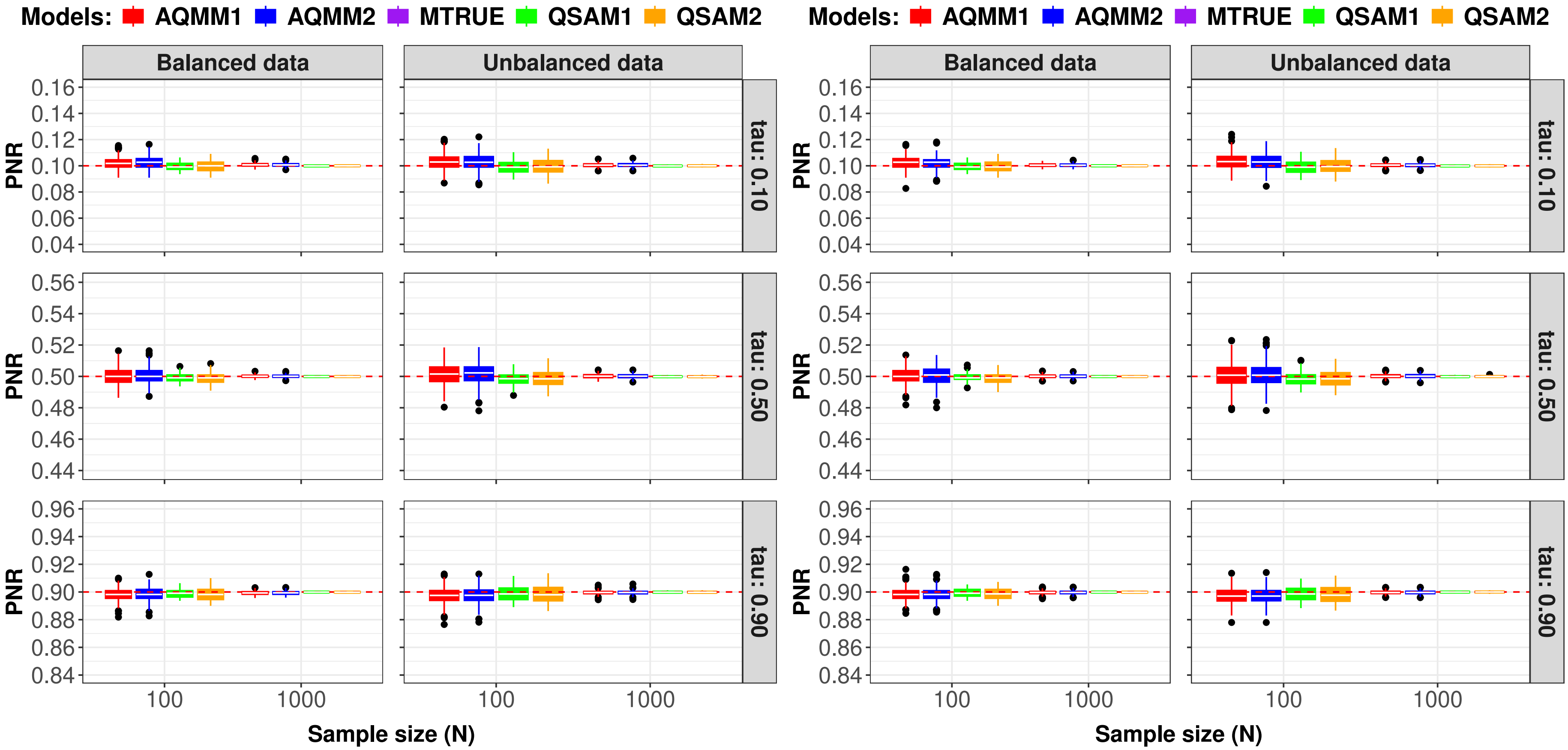


Figure B.6: The PNR of the four models in the *homogeneous* scenario of Study 4.2. The left figure represents the normal error case, while the right figure represents the t ($\mathcal{T}_4$) error case. The three rows in each figure contain the results for quantile levels at 0.10, 0.50 and 0.90, respectively.

Table B.1: The MBE and RMSE concerning the simulated data under Study 4.2, specifically focusing on the **balanced** data design, **homogeneous** variance-covariance of errors with two distinct error distributions, and a sample size of **100**.

Error	**Model**	τ	**MBE**	**RMSE**	**MBE**	**RMSE**
			β_1	β_1	β_2	β_2
$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i)$	AQMM1	0.10	0.0115	0.2512	-0.0022	0.0253
	AQMM2	0.10	0.0115	0.2513	-0.0021	0.0252
	QSAM1	0.10	-0.7832	0.7883	0.0002	0.0357
	QSAM2	0.10	-0.7825	0.7876	-0.0011	0.0370
	MTRUE	Mean	0.0120	0.2497	-0.0008	0.0160
	AQMM1	0.50	0.0125	0.2507	-0.0006	0.0185
	AQMM2	0.50	0.0125	0.2507	-0.0006	0.0184
	QSAM1	0.50	-0.7727	0.7767	0.0007	0.0273
	QSAM2	0.50	-0.7800	0.7837	-0.0002	0.0258
	AQMM1	0.90	0.0101	0.2545	-0.0005	0.0241
	AQMM2	0.90	0.0101	0.2546	-0.0005	0.0241
	QSAM1	0.90	-0.7522	0.7594	0.0016	0.0373
	QSAM2	0.90	-0.7762	0.7818	0.0007	0.0368
$\boldsymbol{\epsilon}_i \sim \mathcal{T}_{n_i,4}(\mathbf{0}, \mathbf{\Sigma}_i)$	AQMM1	0.10	0.0111	0.2476	0.0004	0.0263
	AQMM2	0.10	0.0111	0.2476	0.0005	0.0263
	QSAM1	0.10	-0.7824	0.7892	0.0029	0.0357
	QSAM2	0.10	-0.7801	0.7868	0.0012	0.0366
	MTRUE	Mean	0.0084	0.2413	-0.0003	0.0167
	AQMM1	0.50	0.0080	0.2413	-0.0009	0.0157
	AQMM2	0.50	0.0080	0.2412	-0.0009	0.0157
	QSAM1	0.50	-0.7466	0.7501	0.0010	0.0254
	QSAM2	0.50	-0.7589	0.7622	-0.0003	0.0253
	AQMM1	0.90	0.0067	0.2512	-0.0004	0.0252
	AQMM2	0.90	0.0066	0.2512	-0.0005	0.0252
	QSAM1	0.90	-0.7533	0.7611	-0.0012	0.0389
	QSAM2	0.90	-0.7825	0.7891	-0.0018	0.0382

Table B.2: The MBE and RMSE concerning the simulated data under Study 4.2, specifically focusing on the **balanced** data design, **homogeneous** variance-covariance of errors with two distinct error distributions, and a sample size of **1000**.

Error	**Model**	τ	**MBE**	**RMSE**	**MBE**	**RMSE**
			β_1	β_1	β_2	β_2
$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i)$	AQMM1	0.10	-0.0002	0.0836	-0.0002	0.0077
	AQMM2	0.10	-0.0002	0.0836	-0.0002	0.0077
	QSAM1	0.10	-0.7923	0.7928	0.0007	0.0112
	QSAM2	0.10	-0.7917	0.7922	0.0008	0.0112
	MTRUE	Mean	0.0006	0.0813	-0.0002	0.0052
	AQMM1	0.50	0.0004	0.0814	-0.0001	0.0060
	AQMM2	0.50	0.0004	0.0814	-0.0001	0.0060
	QSAM1	0.50	-0.7805	0.7809	0.0002	0.0087
	QSAM2	0.50	-0.7894	0.7898	0.0004	0.0087
	AQMM1	0.90	0.0008	0.0825	0.0001	0.0076
	AQMM2	0.90	0.0008	0.0824	0.0001	0.0077
	QSAM1	0.90	-0.7650	0.7655	0.0002	0.0118
	QSAM2	0.90	-0.7898	0.7902	0.0000	0.0115
$\boldsymbol{\epsilon}_i \sim \mathcal{T}_{n_i,4}(\mathbf{0}, \boldsymbol{\Sigma}_i)$	AQMM1	0.10	-0.0038	0.0809	0.0002	0.0082
	AQMM2	0.10	-0.0038	0.0809	0.0002	0.0083
	QSAM1	0.10	-0.7996	0.8001	0.0010	0.0117
	QSAM2	0.10	-0.7983	0.7989	0.0008	0.0120
	MTRUE	Mean	-0.0046	0.0792	0.0001	0.0050
	AQMM1	0.50	-0.0048	0.0788	0.0001	0.0046
	AQMM2	0.50	-0.0048	0.0788	0.0001	0.0046
	QSAM1	0.50	-0.7585	0.7588	0.0006	0.0079
	QSAM2	0.50	-0.7681	0.7685	0.0001	0.0077
	AQMM1	0.90	-0.0057	0.0823	-0.0004	0.0080
	AQMM2	0.90	-0.0057	0.0823	-0.0004	0.0080
	QSAM1	0.90	-0.7716	0.7723	0.0003	0.0122
	QSAM2	0.90	-0.7967	0.7973	0.0002	0.0117

Table B.3: The MBE and RMSE concerning the simulated data under Study 4.2, specifically focusing on the **unbalanced** data design, **homogeneous** variance-covariance of errors with two distinct error distributions, and a sample size of **100**.

Error	**Model**	τ	**MBE**	**RMSE**	**MBE**	**RMSE**
			β_1	β_1	β_2	β_2
$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i)$	AQMM1	0.10	0.0177	0.2658	0.0008	0.0258
	AQMM2	0.10	0.0177	0.2656	0.0007	0.0258
	QSAM1	0.10	-0.7662	0.7761	0.0037	0.0478
	QSAM2	0.10	-0.7652	0.7751	0.0033	0.0491
	MTRUE	Mean	0.0204	0.2636	-0.0001	0.0175
	AQMM1	0.50	0.0210	0.2644	0.0002	0.0202
	AQMM2	0.50	0.0211	0.2646	0.0003	0.0203
	QSAM1	0.50	-0.7558	0.7609	0.0007	0.0342
	QSAM2	0.50	-0.7601	0.7650	0.0006	0.0338
	AQMM1	0.90	0.0195	0.2669	-0.0003	0.0251
	AQMM2	0.90	0.0196	0.2669	-0.0002	0.0253
	QSAM1	0.90	-0.7514	0.7604	-0.0001	0.0477
	QSAM2	0.90	-0.7574	0.7658	-0.0007	0.0488
$\boldsymbol{\epsilon}_i \sim \mathcal{T}_{n_i,4}(\mathbf{0}, \boldsymbol{\Sigma}_i)$	AQMM1	0.10	-0.0129	0.2943	0.0016	0.0284
	AQMM2	0.10	-0.0128	0.2944	0.0016	0.0284
	QSAM1	0.10	-0.7682	0.7774	0.0036	0.0486
	QSAM2	0.10	-0.7688	0.7781	0.0023	0.0496
	MTRUE	Mean	-0.0158	0.2911	-0.0005	0.0185
	AQMM1	0.50	-0.0149	0.2900	-0.0011	0.0171
	AQMM2	0.50	-0.0151	0.2898	-0.0011	0.0170
	QSAM1	0.50	-0.7370	0.7429	0.0005	0.0320
	QSAM2	0.50	-0.7406	0.7464	0.0001	0.0317
	AQMM1	0.90	-0.0169	0.2977	-0.0020	0.0260
	AQMM2	0.90	-0.0166	0.2977	-0.0021	0.0260
	QSAM1	0.90	-0.7656	0.7753	-0.0012	0.0509
	QSAM2	0.90	-0.7718	0.7819	-0.0013	0.0503

Table B.4: The MBE and RMSE concerning the simulated data under Study 4.2, specifically focusing on the **unbalanced** data design, **homogeneous** variance-covariance of errors with two distinct error distributions, and a sample size of **1000**.

Error	**Model**	τ	**MBE**	**RMSE**	**MBE**	**RMSE**
			β_1	β_1	β_2	β_2
$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i)$	AQMM1	0.10	0.0007	0.0845	-0.0002	0.0081
	AQMM2	0.10	0.0006	0.0845	-0.0002	0.0081
	QSAM1	0.10	-0.7711	0.7719	0.0012	0.0156
	QSAM2	0.10	-0.7722	0.7731	0.0013	0.0156
	MTRUE	Mean	0.0015	0.0827	-0.0002	0.0056
	AQMM1	0.50	0.0022	0.0834	-0.0002	0.0059
	AQMM2	0.50	0.0021	0.0834	-0.0002	0.0059
	QSAM1	0.50	-0.7671	0.7676	0.0001	0.0105
	QSAM2	0.50	-0.7695	0.7700	0.0002	0.0104
	AQMM1	0.90	0.0010	0.0829	0.0002	0.0079
	AQMM2	0.90	0.0009	0.0830	0.0003	0.0079
	QSAM1	0.90	-0.7676	0.7684	0.0004	0.0148
	QSAM2	0.90	-0.7721	0.7729	0.0004	0.0144
$\boldsymbol{\epsilon}_i \sim \mathcal{T}_{n_i,4}(\mathbf{0}, \boldsymbol{\Sigma}_i)$	AQMM1	0.10	0.0062	0.0882	0.0008	0.0084
	AQMM2	0.10	0.0062	0.0882	0.0008	0.0083
	QSAM1	0.10	-0.7775	0.7785	0.0012	0.0149
	QSAM2	0.10	-0.7780	0.7790	0.0014	0.0148
	MTRUE	Mean	0.0050	0.0856	0.0004	0.0054
	AQMM1	0.50	0.0049	0.0855	0.0001	0.0049
	AQMM2	0.50	0.0049	0.0855	0.0001	0.0049
	QSAM1	0.50	-0.7405	0.7410	0.0005	0.0104
	QSAM2	0.50	-0.7425	0.7431	0.0004	0.0102
	AQMM1	0.90	0.0039	0.0873	-0.0001	0.0085
	AQMM2	0.90	0.0039	0.0872	-0.0001	0.0085
	QSAM1	0.90	-0.7726	0.7737	0.0002	0.0147
	QSAM2	0.90	-0.7772	0.7782	0.0002	0.0147

Table B.5: The MBE and RMSE concerning the simulated data under Study 4.2, specifically focusing on the **balanced** data design, **heterogeneous** variance-covariance of errors with two distinct error distributions, and a sample size of **100**.

Error	**Model**	τ	**MBE**	**RMSE**	**MBE**	**RMSE**
			β_1	β_1	β_2	β_2
$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i)$	AQMM1	0.10	0.0115	0.1798	-0.0014	0.0208
	AQMM2	0.10	0.0115	0.1798	-0.0015	0.0208
	QSAM1	0.10	-0.7138	0.7183	-0.0004	0.0305
	QSAM2	0.10	-0.7118	0.7166	-0.0002	0.0308
	MTRUE	Mean	0.0115	0.1776	-0.0006	0.0132
	AQMM1	0.50	0.0116	0.1786	-0.0006	0.0154
	AQMM2	0.50	0.0116	0.1786	-0.0006	0.0154
	QSAM1	0.50	-0.7033	0.7071	0.0010	0.0232
	QSAM2	0.50	-0.7100	0.7136	-0.0001	0.0225
	AQMM1	0.90	0.0111	0.1837	-0.0004	0.0204
	AQMM2	0.90	0.0110	0.1836	-0.0004	0.0203
	QSAM1	0.90	-0.6860	0.6926	0.0024	0.0315
	QSAM2	0.90	-0.7080	0.7135	0.0008	0.0317
$\boldsymbol{\epsilon}_i \sim \mathcal{T}_{n_i,4}(\mathbf{0}, \boldsymbol{\Sigma}_i)$	AQMM1	0.10	0.0097	0.1779	0.0003	0.0219
	AQMM2	0.10	0.0097	0.1780	0.0003	0.0218
	QSAM1	0.10	-0.7128	0.7188	0.0030	0.0311
	QSAM2	0.10	-0.7103	0.7167	0.0003	0.0315
	MTRUE	Mean	0.0071	0.1725	-0.0002	0.0139
	AQMM1	0.50	0.0069	0.1715	-0.0006	0.0134
	AQMM2	0.50	0.0069	0.1714	-0.0006	0.0134
	QSAM1	0.50	-0.6742	0.6779	0.0001	0.0221
	QSAM2	0.50	-0.6839	0.6872	-0.0002	0.0219
	AQMM1	0.90	0.0062	0.1829	-0.0003	0.0211
	AQMM2	0.90	0.0061	0.1830	-0.0002	0.0210
	QSAM1	0.90	-0.6842	0.6909	-0.0016	0.0326
	QSAM2	0.90	-0.7122	0.7181	-0.0020	0.0319

Table B.6: The MBE and RMSE concerning the simulated data under Study 4.2, specifically focusing on the **balanced** data design, **heterogeneous** variance-covariance of errors with two distinct error distributions, and a sample size of **1000**.

Error	**Model**	τ	**MBE**	**RMSE**	**MBE**	**RMSE**
			β_1	β_1	β_2	β_2
$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i)$	AQMM1	0.10	0.0016	0.0592	-0.0003	0.0063
	AQMM2	0.10	0.0016	0.0592	-0.0003	0.0063
	QSAM1	0.10	-0.7231	0.7236	0.0009	0.0095
	QSAM2	0.10	-0.7212	0.7217	0.0008	0.0095
	MTRUE	Mean	0.0020	0.0570	-0.0002	0.0042
	AQMM1	0.50	0.0020	0.0570	-0.0001	0.0048
	AQMM2	0.50	0.0020	0.0570	-0.0001	0.0048
	QSAM1	0.50	-0.7114	0.7118	0.0003	0.0078
	QSAM2	0.50	-0.7196	0.7199	0.0000	0.0076
	AQMM1	0.90	0.0025	0.0584	-0.0002	0.0062
	AQMM2	0.90	0.0024	0.0584	-0.0002	0.0062
	QSAM1	0.90	-0.6963	0.6969	0.0005	0.0106
	QSAM2	0.90	-0.7180	0.7184	0.0002	0.0100
$\boldsymbol{\epsilon}_i \sim \mathcal{T}_{n_i,4}(\mathbf{0}, \boldsymbol{\Sigma}_i)$	AQMM1	0.10	-0.0016	0.0582	0.0001	0.0068
	AQMM2	0.10	-0.0016	0.0581	0.0001	0.0068
	QSAM1	0.10	-0.7301	0.7306	0.0010	0.0101
	QSAM2	0.10	-0.7274	0.7279	0.0005	0.0101
	MTRUE	Mean	-0.0025	0.0563	0.0000	0.0041
	AQMM1	0.50	-0.0026	0.0560	0.0000	0.0037
	AQMM2	0.50	-0.0026	0.0560	0.0000	0.0037
	QSAM1	0.50	-0.6864	0.6868	0.0005	0.0068
	QSAM2	0.50	-0.6953	0.6956	0.0001	0.0068
	AQMM1	0.90	-0.0034	0.0587	-0.0004	0.0067
	AQMM2	0.90	-0.0034	0.0587	-0.0004	0.0067
	QSAM1	0.90	-0.7023	0.7030	0.0002	0.0102
	QSAM2	0.90	-0.7261	0.7267	0.0004	0.0099

Appendix C

Additional results from Chapter 5

In this appendix, the numerical results from two simulation studies, Studies 5.1 to 5.2, are presented, as detailed in Chapter 5. Each table representing these numerical results is listed in the table below.

Table	**Study**	**Metric**	**Error distribution**	**Note**
C.1	5.1	MBE	Standard normal	Comparison in eight different methods
C.2	5.1	RMSE	Standard normal	Comparison in eight different methods
C.3	5.1	MBE	t_3	Comparison in eight different methods
C.4	5.1	RMSE	t_3	Comparison in eight different methods
C.5	5.1	MBE	χ^2_3	Comparison in eight different methods
C.6	5.1	RMSE	χ^2_3	Comparison in eight different methods
C.7	5.1	MBE	Standard normal	Comparison between BGLSSQR and BSGSSQR in two different iteration numbers for estimating λ
C.8	5.1	RMSE	Standard normal	Comparison between BGLSSQR and BSGSSQR in two different iteration numbers for estimating λ
C.9	5.1	MBE	t_3	Comparison between BGLSSQR and BSGSSQR in two different iteration numbers for estimating λ
C.10	5.1	RMSE	t_3	Comparison between BGLSSQR and BSGSSQR in two different iteration numbers for estimating λ
C.11	5.1	MBE	χ^2_3	Comparison between BGLSSQR and BSGSSQR in two different iteration numbers for estimating λ
C.12	5.1	RMSE	χ^2_3	Comparison between BGLSSQR and BSGSSQR in two different iteration numbers for estimating λ
C.13	5.2	MBE	Standard normal	Comparison in eight different methods
C.14	5.2	RMSE	Standard normal	Comparison in eight different methods
C.15	5.2	MBE	t_3	Comparison in eight different methods
C.16	5.2	RMSE	t_3	Comparison in eight different methods
C.17	5.2	MBE	χ^2_3	Comparison in eight different methods
C.18	5.2	RMSE	χ^2_3	Comparison in eight different methods

Table C.1: Estimated **MBE** from eight methods in Study 5.1, assuming a **standard normal** error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	CQR	-0.0115	-0.0080	-0.0003	0.0043	-0.0104	-0.0045	0.0057	0.0011	-0.0021
	LASSOQR	-0.0528	-0.0626	-0.0581	-0.0506	-0.0478	0.0002	0.0079	0.0140	-0.0022
	GLASSOQR	-0.0453	-0.0490	-0.0416	-0.0483	-0.0344	-0.0021	0.0098	0.0080	-0.0004
	BLASSOQR	-0.0323	-0.0337	-0.0228	-0.0183	-0.0314	-0.0042	0.0011	0.0015	-0.0029
	BGLSSQR$_{mean}$	-0.0229	-0.0279	-0.0140	-0.0229	-0.0222	-0.0023	0.0008	-0.0001	-0.0004
	BGLSSQR$_{median}$	-0.0228	-0.0277	-0.0137	-0.0213	-0.0225	-0.0024	0.0005	0.0007	-0.0004
	BSGSSQR$_{mean}$	-0.0239	-0.0263	-0.0149	-0.0227	-0.0527	-0.0029	0.0009	0.0005	-0.0020
	BSGSSQR$_{median}$	-0.0237	-0.0260	-0.0147	-0.0211	-0.0572	-0.0027	0.0011	0.0009	-0.0016
0.50	CQR	-0.0115	-0.0080	-0.0003	0.0043	-0.0104	-0.0045	0.0057	0.0011	-0.0021
	LASSOQR	-0.0726	-0.0694	-0.0534	-0.0574	-0.0695	-0.0029	-0.0014	-0.0036	-0.0062
	GLASSOQR	-0.0636	-0.0634	-0.0520	-0.0746	-0.0534	-0.0031	0.0004	-0.0010	-0.0026
	BLASSOQR	-0.0323	-0.0337	-0.0228	-0.0183	-0.0314	-0.0042	0.0011	0.0015	-0.0029
	BGLSSQR$_{mean}$	-0.0229	-0.0279	-0.0140	-0.0232	-0.0222	-0.0022	0.0008	-0.0002	-0.0003
	BGLSSQR$_{median}$	-0.0228	-0.0278	-0.0138	-0.0214	-0.0225	-0.0023	0.0005	0.0006	-0.0001
	BSGSSQR$_{mean}$	-0.0240	-0.0262	-0.0150	-0.0227	-0.0527	-0.0026	0.0009	0.0005	-0.0020
	BSGSSQR$_{median}$	-0.0238	-0.0259	-0.0147	-0.0211	-0.0570	-0.0025	0.0012	0.0009	-0.0015
0.90	CQR	-0.0115	-0.0080	-0.0003	0.0043	-0.0104	-0.0045	0.0057	0.0011	-0.0021
	LASSOQR	-0.0466	-0.0558	-0.0412	-0.0462	-0.0491	-0.0085	0.0021	-0.0009	-0.0122
	GLASSOQR	-0.0355	-0.0464	-0.0304	-0.0484	-0.0391	-0.0055	-0.0024	-0.0038	-0.0091
	BLASSOQR	-0.0323	-0.0337	-0.0228	-0.0183	-0.0314	-0.0042	0.0011	0.0015	-0.0029
	BGLSSQR$_{mean}$	-0.0229	-0.0280	-0.0141	-0.0229	-0.0212	-0.0022	0.0008	-0.0002	-0.0004
	BGLSSQR$_{median}$	-0.0228	-0.0278	-0.0139	-0.0213	-0.0224	-0.0024	0.0004	0.0006	-0.0002
	BSGSSQR$_{mean}$	-0.0240	-0.0263	-0.0149	-0.0228	-0.0527	-0.0028	0.0009	0.0004	-0.0019
	BSGSSQR$_{median}$	-0.0238	-0.0259	-0.0147	-0.0211	-0.0571	-0.0026	0.0012	0.0010	-0.0015

Table C.2: Estimated **RMSE** from eight methods in Study 5.1, assuming a **standard normal** error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	CQR	0.1318	0.1319	0.1317	0.1299	0.1361	0.1260	0.1267	0.1318	0.1335
	LASSOQR	0.1835	0.1921	0.1945	0.1884	0.1880	0.1412	0.1402	0.1428	0.1393
	GLASSOQR	0.1706	0.1736	0.1784	0.1743	0.1683	0.1502	0.1421	0.1403	0.1442
	BLASSOQR	0.1195	0.1229	0.1202	0.1213	0.1224	0.0995	0.1030	0.1029	0.1045
	BGLSSQR$_{mean}$	0.1148	0.1200	0.1188	0.1281	0.1188	0.1055	0.0407	0.0361	0.0405
	BGLSSQR$_{median}$	0.1154	0.1206	0.1193	0.1276	0.1209	0.1058	0.0377	0.0303	0.0357
	BSGSSQR$_{mean}$	0.1170	0.1207	0.1177	0.1359	0.1439	0.0676	0.0482	0.0431	0.0500
	BSGSSQR$_{median}$	0.1176	0.1214	0.1181	0.1367	0.1562	0.0588	0.0439	0.0371	0.0455
0.50	CQR	0.1318	0.1319	0.1317	0.1299	0.1361	0.1260	0.1267	0.1318	0.1335
	LASSOQR	0.1517	0.1487	0.1443	0.1478	0.1535	0.0861	0.0874	0.0931	0.0942
	GLASSOQR	0.1366	0.1446	0.1383	0.1577	0.1318	0.1040	0.0944	0.0931	0.0929
	BLASSOQR	0.1195	0.1229	0.1203	0.1213	0.1224	0.0995	0.1030	0.1029	0.1044
	BGLSSQR$_{mean}$	0.1148	0.1199	0.1188	0.1279	0.1189	0.1056	0.0406	0.0362	0.0406
	BGLSSQR$_{median}$	0.1153	0.1253	0.1193	0.1275	0.1209	0.1058	0.0378	0.0308	0.0360
	BSGSSQR$_{mean}$	0.1171	0.1207	0.1178	0.1357	0.1439	0.0675	0.0482	0.0434	0.0501
	BSGSSQR$_{median}$	0.1176	0.1214	0.1182	0.1366	0.1560	0.0586	0.0439	0.0375	0.0454
0.90	CQR	0.1318	0.1319	0.1317	0.1299	0.1361	0.1260	0.1267	0.1318	0.1335
	LASSOQR	0.1885	0.2015	0.1832	0.1841	0.1774	0.1415	0.1505	0.1488	0.1473
	GLASSOQR	0.1741	0.1813	0.1665	0.1769	0.1567	0.1461	0.1470	0.1409	0.1430
	BLASSOQR	0.1195	0.1229	0.1203	0.1214	0.1224	0.0995	0.1030	0.1029	0.1045
	BGLSSQR$_{mean}$	0.1147	0.1199	0.1188	0.1280	0.1188	0.1056	0.0407	0.0361	0.0404
	BGLSSQR$_{median}$	0.1153	0.1206	0.1193	0.1276	0.1208	0.1058	0.0378	0.0304	0.0362
	BSGSSQR$_{mean}$	0.1171	0.1207	0.1178	0.1359	0.1441	0.0675	0.0481	0.0433	0.0500
	BSGSSQR$_{median}$	0.1176	0.1214	0.1182	0.1367	0.1563	0.0586	0.0439	0.0372	0.0453

Table C.3: Estimated **MBE** from eight methods in Study 5.1, assuming a t_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	CQR	0.0125	0.0062	0.0106	-0.0014	0.0019	-0.0044	0.0006	-0.0049	0.0063
	LASSOQR	-0.0584	-0.0790	-0.1075	-0.0872	-0.0417	-0.0059	0.0035	0.0008	-0.0206
	GLASSOQR	-0.0495	-0.0640	-0.0889	-0.0851	-0.0312	-0.0025	0.0104	0.0038	-0.0158
	BLASSOQR	-0.0276	-0.0344	-0.0274	-0.0353	-0.0294	-0.0060	0.0012	-0.0030	0.0067
	BGLSSQR$_{mean}$	-0.0182	-0.0218	-0.0188	-0.0674	-0.0401	-0.0086	-0.0015	0.0000	0.0021
	BGLSSQR$_{median}$	-0.0178	-0.0213	-0.0184	-0.0670	-0.0426	-0.0074	-0.0013	0.0007	0.0005
	BSGSSQR$_{mean}$	-0.0208	-0.0242	-0.0182	-0.0761	-0.0800	-0.0042	0.0006	-0.0005	0.0023
	BSGSSQR$_{median}$	-0.0203	-0.0236	-0.0176	-0.0773	-0.0915	-0.0024	-0.0005	-0.0004	0.0010
0.50	CQR	0.0125	0.0062	0.0106	-0.0014	0.0019	-0.0044	0.0006	-0.0049	0.0063
	LASSOQR	-0.0695	-0.0804	-0.0694	-0.0833	-0.0751	-0.0033	-0.0012	-0.0025	0.0058
	GLASSOQR	-0.0632	-0.0682	-0.0663	-0.1069	-0.0648	-0.0053	-0.0005	0.0024	0.0075
	BLASSOQR	-0.0276	-0.0344	-0.0274	-0.0353	-0.0294	-0.0060	0.0012	-0.0031	0.0067
	BGLSSQR$_{mean}$	-0.0182	-0.0219	-0.0188	-0.0673	-0.0402	-0.0086	-0.0015	-0.0001	0.0020
	BGLSSQR$_{median}$	-0.0178	-0.0215	-0.0184	-0.0668	-0.0428	-0.0075	-0.0012	0.0008	0.0006
	BSGSSQR$_{mean}$	-0.0207	-0.0243	-0.0180	-0.0760	-0.0801	-0.0042	0.0006	-0.0005	0.0023
	BSGSSQR$_{median}$	-0.0202	-0.0237	-0.0174	-0.0770	-0.0913	-0.0025	-0.0006	-0.0002	0.0010
0.90	CQR	0.0125	0.0062	0.0106	-0.0014	0.0019	-0.0044	0.0006	-0.0049	0.0063
	LASSOQR	-0.0786	-0.0864	-0.0724	-0.0727	-0.0535	-0.0028	0.0018	-0.0026	0.0021
	GLASSOQR	-0.0567	-0.0671	-0.0553	-0.0745	-0.0354	-0.0095	0.0034	-0.0076	0.0047
	BLASSOQR	-0.0276	-0.0344	-0.0274	-0.0353	-0.0294	-0.0060	0.0012	-0.0031	0.0067
	BGLSSQR$_{mean}$	-0.0181	-0.0220	-0.0188	-0.0676	-0.0403	-0.0085	-0.0016	0.0001	0.0020
	BGLSSQR$_{median}$	-0.0177	-0.0215	-0.0184	-0.0669	-0.0430	-0.0074	-0.0014	0.0009	0.0004
	BSGSSQR$_{mean}$	-0.0209	-0.0243	-0.0180	-0.0762	-0.0800	-0.0041	0.0006	-0.0005	0.0023
	BSGSSQR$_{median}$	-0.0203	-0.0236	-0.0174	-0.0773	-0.0916	-0.0025	-0.0005	-0.0004	0.0009

Table C.4: Estimated **RMSE** from eight methods in Study 5.1, assuming a t_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	CQR	0.1388	0.1401	0.1584	0.1486	0.1481	0.1618	0.1457	0.1416	0.1479
	LASSOQR	0.2990	0.3100	0.3230	0.2943	0.2599	0.2293	0.2309	0.2243	0.2201
	GLASSOQR	0.2952	0.2732	0.2948	0.3025	0.2511	0.2386	0.2423	0.2388	0.2299
	BLASSOQR	0.1259	0.1316	0.1458	0.1395	0.1313	0.1263	0.1114	0.1098	0.1151
	BGLSSQR$_{mean}$	0.1218	0.1308	0.1412	0.1813	0.1403	0.1265	0.0395	0.0346	0.0404
	BGLSSQR$_{median}$	0.1220	0.1309	0.1417	0.1895	0.1481	0.1252	0.0380	0.0267	0.0346
	BSGSSQR$_{mean}$	0.1234	0.1325	0.1417	0.1903	0.1638	0.0875	0.0437	0.0403	0.0450
	BSGSSQR$_{median}$	0.1235	0.1326	0.1422	0.2026	0.1853	0.0777	0.0375	0.0299	0.0372
0.50	CQR	0.1388	0.1401	0.1584	0.1486	0.1481	0.1618	0.1457	0.1416	0.1479
	LASSOQR	0.1589	0.1660	0.1753	0.1768	0.1643	0.1111	0.0908	0.0868	0.0908
	GLASSOQR	0.1500	0.1530	0.1645	0.2036	0.1482	0.1228	0.0929	0.0932	0.1017
	BLASSOQR	0.1259	0.1316	0.1458	0.1395	0.1313	0.1264	0.1114	0.1098	0.1151
	BGLSSQR$_{mean}$	0.1218	0.1308	0.1411	0.1811	0.1405	0.1265	0.0393	0.0346	0.0401
	BGLSSQR$_{median}$	0.1220	0.1311	0.1416	0.1893	0.1483	0.1253	0.0371	0.0269	0.0342
	BSGSSQR$_{mean}$	0.1232	0.1325	0.1418	0.1904	0.1638	0.0875	0.0440	0.0404	0.0447
	BSGSSQR$_{median}$	0.1233	0.1326	0.1423	0.2024	0.1852	0.0777	0.0378	0.0299	0.0368
0.90	CQR	0.1388	0.1401	0.1584	0.1486	0.1481	0.1618	0.1457	0.1416	0.1479
	LASSOQR	0.2989	0.3155	0.3201	0.2905	0.2593	0.2296	0.2238	0.2242	0.2343
	GLASSOQR	0.2838	0.2922	0.3014	0.2979	0.2552	0.2420	0.2285	0.2305	0.2513
	BLASSOQR	0.1259	0.1316	0.1458	0.1395	0.1313	0.1264	0.1114	0.1098	0.1151
	BGLSSQR$_{mean}$	0.1216	0.1308	0.1410	0.1812	0.1405	0.1265	0.0400	0.0345	0.0401
	BGLSSQR$_{median}$	0.1218	0.1310	0.1415	0.1890	0.1486	0.1253	0.0381	0.0266	0.0342
	BSGSSQR$_{mean}$	0.1233	0.1327	0.1417	0.1904	0.1638	0.0874	0.0438	0.0405	0.0446
	BSGSSQR$_{median}$	0.1234	0.1328	0.1421	0.2028	0.1854	0.0777	0.0377	0.0302	0.0366

Table C.5: Estimated **MBE** from eight methods in Study 5.1, assuming a χ^2_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	CQR	-0.0195	0.0014	0.0055	-0.0079	-0.0010	-0.0433	-0.0154	0.0014	0.0085
	LASSOQR	-0.0627	-0.0712	-0.0691	-0.0558	-0.0733	-0.0055	-0.0090	0.0012	-0.0025
	GLASSOQR	-0.0670	-0.0694	-0.0578	-0.0900	-0.0587	-0.0023	-0.0036	-0.0042	-0.0097
	BLASSOQR	-0.1307	-0.1218	-0.1092	-0.0942	-0.0632	-0.0290	-0.0187	-0.0038	0.0060
	BGLSSQR$_{mean}$	-0.0937	-0.0873	-0.0803	-0.2041	-0.1161	-0.0173	-0.0083	-0.0056	0.0042
	BGLSSQR$_{median}$	-0.0940	-0.0875	-0.0801	-0.2333	-0.1400	-0.0154	-0.0076	-0.0055	0.0027
	BSGSSQR$_{mean}$	-0.1249	-0.1167	-0.1099	-0.2167	-0.1514	-0.0138	-0.0103	-0.0052	0.0059
	BSGSSQR$_{median}$	-0.1235	-0.1147	-0.1077	-0.2571	-0.1883	-0.0082	-0.0063	-0.0043	0.0062
0.50	CQR	-0.0195	0.0014	0.0055	-0.0079	-0.0010	-0.0433	-0.0154	0.0014	0.0085
	LASSOQR	-0.1309	-0.1727	-0.1541	-0.1449	-0.1097	-0.0117	-0.0132	-0.0124	0.0112
	GLASSOQR	-0.1852	-0.1743	-0.1649	-0.2293	-0.1369	-0.0142	-0.0102	-0.0144	0.0117
	BLASSOQR	-0.1307	-0.1218	-0.1091	-0.0942	-0.0633	-0.0290	-0.0186	-0.0038	0.0060
	BGLSSQR$_{mean}$	-0.0935	-0.0872	-0.0804	-0.2038	-0.1159	-0.0173	-0.0084	-0.0058	0.0039
	BGLSSQR$_{median}$	-0.0938	-0.0873	-0.0802	-0.2331	-0.1407	-0.0152	-0.0076	-0.0055	0.0025
	BSGSSQR$_{mean}$	-0.1250	-0.1165	-0.1098	-0.2170	-0.1514	-0.0139	-0.0103	-0.0054	0.0059
	BSGSSQR$_{median}$	-0.1235	-0.1145	-0.1077	-0.2575	-0.1879	-0.0083	-0.0065	-0.0045	0.0060
0.90	CQR	-0.0195	0.0014	0.0055	-0.0079	-0.0010	-0.0433	-0.0154	0.0014	0.0085
	LASSOQR	-0.0181	-0.1984	-0.1988	-0.1238	-0.0680	-0.0386	-0.0057	0.0028	0.0288
	GLASSOQR	-0.0392	-0.0713	-0.0827	-0.1165	-0.0332	-0.0542	0.0048	0.0008	0.0230
	BLASSOQR	-0.1307	-0.1218	-0.1092	-0.0942	-0.0633	-0.0289	-0.0186	-0.0037	0.0061
	BGLSSQR$_{mean}$	-0.0933	-0.0868	-0.0803	-0.2039	-0.1160	-0.0168	-0.0084	-0.0056	0.0040
	BGLSSQR$_{median}$	-0.0936	-0.0869	-0.0800	-0.2334	-0.1403	-0.0151	-0.0076	-0.0055	0.0030
	BSGSSQR$_{mean}$	-0.1250	-0.1170	-0.1098	-0.2173	-0.1516	-0.0138	-0.0104	-0.0054	0.0059
	BSGSSQR$_{median}$	-0.1236	-0.1150	-0.1076	-0.2578	-0.1884	-0.0081	-0.0065	-0.0045	0.0060

Table C.6: Estimated **RMSE** from eight methods in Study 5.1, assuming a χ^2_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	CQR	0.3170	0.3055	0.3104	0.3019	0.3135	0.3284	0.3189	0.2968	0.2992
	LASSOQR	0.1802	0.1734	0.1810	0.1659	0.1749	0.1063	0.1114	0.1080	0.1117
	GLASSOQR	0.1682	0.1629	0.1577	0.1951	0.1569	0.1239	0.1125	0.1030	0.1122
	BLASSOQR	0.3135	0.2986	0.2918	0.2589	0.2379	0.2268	0.2117	0.1986	0.2059
	BGLSSQR$_{mean}$	0.2962	0.2756	0.2736	0.3292	0.2478	0.1857	0.1124	0.1034	0.1109
	BGLSSQR$_{median}$	0.2987	0.2782	0.2760	0.3729	0.2734	0.1800	0.0989	0.0893	0.0970
	BSGSSQR$_{mean}$	0.3418	0.3176	0.3165	0.3422	0.2487	0.1455	0.1074	0.0920	0.1085
	BSGSSQR$_{median}$	0.3479	0.3234	0.3209	0.3972	0.2837	0.1293	0.0940	0.0728	0.0943
0.50	CQR	0.3170	0.3055	0.3104	0.3019	0.3135	0.3284	0.3189	0.2968	0.2992
	LASSOQR	0.3040	0.3262	0.3189	0.3089	0.2561	0.1929	0.1700	0.1691	0.1715
	GLASSOQR	0.3268	0.3171	0.2981	0.3647	0.2616	0.1855	0.1514	0.1460	0.1418
	BLASSOQR	0.3135	0.2986	0.2917	0.2589	0.2379	0.2269	0.2117	0.1985	0.2059
	BGLSSQR$_{mean}$	0.2963	0.2759	0.2734	0.3293	0.2475	0.1856	0.1129	0.1040	0.1120
	BGLSSQR$_{median}$	0.2989	0.2784	0.2758	0.3732	0.2739	0.1796	0.0999	0.0901	0.0983
	BSGSSQR$_{mean}$	0.3416	0.3179	0.3164	0.3427	0.2485	0.1459	0.1074	0.0922	0.1090
	BSGSSQR$_{median}$	0.3476	0.3236	0.3210	0.3975	0.2834	0.1298	0.0940	0.0731	0.0952
0.90	CQR	0.3170	0.3055	0.3104	0.3019	0.3135	0.3284	0.3189	0.2968	0.2992
	LASSOQR	0.5142	0.6614	0.6616	0.5630	0.5176	0.4945	0.4904	0.4326	0.4739
	GLASSOQR	0.5653	0.6062	0.5620	0.5866	0.5594	0.5343	0.5579	0.4930	0.5065
	BLASSOQR	0.3135	0.2987	0.2917	0.2589	0.2378	0.2268	0.2117	0.1986	0.2059
	BGLSSQR$_{mean}$	0.2965	0.2756	0.2735	0.3291	0.2479	0.1854	0.1134	0.1041	0.1119
	BGLSSQR$_{median}$	0.2992	0.2780	0.2758	0.3731	0.2738	0.1796	0.1002	0.0903	0.0985
	BSGSSQR$_{mean}$	0.3418	0.3180	0.3165	0.3426	0.2486	0.1457	0.1074	0.0923	0.1084
	BSGSSQR$_{median}$	0.3479	0.3239	0.3212	0.3976	0.2836	0.1293	0.0940	0.0730	0.0943

Table C.7: Estimated **MBE** from four methods using different iteration numbers in the Monte Carlo EM method to estimate λ in Study 5.1, assuming a standard normal error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	BGLSSQR$_{mean,100}$	-0.0229	-0.0279	-0.0140	-0.0229	-0.0222	-0.0023	0.0008	-0.0001	-0.0004
	BGLSSQR$_{mean,1000}$	-0.0237	-0.0273	-0.0170	-0.0262	-0.0231	-0.0041	0.0009	0.0002	-0.0011
	BGLSSQR$_{median,100}$	-0.0228	-0.0277	-0.0137	-0.0213	-0.0225	-0.0024	0.0005	0.0007	-0.0004
	BGLSSQR$_{median,1000}$	-0.0236	-0.0271	-0.0168	-0.0243	-0.0235	-0.0044	-0.0002	0.0009	-0.0007
	BSGSSQR$_{mean,100}$	-0.0239	-0.0263	-0.0149	-0.0227	-0.0527	-0.0029	0.0009	0.0005	-0.0020
	BSGSSQR$_{mean,1000}$	-0.0239	-0.0263	-0.0149	-0.0227	-0.0527	-0.0029	0.0009	0.0005	-0.0020
	BSGSSQR$_{median,100}$	-0.0237	-0.0260	-0.0147	-0.0211	-0.0572	-0.0027	0.0011	0.0009	-0.0016
	BSGSSQR$_{median,1000}$	-0.0237	-0.0260	-0.0147	-0.0211	-0.0572	-0.0027	0.0012	0.0010	-0.0016
0.50	BGLSSQR$_{mean,100}$	-0.0229	-0.0279	-0.0140	-0.0232	-0.0222	-0.0022	0.0008	-0.0002	-0.0003
	BGLSSQR$_{mean,1000}$	-0.0238	-0.0275	-0.0171	-0.0264	-0.0232	-0.0042	0.0008	0.0000	-0.0010
	BGLSSQR$_{median,100}$	-0.0228	-0.0278	-0.0138	-0.0214	-0.0225	-0.0023	0.0005	0.0006	-0.0001
	BGLSSQR$_{median,1000}$	-0.0236	-0.0273	-0.0169	-0.0246	-0.0237	-0.0045	-0.0001	0.0008	-0.0006
	BSGSSQR$_{mean,100}$	-0.0240	-0.0262	-0.0150	-0.0227	-0.0527	-0.0026	0.0009	0.0005	-0.0020
	BSGSSQR$_{mean,1000}$	-0.0240	-0.0262	-0.0150	-0.0227	-0.0527	-0.0026	0.0009	0.0005	-0.0020
	BSGSSQR$_{median,100}$	-0.0238	-0.0259	-0.0147	-0.0211	-0.0570	-0.0025	0.0012	0.0009	-0.0015
	BSGSSQR$_{median,1000}$	-0.0238	-0.0259	-0.0148	-0.0211	-0.0570	-0.0025	0.0012	0.0009	-0.0015
0.90	BGLSSQR$_{mean,100}$	-0.0229	-0.0280	-0.0141	-0.0229	-0.0212	-0.0022	0.0008	-0.0002	-0.0004
	BGLSSQR$_{mean,1000}$	-0.0238	-0.0274	-0.0170	-0.0264	-0.0232	-0.0039	0.0009	0.0002	-0.0009
	BGLSSQR$_{median,100}$	-0.0228	-0.0278	-0.0139	-0.0213	-0.0224	-0.0024	0.0004	0.0006	-0.0002
	BGLSSQR$_{median,1000}$	-0.0237	-0.0272	-0.0169	-0.0245	-0.0236	-0.0042	-0.0003	0.0011	-0.0008
	BSGSSQR$_{mean,100}$	-0.0240	-0.0263	-0.0149	-0.0228	-0.0527	-0.0028	0.0009	0.0004	-0.0019
	BSGSSQR$_{mean,1000}$	-0.0240	-0.0263	-0.0150	-0.0228	-0.0527	-0.0028	0.0009	0.0004	-0.0019
	BSGSSQR$_{median,100}$	-0.0238	-0.0259	-0.0147	-0.0211	-0.0571	-0.0026	0.0012	0.0010	-0.0015
	BSGSSQR$_{median,1000}$	-0.0238	-0.0259	-0.0147	-0.0211	-0.0571	-0.0026	0.0012	0.0009	-0.0015

Table C.8: Estimated **RMSE** from four methods using different iteration numbers in the Monte Carlo EM method to estimate λ in Study 5.1, assuming a standard normal error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	BGLSSQR$_{mean,100}$	0.1148	0.1200	0.1188	0.1281	0.1188	0.1055	0.0407	0.0361	0.0400
	BGLSSQR$_{mean,1000}$	0.1150	0.1206	0.1177	0.1338	0.1218	0.1065	0.0417	0.0364	0.0433
	BGLSSQR$_{median,100}$	0.1154	0.1206	0.1193	0.1276	0.1209	0.1058	0.0377	0.0303	0.0357
	BGLSSQR$_{median,1000}$	0.1156	0.1213	0.1181	0.1330	0.1241	0.1067	0.0384	0.0303	0.0393
	BSGSSQR$_{mean,100}$	0.1170	0.1207	0.1177	0.1359	0.1439	0.0676	0.0482	0.0431	0.0500
	BSGSSQR$_{mean,1000}$	0.1170	0.1207	0.1177	0.1359	0.1439	0.0676	0.0482	0.0431	0.0500
	BSGSSQR$_{median,100}$	0.1176	0.1214	0.1181	0.1367	0.1562	0.0588	0.0439	0.0371	0.0455
	BSGSSQR$_{median,1000}$	0.1176	0.1214	0.1181	0.1367	0.1562	0.0588	0.0439	0.0371	0.0455
0.50	BGLSSQR$_{mean,100}$	0.1148	0.1199	0.1188	0.1279	0.1189	0.1056	0.0406	0.0362	0.0406
	BGLSSQR$_{mean,1000}$	0.1149	0.1207	0.1175	0.1340	0.1219	0.1065	0.0418	0.0366	0.0432
	BGLSSQR$_{median,100}$	0.1153	0.1253	0.1193	0.1275	0.1209	0.1058	0.0378	0.0308	0.0360
	BGLSSQR$_{median,1000}$	0.1154	0.1213	0.1180	0.1332	0.1244	0.1068	0.0385	0.0303	0.0393
	BSGSSQR$_{mean,100}$	0.1171	0.1207	0.1178	0.1357	0.1439	0.0675	0.0482	0.0434	0.0501
	BSGSSQR$_{mean,1000}$	0.1171	0.1207	0.1178	0.1357	0.1439	0.0675	0.0482	0.0434	0.0501
	BSGSSQR$_{median,100}$	0.1176	0.1214	0.1182	0.1366	0.1560	0.0586	0.0439	0.0375	0.0454
	BSGSSQR$_{median,1000}$	0.1176	0.1214	0.1182	0.1366	0.1559	0.0586	0.0439	0.0375	0.0454
0.90	BGLSSQR$_{mean,100}$	0.1147	0.1199	0.1188	0.1280	0.1188	0.1056	0.0407	0.0361	0.0404
	BGLSSQR$_{mean,1000}$	0.1151	0.1206	0.1175	0.1339	0.1219	0.1064	0.0419	0.0366	0.0434
	BGLSSQR$_{median,100}$	0.1153	0.1206	0.1193	0.1276	0.1208	0.1058	0.0378	0.0304	0.0362
	BGLSSQR$_{median,1000}$	0.1156	0.1212	0.1179	0.1330	0.1242	0.1067	0.0386	0.0304	0.0390
	BSGSSQR$_{mean,100}$	0.1171	0.1207	0.1178	0.1359	0.1441	0.0675	0.0481	0.0433	0.0500
	BSGSSQR$_{mean,1000}$	0.1170	0.1207	0.1178	0.1359	0.1440	0.0675	0.0481	0.0433	0.0500
	BSGSSQR$_{median,100}$	0.1176	0.1214	0.1182	0.1367	0.1563	0.0586	0.0439	0.0372	0.0453
	BSGSSQR$_{median,1000}$	0.1176	0.1214	0.1182	0.1367	0.1563	0.0586	0.0440	0.0373	0.0454

Table C.9: Estimated **MBE** from four methods using different iteration numbers in the Monte Carlo EM method to estimate λ in Study 5.1, assuming a t_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	BGLSSQR$_{mean,100}$	-0.0182	-0.0218	-0.0188	-0.0674	-0.0401	-0.0086	-0.0015	0.0000	0.0021
	BGLSSQR$_{mean,1000}$	-0.0126	-0.0204	-0.0143	-0.0678	-0.0361	-0.0054	-0.0007	-0.0006	0.0026
	BGLSSQR$_{median,100}$	-0.0178	-0.0213	-0.0184	-0.0670	-0.0426	-0.0074	-0.0013	0.0007	0.0005
	BGLSSQR$_{median,1000}$	-0.0122	-0.0200	-0.0139	-0.0682	-0.0390	-0.0040	-0.0009	0.0003	0.0018
	BSGSSQR$_{mean,100}$	-0.0208	-0.0242	-0.0182	-0.0761	-0.0800	-0.0042	0.0006	-0.0005	0.0023
	BSGSSQR$_{mean,1000}$	0.0168	-0.0259	-0.0231	-0.0687	-0.0813	-0.0058	-0.0011	-0.0003	0.0032
	BSGSSQR$_{median,100}$	-0.0203	-0.0236	-0.0176	-0.0773	-0.0915	-0.0024	-0.0005	-0.0004	0.0010
	BSGSSQR$_{median,1000}$	-0.0162	-0.0252	-0.0225	-0.0692	-0.0928	-0.0050	-0.0016	0.0002	0.0014
0.50	BGLSSQR$_{mean,100}$	-0.0182	-0.0219	-0.0188	-0.0673	-0.0402	-0.0086	-0.0015	-0.0001	0.0020
	BGLSSQR$_{mean,1000}$	-0.0126	-0.0201	-0.0141	-0.0678	-0.0361	-0.0054	-0.0007	-0.0005	0.0025
	BGLSSQR$_{median,100}$	-0.0178	-0.0215	-0.0184	-0.0668	-0.0428	-0.0075	-0.0012	0.0008	0.0006
	BGLSSQR$_{median,1000}$	-0.0123	-0.0197	-0.0137	-0.0683	-0.0389	-0.0040	-0.0008	0.0004	0.0018
	BSGSSQR$_{mean,100}$	-0.0207	-0.0243	-0.0180	-0.0760	-0.0801	-0.0042	0.0006	-0.0005	0.0023
	BSGSSQR$_{mean,1000}$	-0.0167	-0.0258	-0.0232	-0.0689	-0.0812	-0.0058	-0.0010	-0.0002	0.0032
	BSGSSQR$_{median,100}$	-0.0202	-0.0237	-0.0174	-0.0770	-0.0913	-0.0025	-0.0006	-0.0002	0.0010
	BSGSSQR$_{median,1000}$	-0.0161	-0.0251	-0.0225	-0.0695	-0.0927	-0.0049	-0.0016	0.0003	0.0014
0.90	BGLSSQR$_{mean,100}$	-0.0181	-0.0220	-0.0188	-0.0676	-0.0403	-0.0085	-0.0016	0.0001	0.0020
	BGLSSQR$_{mean,1000}$	-0.0125	-0.0200	-0.0141	-0.0680	-0.0363	-0.0055	-0.0006	-0.0004	0.0025
	BGLSSQR$_{median,100}$	-0.0177	-0.0215	-0.0184	-0.0669	-0.0430	-0.0074	-0.0014	0.0009	0.0004
	BGLSSQR$_{median,1000}$	-0.0122	-0.0197	-0.0137	-0.0685	-0.0393	-0.0042	-0.0008	0.0005	0.0017
	BSGSSQR$_{mean,100}$	-0.0209	-0.0243	-0.0180	-0.0762	-0.0800	-0.0041	0.0006	-0.0005	0.0023
	BSGSSQR$_{mean,1000}$	-0.0167	-0.0257	-0.0231	-0.0689	-0.0814	-0.0058	-0.0011	-0.0002	0.0031
	BSGSSQR$_{median,100}$	-0.0203	-0.0236	-0.0174	-0.0773	-0.0916	-0.0025	-0.0005	-0.0004	0.0009
	BSGSSQR$_{median,1000}$	-0.0161	-0.0250	-0.0225	-0.0696	-0.0930	-0.0050	-0.0016	0.0003	0.0012

Table C.10: Estimated **RMSE** from four methods using different iteration numbers in the Monte Carlo EM method to estimate λ in Study 5.1, assuming a t_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	BGLSSQR$_{mean,100}$	0.1218	0.1308	0.1412	0.1813	0.1403	0.1265	0.0395	0.0346	0.0404
	BGLSSQR$_{mean,1000}$	0.1210	0.1280	0.1388	0.1832	0.1383	0.1270	0.0387	0.0310	0.0344
	BGLSSQR$_{median,100}$	0.1220	0.1309	0.1417	0.1895	0.1481	0.1252	0.0380	0.0267	0.0346
	BGLSSQR$_{median,1000}$	0.1213	0.1283	0.1394	0.1931	0.1468	0.1258	0.0369	0.0217	0.0276
	BSGSSQR$_{mean,100}$	0.1234	0.1325	0.1417	0.1903	0.1638	0.0875	0.0437	0.0403	0.0450
	BSGSSQR$_{mean,1000}$	0.1219	0.1295	0.1439	0.1853	0.1639	0.0875	0.0477	0.0387	0.0459
	BSGSSQR$_{median,100}$	0.1235	0.1326	0.1422	0.2026	0.1853	0.0777	0.0375	0.0299	0.0372
	BSGSSQR$_{median,1000}$	0.1222	0.1296	0.1442	0.1967	0.1854	0.0780	0.0423	0.0278	0.0385
0.50	BGLSSQR$_{mean,100}$	0.1218	0.1308	0.1411	0.1811	0.1405	0.1265	0.0393	0.0346	0.0401
	BGLSSQR$_{mean,1000}$	0.1209	0.1279	0.1386	0.1835	0.1385	0.1272	0.0387	0.0310	0.0344
	BGLSSQR$_{median,100}$	0.1220	0.1311	0.1416	0.1893	0.1483	0.1253	0.0371	0.0269	0.0342
	BGLSSQR$_{median,1000}$	0.1213	0.1282	0.1391	0.1933	0.1468	0.1260	0.0367	0.0215	0.0274
	BSGSSQR$_{mean,100}$	0.1232	0.1325	0.1418	0.1904	0.1638	0.0875	0.0440	0.0404	0.0447
	BSGSSQR$_{mean,1000}$	0.1218	0.1294	0.1437	0.1853	0.1639	0.0874	0.0476	0.0388	0.0458
	BSGSSQR$_{median,100}$	0.1233	0.1326	0.1423	0.2024	0.1852	0.0777	0.0378	0.0299	0.0368
	BSGSSQR$_{median,1000}$	0.1221	0.1296	0.1441	0.1969	0.1854	0.0779	0.0421	0.0278	0.0384
0.90	BGLSSQR$_{mean,100}$	0.1216	0.1308	0.1410	0.1812	0.1405	0.1265	0.0400	0.0345	0.0401
	BGLSSQR$_{mean,1000}$	0.1211	0.1280	0.1385	0.1837	0.1385	0.1270	0.0388	0.0310	0.0342
	BGLSSQR$_{median,100}$	0.1218	0.1310	0.1415	0.1890	0.1486	0.1253	0.0381	0.0266	0.0342
	BGLSSQR$_{median,1000}$	0.1214	0.1283	0.1391	0.1937	0.1473	0.1258	0.0370	0.0221	0.0275
	BSGSSQR$_{mean,100}$	0.1233	0.1327	0.1417	0.1904	0.1638	0.0874	0.0438	0.0405	0.0446
	BSGSSQR$_{mean,1000}$	0.1219	0.1295	0.1439	0.1855	0.1641	0.0875	0.0474	0.0388	0.0458
	BSGSSQR$_{median,100}$	0.1234	0.1328	0.1421	0.2028	0.1854	0.0777	0.0377	0.0302	0.0366
	BSGSSQR$_{median,1000}$	0.1221	0.1297	0.1443	0.1974	0.1855	0.0780	0.0418	0.0277	0.0383

Table C.11: Estimated **MBE** from four methods using different iteration numbers in the Monte Carlo EM method to estimate λ in Study 5.1, assuming a χ_3^2 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	BGLSSQR$_{mean,100}$	-0.0937	-0.0873	-0.0803	-0.2041	-0.1161	-0.0173	-0.0083	-0.0056	0.0042
	BGLSSQR$_{mean,1000}$	-0.0929	-0.0883	-0.0830	-0.2054	-0.1183	-0.0210	-0.0100	-0.0054	0.0046
	BGLSSQR$_{median,100}$	-0.0940	-0.0875	-0.0801	-0.2333	-0.1400	-0.0154	-0.0076	-0.0055	0.0027
	BGLSSQR$_{median,1000}$	-0.0933	-0.0884	-0.0831	-0.2361	-0.1430	-0.0184	-0.0077	-0.0048	0.0041
	BSGSSQR$_{mean,100}$	-0.1249	-0.1167	-0.1099	-0.2167	-0.1514	-0.0138	-0.0103	-0.0052	0.0059
	BSGSSQR$_{mean,1000}$	-0.1241	-0.1181	-0.1104	-0.2143	-0.1449	-0.0147	-0.0109	-0.0056	0.0022
	BSGSSQR$_{median,100}$	-0.1235	-0.1147	-0.1077	-0.2571	-0.1883	-0.0082	-0.0063	-0.0043	0.0062
	BSGSSQR$_{median,1000}$	-0.1228	-0.1165	-0.1083	-0.2541	-0.1807	-0.0078	-0.0058	-0.0040	0.0033
0.50	BGLSSQR$_{mean,100}$	-0.0935	-0.0872	-0.0804	-0.2038	-0.1159	-0.0173	-0.0084	-0.0058	0.0039
	BGLSSQR$_{mean,1000}$	-0.0933	-0.0884	-0.0831	-0.2054	-0.1183	-0.0210	-0.0098	-0.0055	0.0043
	BGLSSQR$_{median,100}$	-0.0938	-0.0873	-0.0802	-0.2331	-0.1407	-0.0152	-0.0076	-0.0055	0.0025
	BGLSSQR$_{median,1000}$	-0.0937	-0.0885	-0.0832	-0.2357	-0.1428	-0.0184	-0.0072	-0.0046	0.0041
	BSGSSQR$_{mean,100}$	-0.1250	-0.1165	-0.1098	-0.2170	-0.1514	-0.0139	-0.0103	-0.0054	0.0059
	BSGSSQR$_{mean,1000}$	-0.1240	-0.1183	-0.1105	-0.2145	-0.1447	-0.0145	-0.0110	-0.0055	0.0022
	BSGSSQR$_{median,100}$	-0.1235	-0.1145	-0.1077	-0.2575	-0.1879	-0.0083	-0.0065	-0.0045	0.0060
	BSGSSQR$_{median,1000}$	-0.1228	-0.1167	-0.1083	-0.2544	-0.1808	-0.0077	-0.0058	-0.0040	0.0034
0.90	BGLSSQR$_{mean,100}$	-0.0933	-0.0868	-0.0803	-0.2039	-0.1160	-0.0168	-0.0084	-0.0056	0.0040
	BGLSSQR$_{mean,1000}$	-0.0931	-0.0886	-0.0832	-0.2056	-0.1185	-0.0211	-0.0098	-0.0052	0.0042
	BGLSSQR$_{median,100}$	-0.0936	-0.0869	-0.0800	-0.2334	-0.1403	-0.0151	-0.0076	-0.0055	0.0030
	BGLSSQR$_{median,1000}$	-0.0934	-0.0885	-0.0833	-0.2360	-0.1432	-0.0181	-0.0070	-0.0045	0.0041
	BSGSSQR$_{mean,100}$	-0.1250	-0.1170	-0.1098	-0.2173	-0.1516	-0.0138	-0.0104	-0.0054	0.0059
	BSGSSQR$_{mean,1000}$	-0.1241	-0.1181	-0.1106	-0.2140	-0.1448	-0.0145	-0.0109	-0.0055	0.0021
	BSGSSQR$_{median,100}$	-0.1236	-0.1150	-0.1076	-0.2578	-0.1884	-0.0081	-0.0065	-0.0045	0.0060
	BSGSSQR$_{median,1000}$	-0.1229	-0.1165	-0.1083	-0.2537	-0.1811	-0.0077	-0.0058	-0.0041	0.0032

Table C.12: Estimated **RMSE** from four methods using different iteration numbers in the Monte Carlo EM method to estimate λ in Study 5.1, assuming a χ^2_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
0.10	BGLSSQR$_{mean,100}$	0.2962	0.2756	0.2736	0.3292	0.2478	0.1857	0.1124	0.1034	0.1109
	BGLSSQR$_{mean,1000}$	0.2938	0.2760	0.2725	0.3310	0.2486	0.1906	0.1094	0.1031	0.1074
	BGLSSQR$_{median,100}$	0.2987	0.2782	0.2760	0.3729	0.2734	0.1800	0.0989	0.0893	0.0970
	BGLSSQR$_{median,1000}$	0.2962	0.2785	0.2749	0.3762	0.2750	0.1858	0.0965	0.0898	0.0952
	BSGSSQR$_{mean,100}$	0.3418	0.3176	0.3165	0.3422	0.2487	0.1455	0.1074	0.0920	0.1085
	BSGSSQR$_{mean,1000}$	0.3401	0.3198	0.3155	0.3425	0.2493	0.1445	0.1107	0.0934	0.1060
	BSGSSQR$_{median,100}$	0.3479	0.3234	0.3209	0.3972	0.2837	0.1293	0.0940	0.0728	0.0943
	BSGSSQR$_{median,1000}$	0.3466	0.3262	0.3205	0.3966	0.2841	0.1273	0.0965	0.0746	0.0896
0.50	BGLSSQR$_{mean,100}$	0.2963	0.2759	0.2734	0.3293	0.2475	0.1856	0.1129	0.1040	0.1120
	BGLSSQR$_{mean,1000}$	0.2938	0.2762	0.2726	0.3309	0.2483	0.1904	0.1089	0.1029	0.1078
	BGLSSQR$_{median,100}$	0.2989	0.2784	0.2758	0.3732	0.2739	0.1796	0.0999	0.0901	0.0983
	BGLSSQR$_{median,1000}$	0.2963	0.2787	0.2750	0.3756	0.2743	0.1854	0.0958	0.0895	0.0953
	BSGSSQR$_{mean,100}$	0.3416	0.3179	0.3164	0.3427	0.2485	0.1459	0.1074	0.0922	0.1090
	BSGSSQR$_{mean,1000}$	0.3400	0.3200	0.3162	0.3426	0.2489	0.1443	0.1105	0.0935	0.1057
	BSGSSQR$_{median,100}$	0.3476	0.3236	0.3210	0.3975	0.2834	0.1298	0.0940	0.0731	0.0952
	BSGSSQR$_{median,1000}$	0.3465	0.3262	0.3212	0.3969	0.2837	0.1270	0.0963	0.0744	0.0891
0.90	BGLSSQR$_{mean,100}$	0.2965	0.2756	0.2735	0.3291	0.2479	0.1854	0.1134	0.1041	0.1119
	BGLSSQR$_{mean,1000}$	0.2936	0.2758	0.2722	0.3310	0.2483	0.1904	0.1093	0.1029	0.1075
	BGLSSQR$_{median,100}$	0.2992	0.2780	0.2758	0.3731	0.2738	0.1796	0.1002	0.0903	0.0985
	BGLSSQR$_{median,1000}$	0.2960	0.2782	0.2746	0.3760	0.2745	0.1853	0.0958	0.0895	0.0951
	BSGSSQR$_{mean,100}$	0.3418	0.3180	0.3165	0.3426	0.2486	0.1457	0.1074	0.0923	0.1084
	BSGSSQR$_{mean,1000}$	0.3400	0.3199	0.3162	0.3425	0.2489	0.1443	0.1107	0.0935	0.1058
	BSGSSQR$_{median,100}$	0.3479	0.3239	0.3212	0.3976	0.2836	0.1293	0.0940	0.0730	0.0943
	BSGSSQR$_{median,1000}$	0.3465	0.3262	0.3211	0.3965	0.2840	0.1269	0.0965	0.0745	0.0892

Table C.13: Estimated **MBE** from eight methods in Study 5.2, assuming a **standard normal** error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
0.10	CQR	-0.0017	0.0172	0.0036	-0.0091	-0.0118	0.0129	-0.0039	-0.0040	0.0038	0.0019	0.0119
	LASSOQR	0.4983	0.0865	-0.2089	-0.0489	-0.0538	0.0040	-0.0089	0.0011	0.0079	-0.4663	-0.2344
	GLASSOQR	0.1862	0.0795	-0.0470	-0.0424	-0.0396	0.0050	-0.0057	-0.0051	0.0030	-0.2245	-0.0928
	BLASSOQR	0.1868	0.0320	-0.0806	-0.0303	-0.0266	0.0084	-0.0004	-0.0049	0.0039	-0.1807	-0.0936
	BGLSSQR$_{mean}$	0.1865	0.0319	-0.0724	-0.0309	-0.0266	0.0091	0.0021	-0.0036	0.0018	-0.1780	-0.0816
	BGLSSQR$_{median}$	0.1791	0.0316	-0.0699	-0.0302	-0.0270	0.0093	0.0027	-0.0028	0.0027	-0.1729	-0.0797
	BSGSSQR$_{mean}$	0.2248	0.0265	-0.1101	-0.0362	-0.0577	0.0020	0.0020	-0.0055	0.0030	-0.1990	-0.1039
	BSGSSQR$_{median}$	0.2171	0.0260	-0.1291	-0.0353	-0.0613	-0.0001	0.0027	-0.0046	0.0028	-0.1925	-0.1026
0.50	CQR	-0.0017	0.0172	0.0036	-0.0091	-0.0118	0.0129	-0.0039	-0.0040	0.0038	0.0019	0.0119
	LASSOQR	0.5181	0.0419	-0.2252	-0.0625	-0.0625	0.0025	0.0044	-0.0051	0.0023	-0.4852	-0.2708
	GLASSOQR	0.1714	0.0635	-0.0469	-0.0632	-0.0395	0.0070	0.0036	-0.0085	-0.0009	-0.2315	-0.1094
	BLASSOQR	0.1868	0.0319	-0.0806	-0.0303	-0.0265	0.0083	-0.0004	-0.0049	0.0039	-0.1807	-0.0936
	BGLSSQR$_{mean}$	0.1872	0.0319	-0.0725	-0.0309	-0.0266	0.0090	0.0024	-0.0035	0.0020	-0.1787	-0.0820
	BGLSSQR$_{median}$	0.1803	0.0316	-0.0699	-0.0304	-0.0270	0.0093	0.0025	-0.0025	0.0025	-0.1741	-0.0802
	BSGSSQR$_{mean}$	0.2244	0.0262	-0.1099	-0.0362	-0.0576	0.0020	0.0019	-0.0055	0.0030	-0.1987	-0.1039
	BSGSSQR$_{median}$	0.2176	0.0257	-0.1292	-0.0353	-0.0612	-0.0001	0.0026	-0.0046	0.0029	-0.1928	-0.1029
0.90	CQR	-0.0017	0.0172	0.0036	-0.0091	-0.0118	0.0129	-0.0039	-0.0040	0.0038	0.0019	0.0119
	LASSOQR	0.4954	-0.0104	-0.2032	-0.0529	-0.0456	0.0042	0.0041	-0.0041	0.0060	-0.4702	-0.2763
	GLASSOQR	0.1552	0.0279	-0.0515	-0.0542	-0.0240	0.0082	-0.0057	0.0014	0.0002	-0.1798	-0.0943
	BLASSOQR	0.1868	0.0320	-0.0806	-0.0303	-0.0265	0.0083	-0.0004	-0.0049	0.0038	-0.1807	-0.0935
	BGLSSQR$_{mean}$	0.1854	0.0319	-0.0719	-0.0309	-0.0267	0.0091	0.0023	-0.0037	0.0019	-0.1770	-0.0812
	BGLSSQR$_{median}$	0.1778	0.0317	-0.0692	-0.0303	-0.0271	0.0094	0.0028	-0.0027	0.0027	-0.1719	-0.0792
	BSGSSQR$_{mean}$	0.2234	0.0264	-0.1095	-0.0362	-0.0576	0.0020	0.0020	-0.0055	0.0030	-0.1979	-0.1035
	BSGSSQR$_{median}$	0.2148	0.0260	-0.1282	-0.0352	-0.0611	-0.0002	0.0027	-0.0046	0.0027	-0.1907	-0.1019

Table C.14: Estimated **RMSE** from eight methods in Study 5.2, assuming a **standard normal** error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
0.10	CQR	0.5695	0.1778	0.3209	0.1232	0.1267	0.1310	0.1350	0.1251	0.1342	0.4558	0.2747
	LASSOQR	0.7459	0.2407	0.3436	0.1885	0.1871	0.1467	0.1479	0.1440	0.1469	0.7027	0.4139
	GLASSOQR	0.4826	0.2187	0.2622	0.1704	0.1646	0.1488	0.1359	0.1402	0.1366	0.4782	0.2916
	BLASSOQR	0.4426	0.1529	0.2307	0.1132	0.1153	0.1054	0.1107	0.1016	0.1067	0.3961	0.2456
	BGLSSQR$_{mean}$	0.4539	0.1522	0.2451	0.1239	0.1203	0.1133	0.0486	0.0472	0.0431	0.3981	0.2430
	BGLSSQR$_{median}$	0.4489	0.1526	0.2444	0.1262	0.1226	0.1136	0.0438	0.0448	0.0377	0.3942	0.2430
	BSGSSQR$_{mean}$	0.5460	0.1582	0.2634	0.1259	0.1424	0.0805	0.0532	0.0512	0.0476	0.4855	0.2928
	BSGSSQR$_{median}$	0.5713	0.1591	0.2857	0.1273	0.1532	0.0746	0.0473	0.0476	0.0397	0.4938	0.3001
0.50	CQR	0.5695	0.1778	0.3209	0.1232	0.1267	0.1310	0.1350	0.1251	0.1342	0.4558	0.2747
	LASSOQR	0.6810	0.1837	0.3045	0.1439	0.1509	0.0939	0.1002	0.0904	0.0959	0.6579	0.3976
	GLASSOQR	0.3861	0.1743	0.2091	0.1387	0.1244	0.1070	0.1050	0.0964	0.1022	0.4248	0.2527
	BLASSOQR	0.4426	0.1529	0.2307	0.1132	0.1153	0.1054	0.1107	0.1016	0.1066	0.3961	0.2456
	BGLSSQR$_{mean}$	0.4555	0.1524	0.2457	0.1239	0.1202	0.1134	0.0491	0.0473	0.0431	0.3987	0.2433
	BGLSSQR$_{median}$	0.4505	0.1529	0.2449	0.1266	0.1225	0.1137	0.0446	0.0445	0.0377	0.3948	0.2432
	BSGSSQR$_{mean}$	0.5469	0.1585	0.2636	0.1259	0.1425	0.0805	0.0532	0.0509	0.0477	0.4861	0.2931
	BSGSSQR$_{median}$	0.5727	0.1592	0.2861	0.1272	0.1531	0.0746	0.0472	0.0474	0.0403	0.4954	0.3009
0.90	CQR	0.5695	0.1778	0.3209	0.1232	0.1267	0.1310	0.1350	0.1251	0.1342	0.4558	0.2747
	LASSOQR	0.7694	0.2405	0.3595	0.1945	0.1793	0.1437	0.1524	0.1473	0.1406	0.7419	0.4640
	GLASSOQR	0.4505	0.2140	0.2622	0.1669	0.1493	0.1456	0.1523	0.1359	0.1441	0.4507	0.2907
	BLASSOQR	0.4426	0.1529	0.2307	0.1132	0.1154	0.1054	0.1107	0.1016	0.1066	0.3960	0.2456
	BGLSSQR$_{mean}$	0.4524	0.1527	0.2449	0.1242	0.1203	0.1133	0.0487	0.0472	0.0430	0.3964	0.2425
	BGLSSQR$_{median}$	0.4470	0.1530	0.2438	0.1266	0.1227	0.1135	0.0439	0.0447	0.0382	0.3928	0.2426
	BSGSSQR$_{mean}$	0.5455	0.1588	0.2633	0.1260	0.1424	0.0805	0.0532	0.0512	0.0481	0.4855	0.2928
	BSGSSQR$_{median}$	0.5709	0.1595	0.2858	0.1273	0.1530	0.0745	0.0471	0.0477	0.0406	0.4936	0.3000

Table C.15: Estimated **MBE** from eight methods in Study 5.2, assuming a t_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
0.10	CQR	0.0323	0.0025	-0.0211	0.0121	0.0012	-0.0008	0.0000	-0.0049	0.0018	-0.0259	0.0022
	LASSOQR	0.7362	0.0803	-0.2777	-0.0792	-0.0681	-0.0267	0.0135	-0.0001	0.0017	-0.7727	-0.4099
	GLASSOQR	0.2774	0.0684	-0.0606	-0.0495	-0.0420	-0.0174	0.0015	-0.0042	-0.0017	-0.3895	-0.1806
	BLASSOQR	0.3021	0.0341	-0.1306	-0.0274	-0.0281	-0.0012	-0.0011	-0.0027	0.0026	-0.3061	-0.1554
	BGLSSQR$_{mean}$	0.2866	0.0372	-0.1227	-0.0585	-0.0374	-0.0020	0.0015	-0.0003	0.0003	-0.2834	-0.1267
	BGLSSQR$_{median}$	0.2711	0.0363	-0.1172	-0.0596	-0.0412	-0.0023	0.0006	-0.0004	-0.0002	-0.2735	-0.1244
	BSGSSQR$_{mean}$	0.4290	0.0296	-0.1943	-0.0675	-0.0827	-0.0016	-0.0004	0.0003	0.0009	-0.4073	-0.2185
	BSGSSQR$_{median}$	0.4187	0.0275	-0.2140	-0.0678	-0.0961	-0.0005	0.0005	0.0011	-0.0003	-0.4044	-0.2237
0.50	CQR	0.0323	0.0025	-0.0211	0.0121	0.0012	-0.0008	0.0000	-0.0049	0.0018	-0.0259	0.0022
	LASSOQR	0.6458	0.0513	-0.2625	-0.0680	-0.0727	-0.0056	0.0015	0.0020	0.0034	-0.6579	-0.3647
	GLASSOQR	0.2353	0.0695	-0.0713	-0.0709	-0.0435	-0.0045	-0.0008	0.0008	0.0047	-0.3299	-0.1572
	BLASSOQR	0.3020	0.0341	-0.1306	-0.0274	-0.0281	-0.0012	-0.0011	-0.0027	0.0026	-0.3060	-0.1554
	BGLSSQR$_{mean}$	0.2876	0.0369	-0.1228	-0.0589	-0.0378	-0.0020	0.0014	-0.0004	0.0003	-0.2840	-0.1272
	BGLSSQR$_{median}$	0.2716	0.0357	-0.1175	-0.0610	-0.0423	-0.0023	0.0008	-0.0006	-0.0002	-0.2739	-0.1245
	BSGSSQR$_{mean}$	0.4301	0.0293	-0.1945	-0.0677	-0.0826	-0.0016	-0.0003	0.0003	0.0010	-0.4086	-0.2193
	BSGSSQR$_{median}$	0.4204	0.0272	-0.2146	-0.0681	-0.0964	-0.0005	0.0006	0.0010	-0.0002	-0.4058	-0.2244
0.90	CQR	0.0323	0.0025	-0.0211	0.0121	0.0012	-0.0008	0.0000	-0.0049	0.0018	-0.0259	0.0022
	LASSOQR	0.7605	0.0118	-0.3008	-0.0701	-0.0589	0.0054	-0.0058	0.0136	-0.0060	-0.7783	-0.4581
	GLASSOQR	0.2725	0.0224	-0.0901	-0.0288	-0.0112	0.0041	-0.0017	0.0078	0.0113	-0.3312	-0.1976
	BLASSOQR	0.3020	0.0341	-0.1306	-0.0273	-0.0281	-0.0012	-0.0011	-0.0027	0.0026	-0.3060	-0.1555
	BGLSSQR$_{mean}$	0.2871	0.0375	-0.1230	-0.0591	-0.0377	-0.0019	0.0017	-0.0004	0.0001	-0.2834	-0.1268
	BGLSSQR$_{median}$	0.2714	0.0365	-0.1175	-0.0611	-0.0417	-0.0021	0.0010	-0.0004	-0.0004	-0.2738	-0.1245
	BSGSSQR$_{mean}$	0.4283	0.0295	-0.1940	-0.0677	-0.0827	-0.0016	-0.0003	0.0003	0.0010	-0.4067	-0.2183
	BSGSSQR$_{median}$	0.4185	0.0274	-0.2136	-0.0679	-0.0964	-0.0005	0.0007	0.0010	-0.0002	-0.4032	-0.2224

Table C.16: Estimated **RMSE** from eight methods in Study 5.2, assuming a t_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
0.10	CQR	0.7147	0.2149	0.4219	0.1527	0.1538	0.1570	0.1609	0.1522	0.1494	0.5579	0.3306
	LASSOQR	1.0779	0.4407	0.5103	0.3021	0.2658	0.2302	0.2298	0.2228	0.2253	1.1308	0.6469
	GLASSOQR	0.6975	0.3916	0.4312	0.2869	0.2482	0.2571	0.2376	0.2402	0.2279	0.8241	0.5031
	BLASSOQR	0.5361	0.1867	0.2737	0.1399	0.1327	0.1206	0.1231	0.1184	0.1165	0.5017	0.3014
	BGLSSQR$_{mean}$	0.5678	0.1931	0.2981	0.1799	0.1430	0.1214	0.0539	0.0434	0.0399	0.5160	0.3007
	BGLSSQR$_{median}$	0.5612	0.1945	0.2949	0.1929	0.1536	0.1199	0.0482	0.0365	0.0317	0.5155	0.3052
	BSGSSQR$_{mean}$	0.7539	0.1987	0.3096	0.1849	0.1654	0.0769	0.0536	0.0419	0.0435	0.7169	0.4106
	BSGSSQR$_{median}$	0.7826	0.1981	0.3199	0.1964	0.1888	0.0631	0.0450	0.0285	0.0324	0.7526	0.4369
0.50	CQR	0.7147	0.2149	0.4219	0.1527	0.1538	0.1570	0.1609	0.1522	0.1494	0.5579	0.3306
	LASSOQR	0.7989	0.2009	0.3276	0.1709	0.1710	0.0984	0.0989	0.0954	0.0976	0.8270	0.4823
	GLASSOQR	0.4507	0.2081	0.2479	0.1756	0.1401	0.1216	0.1195	0.1091	0.1057	0.5147	0.3033
	BLASSOQR	0.5361	0.1867	0.2737	0.1399	0.1327	0.1206	0.1231	0.1183	0.1164	0.5016	0.3014
	BGLSSQR$_{mean}$	0.5695	0.1920	0.2984	0.1803	0.1430	0.1215	0.0538	0.0428	0.0399	0.5169	0.3008
	BGLSSQR$_{median}$	0.5629	0.1917	0.2955	0.1956	0.1548	0.1199	0.0481	0.0360	0.0318	0.5167	0.3046
	BSGSSQR$_{mean}$	0.7531	0.1982	0.3094	0.1851	0.1655	0.0768	0.0534	0.0417	0.0433	0.7169	0.4109
	BSGSSQR$_{median}$	0.7818	0.1976	0.3199	0.1967	0.1892	0.0631	0.0446	0.0283	0.0321	0.7533	0.4376
0.90	CQR	0.7147	0.2149	0.4219	0.1527	0.1538	0.1570	0.1609	0.1522	0.1494	0.5579	0.3306
	LASSOQR	1.0936	0.3684	0.4942	0.2911	0.2650	0.2282	0.2230	0.2271	0.2169	1.1047	0.6755
	GLASSOQR	0.6758	0.3481	0.4222	0.2696	0.2513	0.2687	0.2378	0.2475	0.2403	0.7258	0.4736
	BLASSOQR	0.5361	0.1867	0.2737	0.1399	0.1327	0.1206	0.1231	0.1184	0.1164	0.5017	0.3014
	BGLSSQR$_{mean}$	0.5692	0.1937	0.2991	0.1812	0.1434	0.1216	0.0541	0.0431	0.0399	0.5158	0.3013
	BGLSSQR$_{median}$	0.5624	0.1945	0.2958	0.1961	0.1544	0.1202	0.0483	0.0361	0.0317	0.5174	0.3066
	BSGSSQR$_{mean}$	0.7514	0.1986	0.3088	0.1850	0.1655	0.0768	0.0537	0.0418	0.0433	0.7155	0.4104
	BSGSSQR$_{median}$	0.7790	0.1981	0.3185	0.1964	0.1891	0.0631	0.0456	0.0283	0.0322	0.7496	0.4348

Table C.17: Estimated **MBE** from eight methods in Study 5.2, assuming a χ^2_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
0.10	CQR	-0.0149	-0.0029	-0.0070	-0.0121	0.0120	-0.0147	-0.0111	0.0047	-0.0030	-0.0023	-0.0228
	LASSOQR	0.6197	0.1388	-0.2411	-0.0735	-0.0670	0.0065	-0.0052	-0.0070	-0.0040	-0.6299	-0.3185
	GLASSOQR	0.2090	0.1298	-0.0349	-0.0657	-0.0390	-0.0012	-0.0002	-0.0035	0.0034	-0.2990	-0.1226
	BLASSOQR	0.5063	0.1653	-0.1507	-0.0830	-0.0495	-0.0143	-0.0050	-0.0067	0.0092	-0.5887	-0.3051
	BGLSSQR$_{mean}$	0.5874	0.1751	-0.2242	-0.2052	-0.1223	-0.0144	-0.0030	-0.0047	0.0013	-0.6216	-0.2791
	BGLSSQR$_{median}$	0.5752	0.1753	-0.2219	-0.2300	-0.1451	-0.0155	-0.0015	-0.0047	-0.0020	-0.6520	-0.3086
	BSGSSQR$_{mean}$	0.9804	0.2079	-0.3081	-0.2194	-0.1568	-0.0088	0.0001	-0.0042	0.0020	-1.0483	-0.5357
	BSGSSQR$_{median}$	0.9922	0.2130	-0.3178	-0.2621	-0.1945	-0.0094	0.0014	-0.0048	0.0007	-1.1443	-0.6027
0.50	CQR	-0.0149	-0.0029	-0.0070	-0.0121	0.0120	-0.0147	-0.0111	0.0047	-0.0030	-0.0023	-0.0228
	LASSOQR	0.9794	0.2200	-0.3081	-0.1862	-0.1431	-0.0059	-0.0005	-0.0035	0.0027	-1.1595	-0.6225
	GLASSOQR	0.3004	0.1802	-0.0111	-0.1728	-0.0938	-0.0100	0.0047	0.0017	0.0071	-0.6063	-0.3096
	BLASSOQR	0.5063	0.1653	-0.1508	-0.0830	-0.0496	-0.0143	-0.0050	-0.0068	0.0092	-0.5887	-0.3052
	BGLSSQR$_{mean}$	0.5888	0.1754	-0.2256	-0.2058	-0.1226	-0.0138	-0.0033	-0.0046	0.0008	-0.6219	-0.2791
	BGLSSQR$_{median}$	0.5783	0.1751	-0.2228	-0.2320	-0.1458	-0.0158	-0.0019	-0.0051	-0.0014	-0.6544	-0.3084
	BSGSSQR$_{mean}$	0.9824	0.2084	-0.3087	-0.2195	-0.1566	-0.0090	0.0000	-0.0040	0.0021	-1.0499	-0.5358
	BSGSSQR$_{median}$	0.9929	0.2134	-0.3181	-0.2620	-0.1945	-0.0094	0.0015	-0.0048	0.0004	-1.1443	-0.6023
0.90	CQR	-0.0149	-0.0029	-0.0070	-0.0121	0.0120	-0.0147	-0.0111	0.0047	-0.0030	-0.0023	-0.0228
	LASSOQR	1.0378	0.1524	-0.2601	-0.1334	-0.0989	-0.0035	0.0083	-0.0221	0.0046	-1.1569	-0.5821
	GLASSOQR	0.4391	-0.0189	-0.0331	-0.0578	-0.0461	-0.0212	0.0163	-0.0306	0.0078	-0.5360	-0.2680
	BLASSOQR	0.5064	0.1654	-0.1508	-0.0830	-0.0496	-0.0143	-0.0050	-0.0068	0.0092	-0.5888	-0.3052
	BGLSSQR$_{mean}$	0.5908	0.1748	-0.2254	-0.2055	-0.1214	-0.0137	-0.0038	-0.0041	0.0011	-0.6250	-0.2811
	BGLSSQR$_{median}$	0.5796	0.1737	-0.2230	-0.2308	-0.1445	-0.0156	-0.0019	-0.0048	-0.0011	-0.6566	-0.3103
	BSGSSQR$_{mean}$	0.9814	0.2084	-0.3085	-0.2196	-0.1560	-0.0088	-0.0001	-0.0041	0.0021	-1.0488	-0.5353
	BSGSSQR$_{median}$	0.9927	0.2132	-0.3186	-0.2617	-0.1940	-0.0092	0.0015	-0.0049	0.0006	-1.1428	-0.6018

Table C.18: Estimated **RMSE** from eight methods in Study 5.2, assuming a χ^2_3 error distribution

τ	**Method**	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
0.10	CQR	1.1445	0.3285	0.6335	0.2409	0.2524	0.2494	0.2535	0.2469	0.2482	0.9369	0.5634
	LASSOQR	0.8216	0.2638	0.3228	0.1911	0.1787	0.1151	0.1173	0.1151	0.1103	0.8479	0.4730
	GLASSOQR	0.4498	0.2434	0.2689	0.1781	0.1495	0.1384	0.1196	0.1185	0.1167	0.5210	0.2967
	BLASSOQR	0.7425	0.3728	0.3621	0.2553	0.2352	0.2186	0.2218	0.2168	0.2043	0.8234	0.4920
	BGLSSQR$_{mean}$	0.9987	0.4211	0.4995	0.3498	0.2558	0.1757	0.1196	0.1248	0.1083	1.0077	0.5920
	BGLSSQR$_{median}$	1.0505	0.4385	0.5008	0.3869	0.2786	0.1683	0.1089	0.1150	0.0926	1.1015	0.6520
	BSGSSQR$_{mean}$	1.2692	0.4410	0.4265	0.3455	0.2502	0.1301	0.1062	0.1037	0.0908	1.3507	0.7350
	BSGSSQR$_{median}$	1.3474	0.4728	0.4120	0.4026	0.2845	0.1136	0.0872	0.0878	0.0675	1.5092	0.8246
0.50	CQR	1.1445	0.3285	0.6335	0.2409	0.2524	0.2494	0.2535	0.2469	0.2482	0.9369	0.5634
	LASSOQR	1.1735	0.4291	0.3666	0.3235	0.2581	0.1481	0.1543	0.1461	0.1411	1.3777	0.7739
	GLASSOQR	0.5947	0.3820	0.3886	0.3386	0.2646	0.2052	0.1857	0.1760	0.1829	0.9247	0.5417
	BLASSOQR	0.7425	0.3727	0.3621	0.2553	0.2352	0.2186	0.2219	0.2167	0.2043	0.8235	0.4920
	BGLSSQR$_{mean}$	1.0008	0.4221	0.5013	0.3501	0.2558	0.1758	0.1205	0.1245	0.1088	1.0097	0.5924
	BGLSSQR$_{median}$	1.0505	0.4391	0.5022	0.3889	0.2788	0.1680	0.1099	0.1151	0.0934	1.1055	0.6518
	BSGSSQR$_{mean}$	1.2700	0.4412	0.4267	0.3458	0.2503	0.1302	0.1064	0.1035	0.0910	1.3514	0.7345
	BSGSSQR$_{median}$	1.3471	0.4730	0.4122	0.4028	0.2846	0.1139	0.0880	0.0881	0.0682	1.5088	0.8238
0.90	CQR	1.1445	0.3285	0.6335	0.2409	0.2524	0.2494	0.2535	0.2469	0.2482	0.9369	0.5634
	LASSOQR	1.7323	0.7502	0.7773	0.5475	0.4805	0.4870	0.4783	0.5031	0.4724	1.7263	0.9931
	GLASSOQR	1.2685	0.7469	0.8015	0.6063	0.5438	0.5609	0.5564	0.5593	0.5352	1.3599	0.9168
	BLASSOQR	0.7425	0.3728	0.3620	0.2553	0.2353	0.2186	0.2218	0.2167	0.2043	0.8234	0.4919
	BGLSSQR$_{mean}$	1.0053	0.4206	0.5044	0.3496	0.2553	0.1762	0.1217	0.1246	0.1092	1.0123	0.5939
	BGLSSQR$_{median}$	1.0565	0.4367	0.5062	0.3875	0.2784	0.1688	0.1113	0.1150	0.0927	1.1090	0.6540
	BSGSSQR$_{mean}$	1.2687	0.4411	0.4265	0.3456	0.2503	0.1305	0.1064	0.1036	0.0903	1.3487	0.7336
	BSGSSQR$_{median}$	1.3463	0.4722	0.4117	0.4023	0.2849	0.1143	0.0876	0.0881	0.0666	1.5055	0.8229

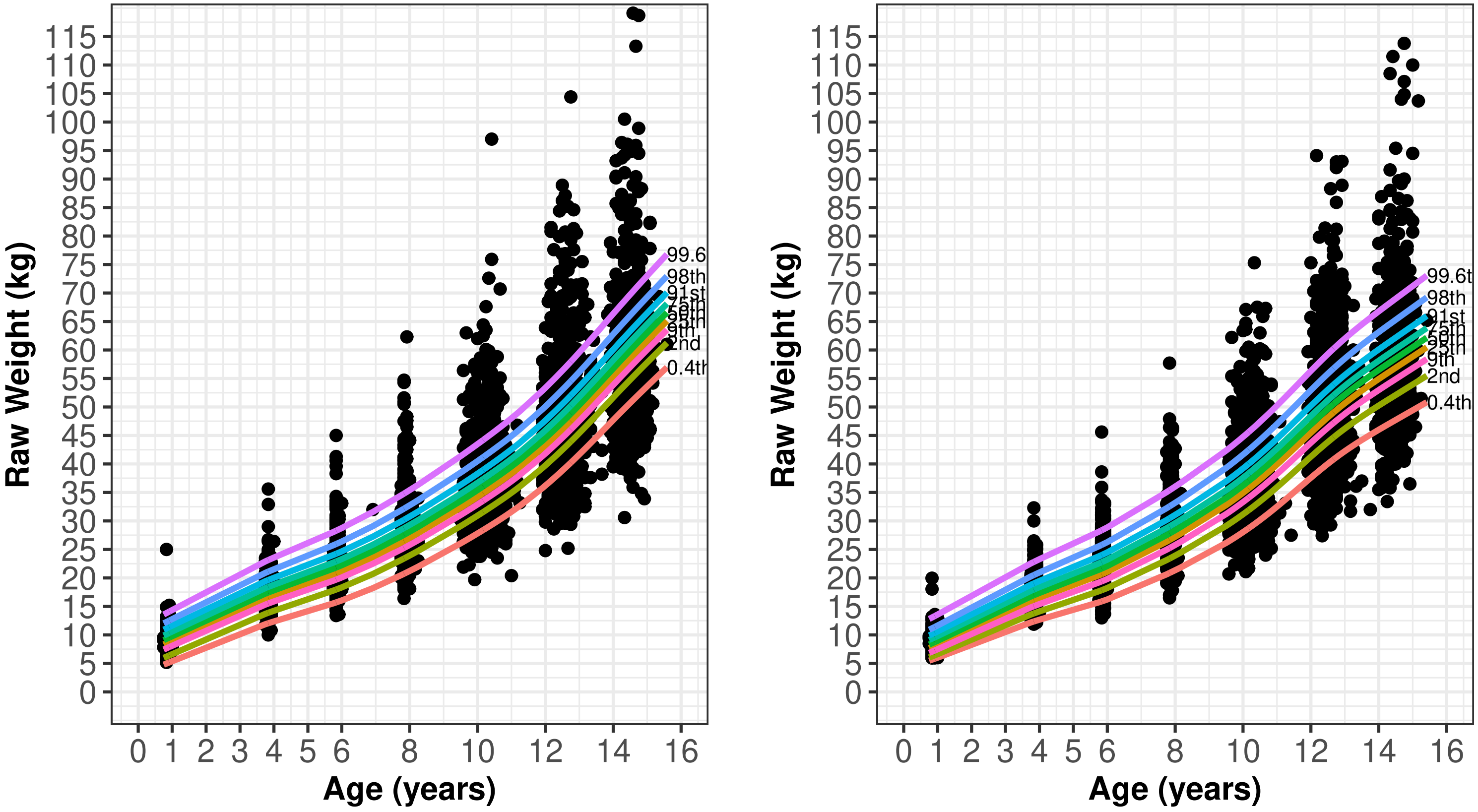
Appendix D

Additional results from Chapter 6

D.1 The quantile growth curves from the AQMM models, fitted with both random intercepts and random slopes.

In this section, the results from the fitted quantile models with random intercepts and random slopes for each child's physical growth measurements, are presented. Each figure representing the results is listed in the table below.

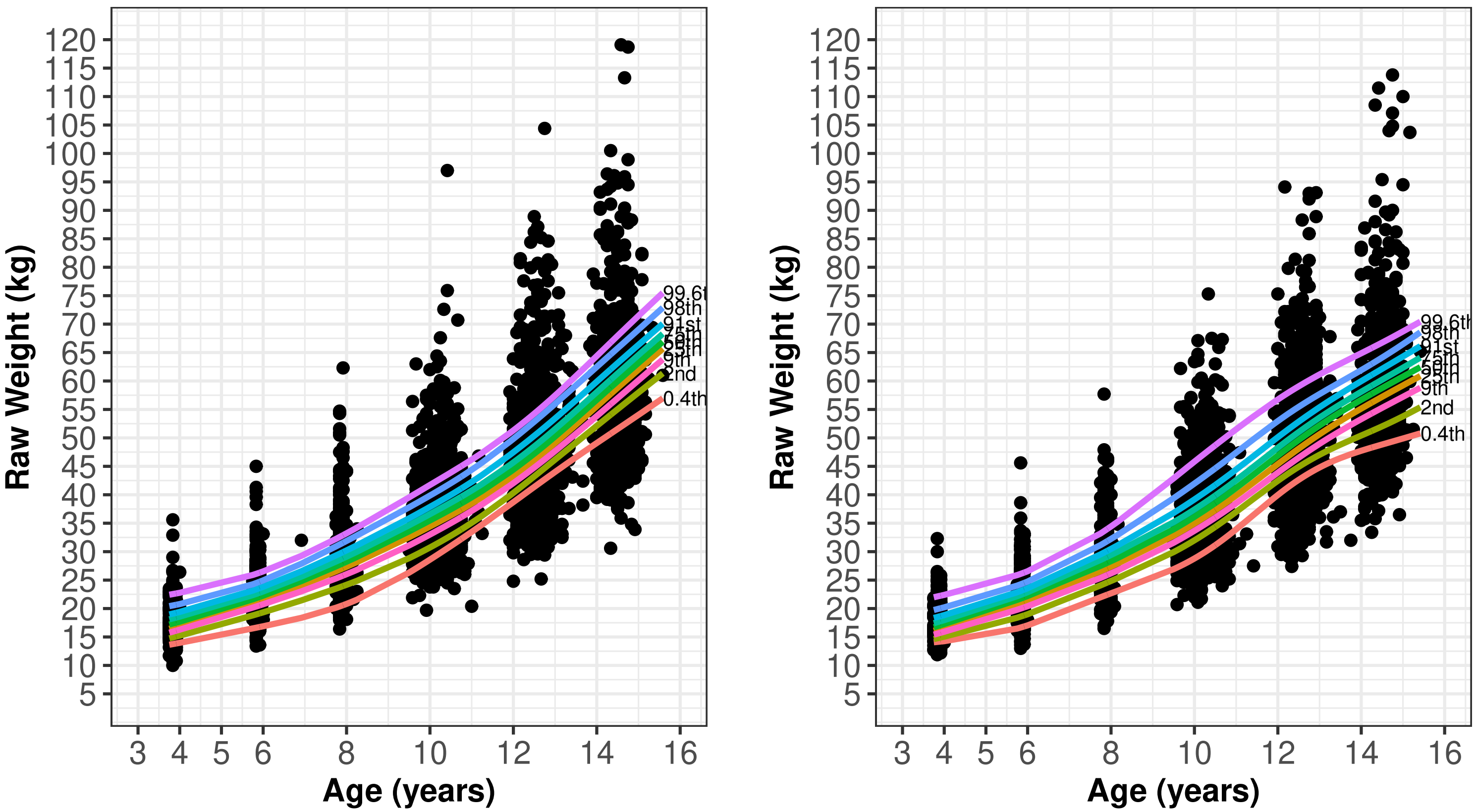
Figure	**Physical growth measurement**	Dataset
D.1	Raw weight	Entire GUS dataset
D.2	Raw weight	Children aged 4 - 14 years GUS dataset
D.3	WAZ	Entire GUS dataset
D.4	WAZ	Children aged 4 - 14 years GUS dataset
D.5	Raw height	Children aged 4 - 14 years GUS dataset
D.6	HAZ	Children aged 4 - 14 years GUS dataset



(a) Male

(b) Female

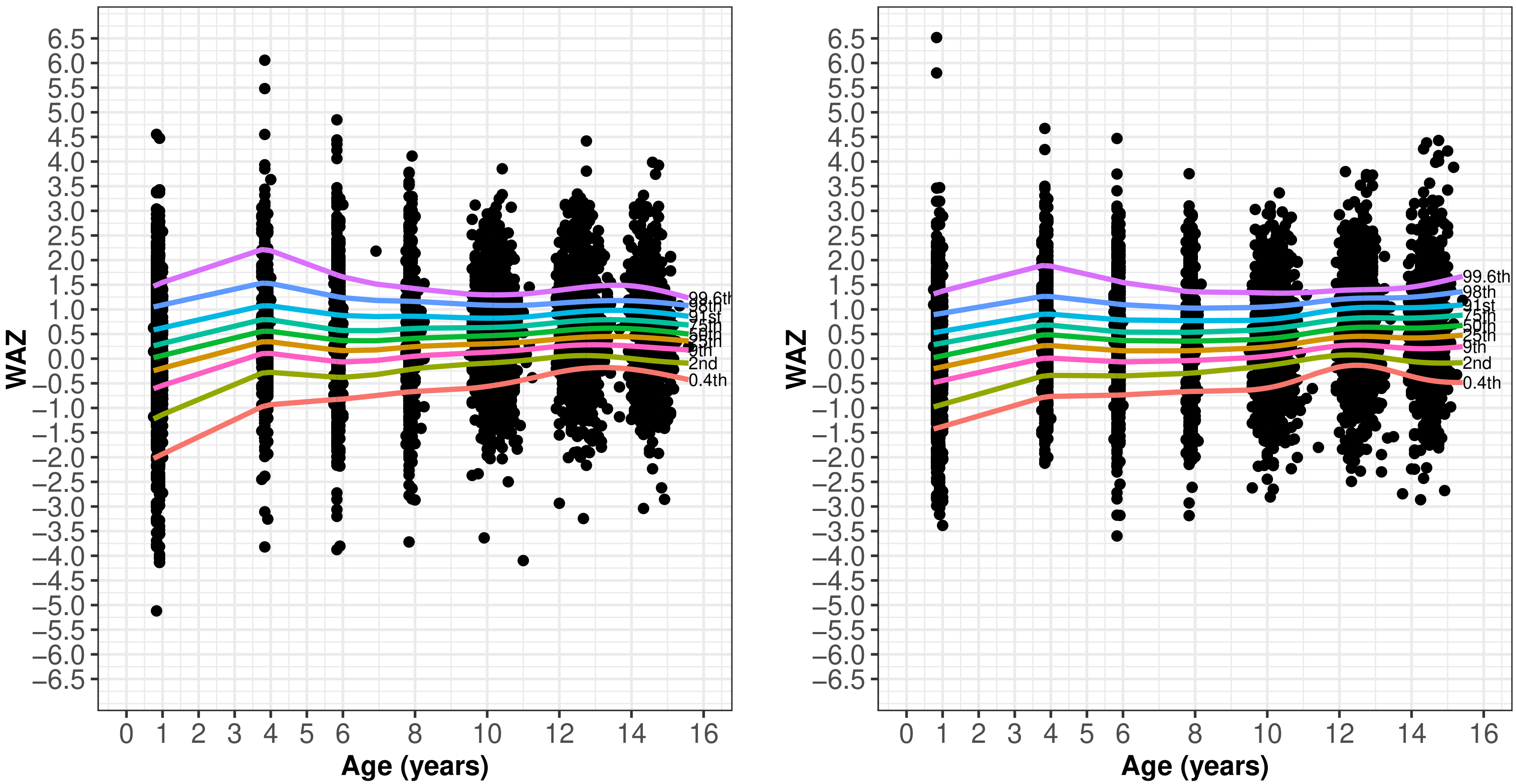
Figure D.1: Conditional quantile curves for raw weight, estimated using the AQMM with cubic P-splines and fitted with **random intercepts and slopes**, for a child in the reference group across nine quantile levels using the **entire GUS dataset**.



(a) Male

(b) Female

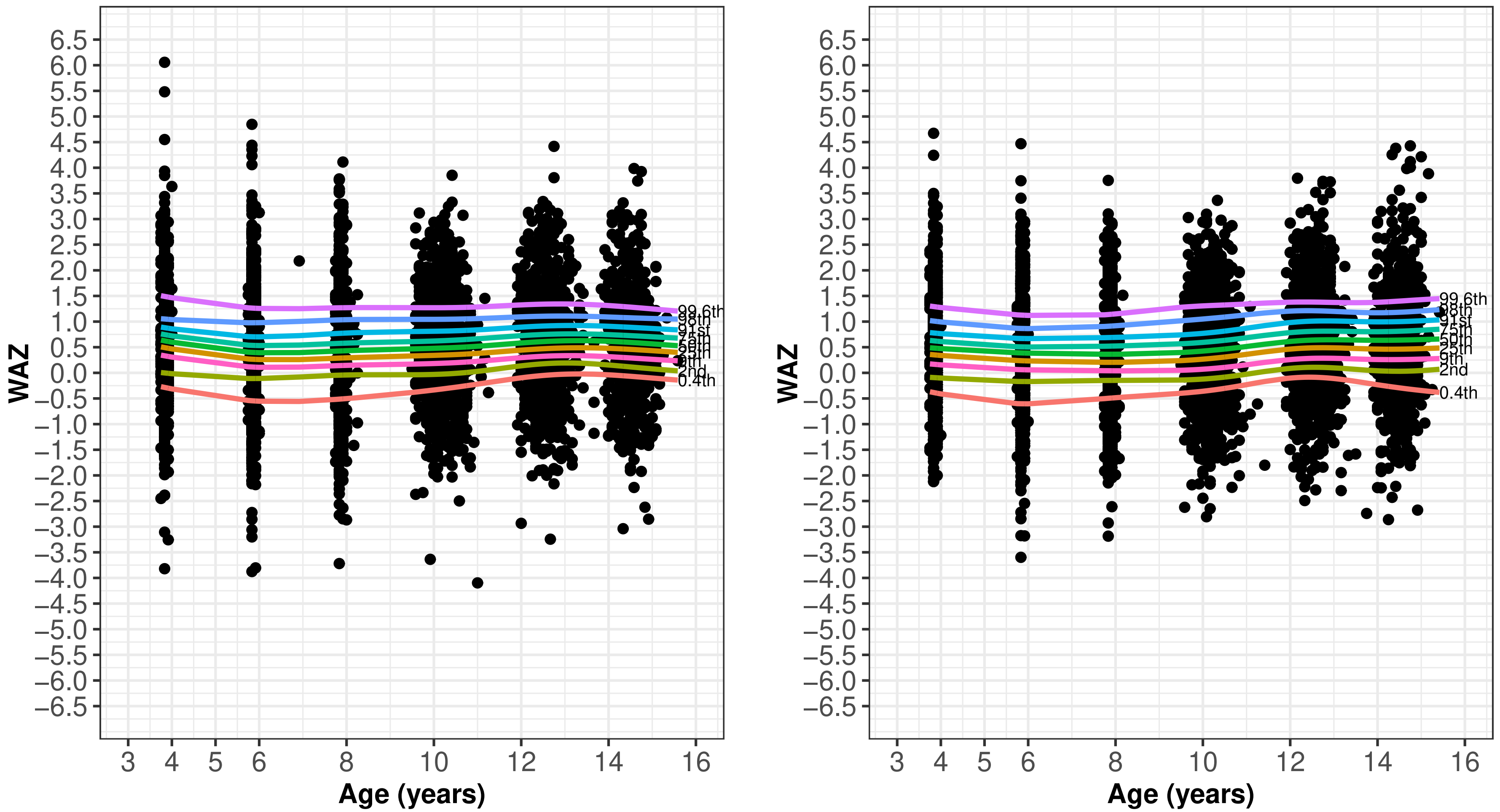
Figure D.2: Conditional quantile curves for raw weight, estimated using the AQMM with cubic P-splines and fitted with **random intercepts and slopes**, for a child in the reference group across nine quantile levels using the **children aged 4-14 years GUS dataset**.



(a) Male

(b) Female

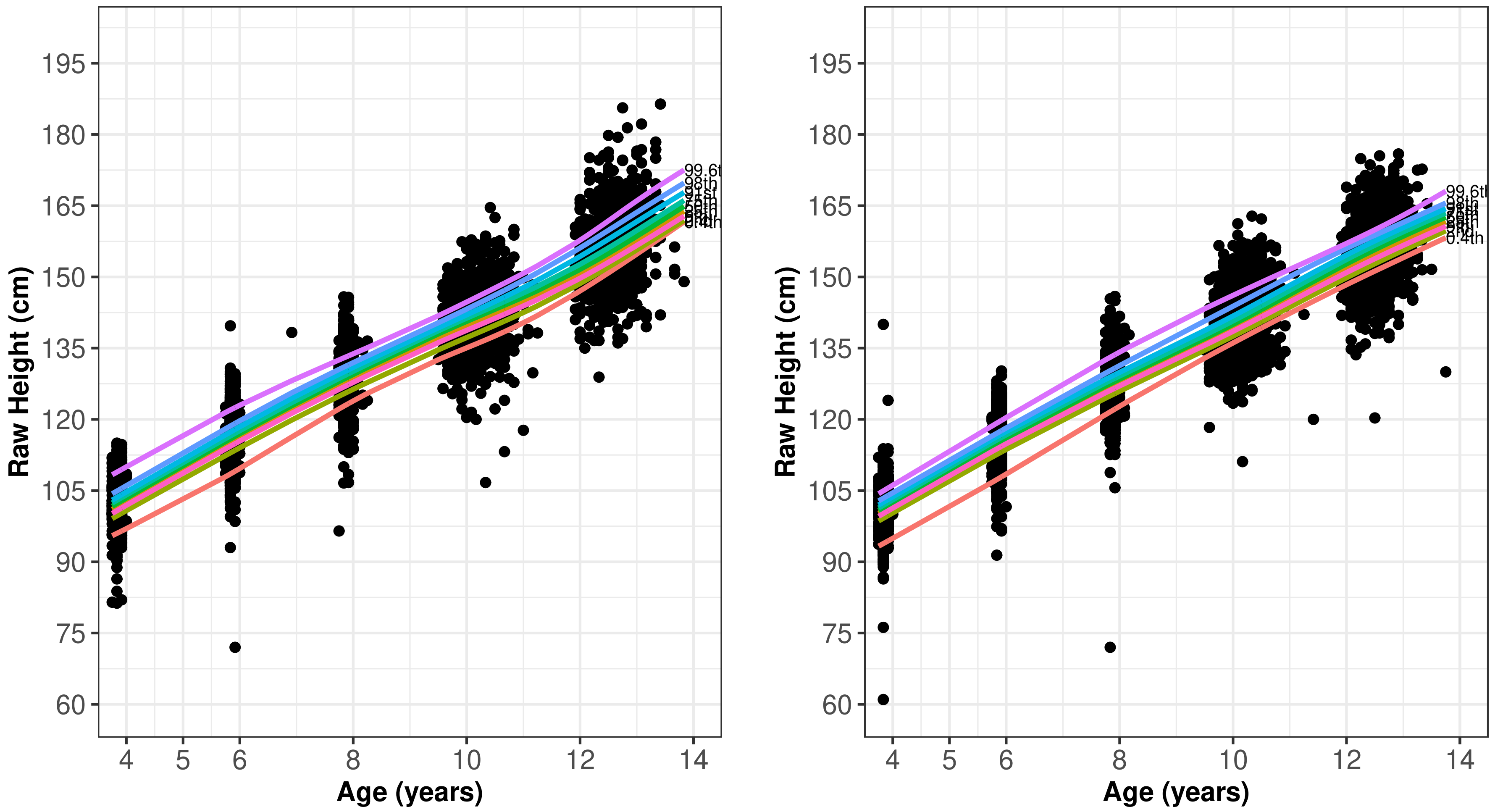
Figure D.3: Conditional quantile curves for WAZ, estimated using the AQMM with cubic P-splines and fitted with **random intercepts and slopes**, for a child in the reference group across nine quantile levels using the **entire GUS dataset**.



(a) Male

(b) Female

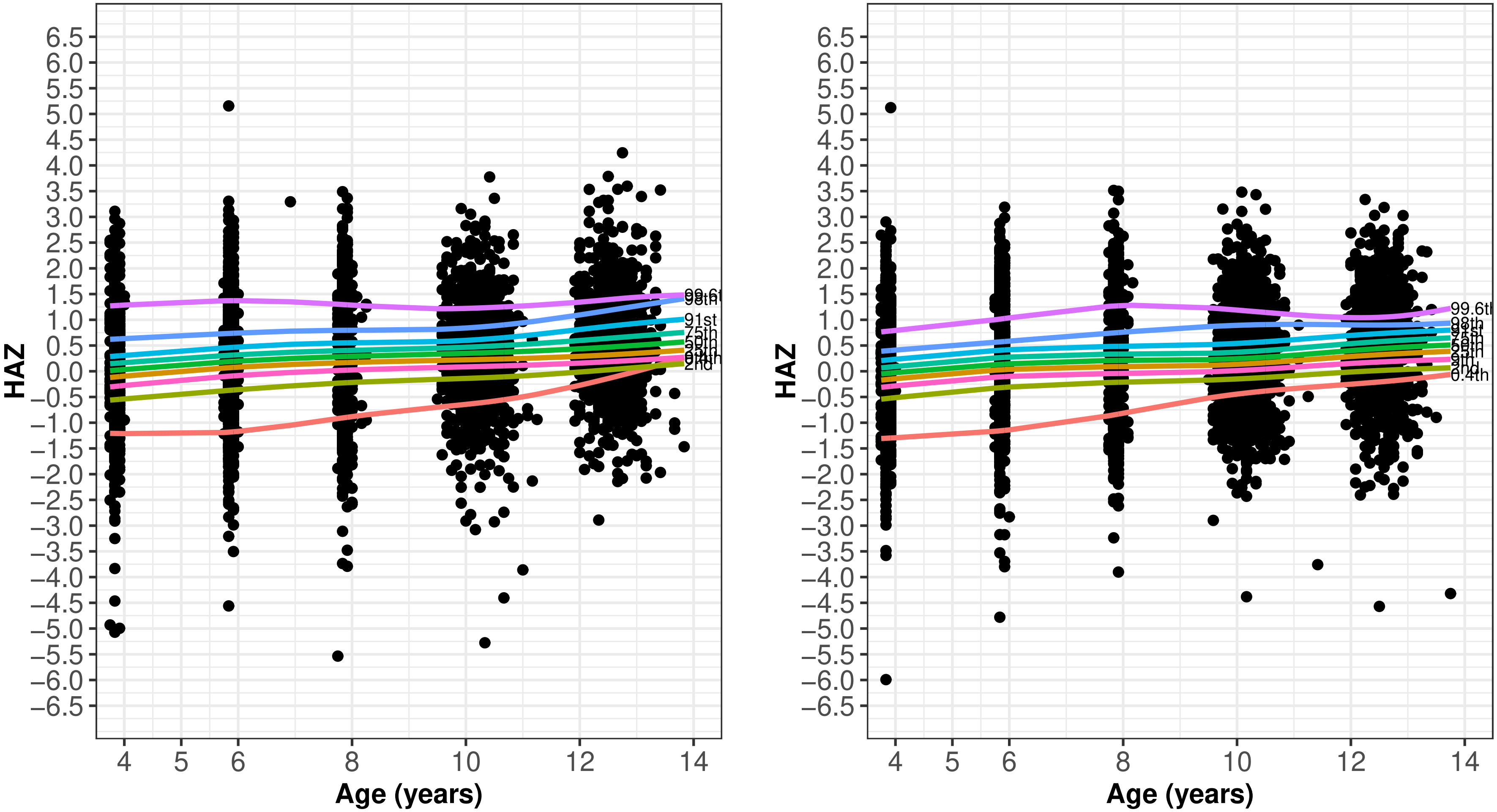
Figure D.4: Conditional quantile curves for WAZ, estimated using the AQMM with cubic P-splines and fitted with **random intercepts and slopes**, for a child in the reference group across nine quantile levels using the **children aged 4 - 14 years GUS dataset**.



(a) Male

(b) Female

Figure D.5: Conditional quantile curves for raw height, estimated using the AQMM with cubic P-splines and fitted with **random intercepts and slopes**, for a child in the reference group across nine quantile levels using the **GUS dataset**.



(a) Male

(b) Female

Figure D.6: Conditional quantile curves for HAZ, estimated using the AQMM with cubic P-splines and fitted with **random intercepts and slopes**, for a child in the reference group across nine quantile levels using the **GUS dataset**.

D.2 The initial fitted quantile models via AQMM for identifying risk factors

In this section, the results from the initially fitted quantile models for each child's physical growth measurements, are presented. Each table representing the results is listed in the table below.

Table	**Physical growth measurement**	Dataset
D.1	(Centred) Raw weight	Children aged 4 - 14 years GUS dataset
D.2	(Centred) WAZ	Children aged 4 - 14 years GUS dataset
D.3	(Centred) Raw height	Children aged 4 - 14 years GUS dataset
D.4	(Centred) HAZ	Children aged 4 - 14 years GUS dataset

Table D.1: Estimates from the initial AQMM with cubic P-splines for the **(centred) raw weight** growth measurement in **children aged 4 - 14 years** from the GUS dataset (standard errors in brackets)†

	$\tau = 0.10$	$\tau = 0.90$
Fixed effects		
(Intercept)	-1.5730 (0.8611)	0.5189 (1.1984)
Sex (Male)	0.2439 (0.1419)	0.2878 (0.1696)
Low birth weight (Yes)	**-0.7644[c] (0.3121)**	**-1.0506[c] (0.4918)**
Ethnicity of a child (White)	**-1.8112[a] (0.4939)**	-0.2500 (1.0306)
Child's health in general (Good)	0.1460 (0.1138)	0.2682 (0.1605)
Child's health in general (Fair, Bad, Very Bad)	-0.4911 (0.3615)	**0.8954[c] (0.3993)**
Number of accidents or injuries of child	0.1329 (0.0669)	-0.0398 (0.0611)
Child's birth order	0.0385 (0.0857)	-0.0966 (0.0705)
Mother's marital status (Single)	0.1753 (0.1710)	0.1856 (0.1716)
Mother's marital status (Other)	**0.6132[c] (0.2449)**	0.4520 (0.2648)
Urban-rural classification (Other urban)	0.2544 (0.1837)	0.2639 (0.2228)
Urban-rural classification (Small, accessible towns)	0.3878 (0.2358)	0.5496 (0.3403)
Urban-rural classification (Small, remote towns)	0.0201 (0.4044)	0.5269 (0.4535)
Urban-rural classification (Accessible rural)	0.3678 (0.2620)	0.4463 (0.2483)
Urban-rural classification (Remote rural)	**0.6765[b] (0.2400)**	**0.9063[b] (0.2841)**
Household size	0.0601 (0.0931)	-0.0230 (0.0907)
Mother's age at first child's birth (<20 years old)	0.4231 (0.2712)	0.2608 (0.2695)
Mother's age at first child's birth (≥ 30 years old)	0.0769 (0.2582)	0.2485 (0.2830)
Respondent's alcoholic drinks (Every day)	0.1338 (0.6117)	**2.5442[a] (0.6300)**
Respondent's alcoholic drinks (4 - 6 times a week)	0.3451 (0.4442)	0.7871 (0.4572)
Respondent's alcoholic drinks (2 - 3 times a week)	0.4216 (0.4587)	**0.9630[c] (0.4397)**
Respondent's alcoholic drinks (Once a week)	0.4228 (0.4603)	0.3681 (0.4210)
Respondent's alcoholic drinks (2 - 3 times a month)	0.6421 (0.4653)	**0.9546[b] (0.4241)**
Respondent's alcoholic drinks (Once a month or less)	0.5578 (0.5430)	0.5538 (0.6114)
Respondent's alcoholic drinks (Not in the last year)	0.1097 (0.5024)	1.4186 (0.4908)
Respondent's current health (Very good)	0.0090 (0.1338)	0.1473 (0.1431)
Respondent's current health (Good)	0.0999 (0.1562)	0.5085 (0.2776)
Respondent's current health (Fair, Poor)	0.3452 (0.2625)	0.5541 (0.3211)
Smoking cigarettes while pregnant (Yes)	-0.0117 (0.2036)	0.0755 (0.3056)
Drinking alcohol while pregnant (≥ 3 - 4 times a week)	**2.6610[a] (0.6045)**	**8.8987[a] (0.5973)**
Drinking alcohol while pregnant (1 - 2 times a week)	0.8922 (0.6103)	0.5750 (0.5746)
Drinking alcohol while pregnant (2 - 3 times a month)	0.5578 (0.5157)	0.4190 (0.4525)
Drinking alcohol while pregnant (<once a month)	0.2469 (0.5141)	0.1961 (0.4682)
Respondent's health problem(s) in a year (Yes)	0.0028 (0.1460)	0.2172 (0.2615)
Respondent's current job (No)	-0.1942 (0.3191)	0.4275 (0.3904)
Deprivation quintile (2)	**0.3622[c] (0.1537)**	**0.5553[c] (0.2320)**
Deprivation quintile (3)	**0.3015[c] (0.1485)**	**0.5256[c] (0.2423)**
Deprivation quintile (4)	**0.6619[b] (0.2195)**	**0.9131[a] (0.2323)**
Deprivation quintile (5)	0.3926 (0.2501)	**0.7185[c] (0.3430)**
Equivalised income	**0.5379[a] (0.0779)**	0.1776 (0.1269)
Linear basis term of Age (in year)	-73.6778 (73.8015)	-77.2178 (78.0153)
Random effects		
$\hat{\sigma}_0$ (SD of intercepts Age in year)	2.3570	1.8232
$\hat{\sigma}_1$ (SD of slopes of the Age in year)	0.7299	0.5840
$\hat{\rho}_{01}$ (Correlation of intercepts and slopes)	-0.9657	-0.9505

[a] p < 0.001, [b] p < 0.005, [c] p < 0.05

† The reference categories are given in Tables 2.9 to 2.11.

Table D.2: Estimates from the initial AQMM with cubic P-splines for **(centred) WAZ** in **children aged 4 - 14 years** form the GUS dataset (standard errors in brackets)†

	$\tau = 0.10$	$\tau = 0.90$
Fixed effects		
(Intercept)	-0.3973 (0.2720)	-0.0355 (0.2860)
Sex (Male)	0.0695 (0.0387)	0.0735 (0.0415)
Low birth weight (Yes)	**-0.5640[a] (0.0994)**	**-0.5203[a] (0.1000)**
Ethnicity of a child (White)	-0.2383 (0.1480)	-0.0817 (0.1800)
Child's health in general (Good)	0.0102 (0.0190)	0.0091 (0.0207)
Child's health in general (Fair, Bad, Very Bad)	0.0177 (0.0638)	0.0256 (0.0476)
Number of accidents or injuries of child	0.0033 (0.0067)	-0.0006 (0.0097)
Child's birth order	0.0114 (0.0222)	0.0050 (0.0201)
Mother's marital status (Single)	0.0089 (0.0217)	-0.0058 (0.0248)
Mother's marital status (Other)	**0.1016[a] (0.0290)**	**0.0777[c] (0.0364)**
Urban-rural classification (Other urban)	0.0289 (0.0430)	0.0283 (0.0372)
Urban-rural classification (Small, accessible towns)	0.0756 (0.0415)	0.0317 (0.0508)
Urban-rural classification (Small, remote towns)	0.1057 (0.0891)	0.0323 (0.0860)
Urban-rural classification (Accessible rural)	0.0921 (0.0538)	0.0642 (0.0483)
Urban-rural classification (Remote rural)	0.1385 (0.0713)	0.0701 (0.0659)
Household size	-0.0042 (0.0165)	-0.0117 (0.0156)
Mother's age at first child's birth ($<$20 years old)	0.0905 (0.0688)	0.0862 (0.0784)
Mother's age at first child's birth ($\geq$ 30 years old)	-0.0049 (0.0732)	0.0810 (0.0707)
Respondent's alcoholic drinks (Every day)	0.0592 (0.1774)	-0.0175 (0.1740)
Respondent's alcoholic drinks (4 - 6 times a week)	0.0869 (0.1517)	0.1265 (0.1516)
Respondent's alcoholic drinks (2 - 3 times a week)	0.0959 (0.1485)	0.1602 (0.1477)
Respondent's alcoholic drinks (Once a week)	0.1078 (0.1504)	0.1406 (0.1485)
Respondent's alcoholic drinks (2 - 3 times a month)	0.1479 (0.1536)	0.1819 (0.1540)
Respondent's alcoholic drinks (Once a month or less)	0.1082 (0.1818)	0.1798 (0.1667)
Respondent's alcoholic drinks (Not in the last year)	**0.3505[c] (0.1724)**	**0.3782[c] (0.1705)**
Respondent's current health (Very good)	0.0321 (0.0200)	**0.0946[a] (0.0165)**
Respondent's current health (Good)	**0.0535[c] (0.0247)**	**0.1154[a] (0.0216)**
Respondent's current health (Fair, Poor)	**0.0816[c] (0.0374)**	**0.1370[b] (0.0414)**
Smoking cigarettes while pregnant (Yes)	**0.1464[c] (0.0561)**	**0.1529[c] (0.0615)**
Drinking alcohol while pregnant ($\geq$ 3 - 4 times a week)	0.0832 (0.2299)	0.3270 (0.2260)
Drinking alcohol while pregnant (1 - 2 times a week)	0.1812 (0.2003)	0.0872 (0.2080)
Drinking alcohol while pregnant (2 - 3 times a month)	0.0499 (0.2104)	-0.0413 (0.2046)
Drinking alcohol while pregnant ($<$once a month)	0.0128 (0.2128)	-0.0420 (0.2172)
Respondent's health problem(s) in a year (Yes)	0.0016 (0.0297)	0.0173 (0.0410)
Respondent's current job (No)	0.0603 (0.0567)	0.0559 (0.0461)
Deprivation quintile (2)	**0.0780[b] (0.0289)**	**0.1002[a] (0.0225)**
Deprivation quintile (3)	**0.0980[b] (0.0359)**	**0.1013[b] (0.0303)**
Deprivation quintile (4)	**0.2025[a] (0.0442)**	**0.1876[a] (0.0408)**
Deprivation quintile (5)	**0.1452[b] (0.0533)**	**0.1327[c] (0.0520)**
Equivalised income	**0.0638[a] (0.0143)**	**0.0296[b] (0.0106)**
Linear basis term of Age (in year)	-0.0582 (0.2014)	-0.2817 (0.2752)
Random effects		
$\hat{\sigma}_0$ (SD of intercepts Age in year)	0.2421	0.2290
$\hat{\sigma}_1$ (SD of slopes of the Age in year)	0.0158	0.0155
$\hat{\rho}_{01}$ (Correlation of intercepts and slopes)	-0.2193	-0.3401

[a] p $<$ 0.001, [b] p $<$ 0.005, [c] p $<$ 0.05

† The reference categories are given in Tables 2.9 to 2.11.

Table D.3: Estimates from the initial AQMM with P-splines for **(centred) raw height** growth measurement in **children aged 4 - 14 years** from the GUS dataset (standard errors in brackets)†

	$\tau = 0.10$	$\tau = 0.90$
Fixed effects		
(Intercept)	-1.6005 (1.4547)	-0.0701 (1.6519)
Sex (Male)	**0.6641^a (0.1887)**	**1.1732^a (0.1882)**
Low birth weight (Yes)	**-2.1866^a (0.4718)**	**-1.3232^b (0.4346)**
Ethnicity of a child (White)	**-1.8065^a (0.5023)**	-0.4995 (1.1266)
Child's health in general (Good)	-0.1595 (0.0903)	0.1260 (0.1623)
Child's health in general (Fair, Bad, Very Bad)	-0.6563 (0.3522)	**1.1967^b (0.3773)**
Number of accidents or injuries of child	0.0156 (0.0364)	0.0825 (0.0423)
Child's birth order	**-0.2347^c (0.1071)**	**-0.2582^c (0.1050)**
Mother's marital status (Single)	-0.0071 (0.1523)	0.1033 (0.0879)
Mother's marital status (Other)	0.4886 (0.2497)	**0.5001^c (0.1877)**
Urban-rural classification (Other urban)	0.1092 (0.2177)	-0.0368 (0.2487)
Urban-rural classification (Small, accessible towns)	0.0455 (0.4200)	-0.0853 (0.4516)
Urban-rural classification (Small, remote towns)	0.7555 (0.4730)	**0.9896^c (0.4328)**
Urban-rural classification (Accessible rural)	-0.1790 (0.2651)	-0.0611 (0.2834)
Urban-rural classification (Remote rural)	0.3167 (0.3019)	0.0124 (0.2822)
Household size	**0.2432^b (0.0892)**	-0.0705 (0.0842)
Mother's age at first child's birth ($<$20 years old)	**0.9058^c (0.3886)**	**1.2333^a (0.3483)**
Mother's age at first child's birth ($\geq$ 30 years old)	-0.1122 (0.3858)	0.1576 (0.4002)
Respondent's alcoholic drinks (Every day)	1.8615 (0.9351)	**2.5695^b (0.9222)**
Respondent's alcoholic drinks (4 - 6 times a week)	0.9724 (0.8821)	1.2598 (0.9126)
Respondent's alcoholic drinks (2 - 3 times a week)	0.7774 (0.9409)	0.8533 (0.9559)
Respondent's alcoholic drinks (Once a week)	1.0236 (0.9279)	0.8180 (0.9723)
Respondent's alcoholic drinks (2 - 3 times a month)	1.0411 (0.9525)	0.9474 (0.9630)
Respondent's alcoholic drinks (Once a month or less)	1.2501 (1.0733)	0.8176 (0.9609)
Respondent's alcoholic drinks (Not in the last year)	-0.2073 (0.9728)	1.3810 (0.9751)
Respondent's current health (Very good)	0.1140 (0.1542)	0.1742 (0.1474)
Respondent's current health (Good)	0.3205 (0.1634)	0.1849 (0.1554)
Respondent's current health (Fair, Poor)	0.2025 (0.2175)	0.1380 (0.1902)
Smoking cigarettes while pregnant (Yes)	**-0.5189^c (0.2409)**	-0.5287 (0.3097)
Drinking alcohol while pregnant ($\geq$ 3 - 4 times a week)	2.2742 (1.1685)	**13.6933^a (1.1170)**
Drinking alcohol while pregnant (1 - 2 times a week)	1.1275 (1.0104)	1.5553 (1.0854)
Drinking alcohol while pregnant (2 - 3 times a month)	0.3728 (1.0028)	0.4252 (1.0230)
Drinking alcohol while pregnant ($<$once a month)	0.1081 (1.0135)	0.2434 (1.0309)
Respondent's health problem(s) in a year (Yes)	-0.0267 (0.1374)	**0.3700^b (0.1178)**
Respondent's current job (No)	0.3846 (0.3192)	**0.7009^c (0.3036)**
Deprivation quintile (2)	**0.8082^a (0.1333)**	0.2422 (0.1825)
Deprivation quintile (3)	0.2347 (0.1777)	0.3256 (0.2714)
Deprivation quintile (4)	**0.4816^c (0.2274)**	**0.6295^b (0.1904)**
Deprivation quintile (5)	0.0740 (0.2723)	0.3199 (0.2411)
Equivalised income	**0.2687^a (0.0644)**	**0.1760^b (0.0657)**
Linear basis term of Age (in year)	-90.3927 (83.8725)	-92.0816 (84.9448)
Random effects		
$\hat{\sigma}_0$ (SD of intercepts Age in year)	1.8536	1.7352
$\hat{\sigma}_1$ (SD of slopes of the Age in year)	0.2908	0.2820
$\hat{\rho}_{01}$ (Correlation of intercepts and slopes)	0.2371	0.3611

^{a} $p < 0.001$, ^{b} $p < 0.005$, ^{c} $p < 0.05$

† The reference categories are given in Tables 2.9 to 2.11.

Table D.4: Estimates from the initial AQMM with P-splines for **(centred) HAZ** in **children aged 4 - 14 years** from the GUS dataset (standard errors in brackets)†

	$\tau = 0.10$	$\tau = 0.90$
Fixed effects		
(Intercept)	-0.3705 (0.3349)	-0.0095 (0.3635)
Sex (Male)	0.0394 (0.0398)	0.0528 (0.0401)
Low birth weight (Yes)	**-0.3873[a] (0.1018)**	**-0.3561[b] (0.1227)**
Ethnicity of a child (White)	**-0.4522[b] (0.1327)**	-0.2133 (0.1458)
Child's health in general (Good)	0.0200 (0.0202)	0.0253 (0.0218)
Child's health in general (Fair, Bad, Very Bad)	**0.1629[b] (0.0544)**	0.0026 (0.0554)
Number of accidents or injuries of child	0.0057 (0.0063)	0.0165 (0.0091)
Child's birth order	-0.0360 (0.0205)	-0.0438 (0.0219)
Mother's marital status (Single)	-0.0002 (0.0204)	0.0323 (0.0218)
Mother's marital status (Other)	**0.1222[b] (0.0357)**	**0.0878[c] (0.0366)**
Urban-rural classification (Other urban)	-0.0558 (0.0537)	-0.0465 (0.0551)
Urban-rural classification (Small, accessible towns)	0.0213 (0.0998)	-0.0220 (0.0992)
Urban-rural classification (Small, remote towns)	**0.1934[c] (0.0817)**	-0.0417 (0.0854)
Urban-rural classification (Accessible rural)	-0.0026 (0.0645)	-0.0453 (0.0649)
Urban-rural classification (Remote rural)	0.0841 (0.0578)	-0.0494 (0.0697)
Household size	**0.0342[c] (0.0150)**	0.0167 (0.0143)
Mother's age at first child's birth (<20 years old)	**0.2299[b] (0.0746)**	**0.1918[c] (0.0788)**
Mother's age at first child's birth (≥ 30 years old)	0.0361 (0.0830)	0.0858 (0.0852)
Respondent's alcoholic drinks (Every day)	0.4011 (0.2185)	0.2770 (0.2189)
Respondent's alcoholic drinks (4 - 6 times a week)	0.3604 (0.1993)	0.3771 (0.2013)
Respondent's alcoholic drinks (2 - 3 times a week)	0.3169 (0.2148)	0.3922 (0.2184)
Respondent's alcoholic drinks (Once a week)	0.2929 (0.2132)	0.3415 (0.2171)
Respondent's alcoholic drinks (2 - 3 times a month)	0.3164 (0.2170)	0.2986 (0.2187)
Respondent's alcoholic drinks (Once a month or less)	0.3140 (0.2192)	0.3527 (0.2262)
Respondent's alcoholic drinks (Not in the last year)	0.4117 (0.2196)	0.4199 (0.2271)
Respondent's current health (Very good)	**0.0665[c] (0.0304)**	**0.0739[b] (0.0264)**
Respondent's current health (Good)	0.0594 (0.0316)	**0.0856[b] (0.0310)**
Respondent's current health (Fair, Poor)	0.0884 (0.0460)	**0.1170[c] (0.0443)**
Smoking cigarettes while pregnant (Yes)	0.0096 (0.0567)	0.0029 (0.0686)
Drinking alcohol while pregnant (≥ 3 - 4 times a week)	0.1485 (0.2677)	-0.0248 (0.2592)
Drinking alcohol while pregnant (1 - 2 times a week)	0.3029 (0.2361)	0.0427 (0.2426)
Drinking alcohol while pregnant (2 - 3 times a month)	0.0822 (0.2299)	-0.0155 (0.2390)
Drinking alcohol while pregnant (<once a month)	0.0403 (0.2411)	-0.0521 (0.2394)
Respondent's health problem(s) in a year (Yes)	0.0427 (0.0316)	0.0616 (0.0386)
Respondent's current job (No)	**0.1758[a] (0.0472)**	-0.0358 (0.0793)
Deprivation quintile (2)	0.0450 (0.0291)	**0.1092[a] (0.0295)**
Deprivation quintile (3)	0.0594 (0.0381)	**0.1249[b] (0.0403)**
Deprivation quintile (4)	**0.1666[a] (0.0432)**	**0.1519[a] (0.0400)**
Deprivation quintile (5)	0.0716 (0.0630)	0.1126 (0.0627)
Equivalised income	**0.0773[a] (0.0120)**	**0.0349[c] (0.0148)**
Linear basis term of Age (in year)	-0.5654 (0.7325)	-0.8248 (0.7906)
Random effects		
$\hat{\sigma}_0$ (SD of intercepts Age in year)	0.2961	0.2544
$\hat{\sigma}_1$ (SD of slopes of the Age in year)	0.0125	0.0111
$\hat{\rho}_{01}$ (Correlation of intercepts and slopes)	-0.0024	-0.2784

[a] $p < 0.001$, [b] $p < 0.005$, [c] $p < 0.05$

† The reference categories are given in Tables 2.9 to 2.11.

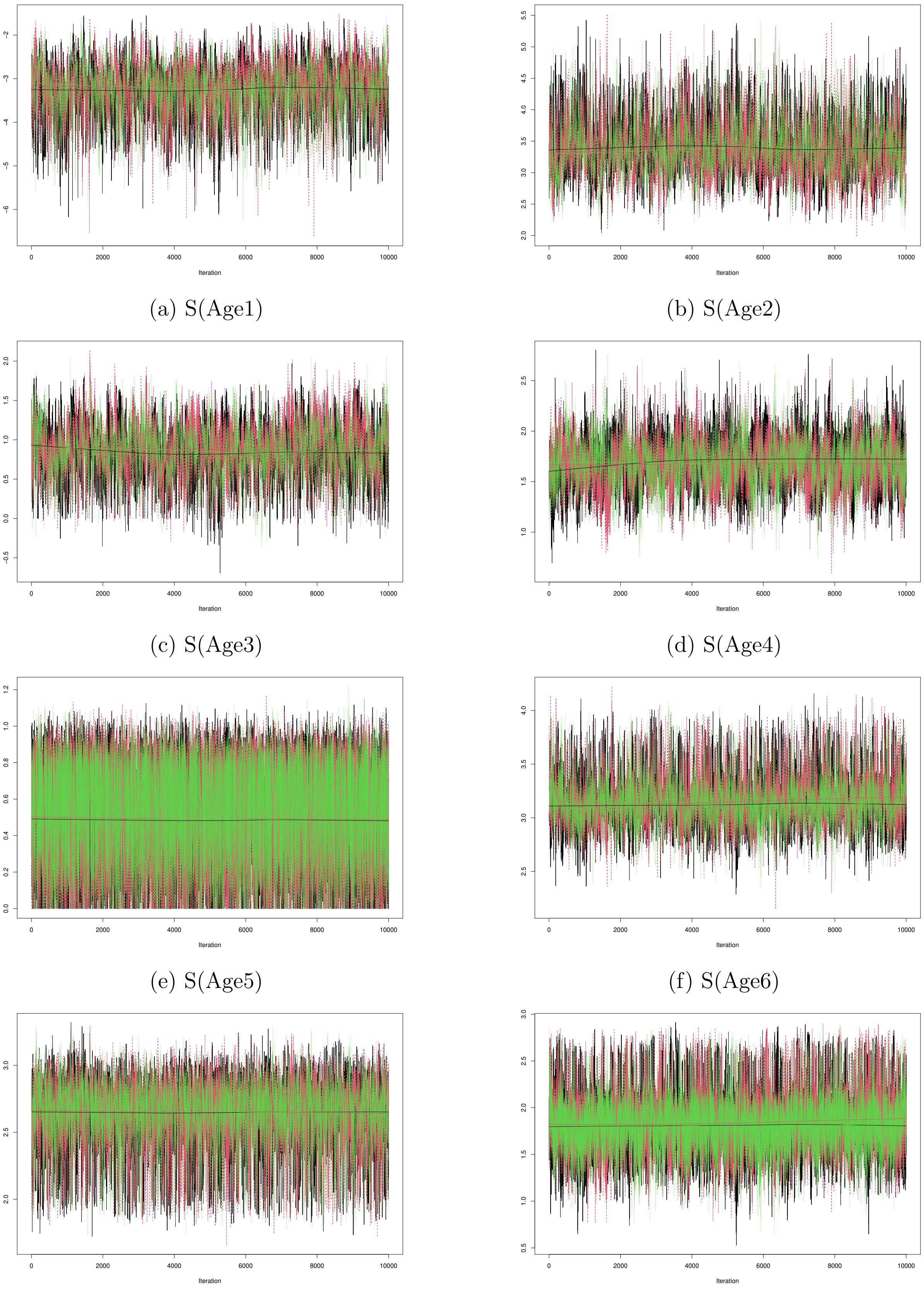
D.3 Convergence diagnostic for basis functions of age

In this section, the trace plots of parameter values sampled across iterations of the MCMC algorithm (Gibbs sampler) for each basis function of age from the BSGSSMQR approach are presented. The corresponding figures and tables are listed in the table below.

Table/Figure	**Physical growth measurement**	**Quantile**
Table D.5	Centred raw weight	0.10
Figure D.7	Centred raw weight	0.10
Table D.6	Centred raw weight	0.90
Figure D.8	Centred raw weight	0.90
Table D.7	Centred WAZ	0.10
Figure D.9	Centred WAZ	0.10
Table D.8	Centred WAZ	0.90
Figure D.10	Centred WAZ	0.90

Table D.5: Summary of the Gelman-Rubin diagnostic ($\hat{R}_{GR}$) and the effective sample size (n_{edf}) for the basis functions of age in the 0.10th quantile model of centred raw weight

Basis functions	$\hat{R}_{GR}$	n_{ESS}	**Basis functions**	$\hat{R}_{GR}$	n_{ESS}
S(Age1)	1.01	1741.96	S(Age13)	1.00	8254.70
S(Age2)	1.02	1767.76	S(Age14)	1.00	4901.42
S(Age3)	1.02	1601.01	S(Age15)	1.00	7452.05
S(Age4)	1.01	1941.03	S(Age16)	1.00	2406.93
S(Age5)	1.00	14994.25	S(Age17)	1.00	7701.98
S(Age6)	1.01	2361.61	S(Age18)	1.00	9210.66
S(Age7)	1.00	6414.97	S(Age19)	1.00	4713.92
S(Age8)	1.00	4135.52	S(Age20)	1.00	6766.25
S(Age9)	1.00	4777.40	S(Age21)	1.00	19606.34
S(Age10)	1.00	2868.32	S(Age22)	1.00	7782.71
S(Age11)	1.00	17641.05	S(Age23)	1.00	3665.70
S(Age12)	1.01	3635.28			



(a) S(Age1)

(b) S(Age2)

(c) S(Age3)

(d) S(Age4)

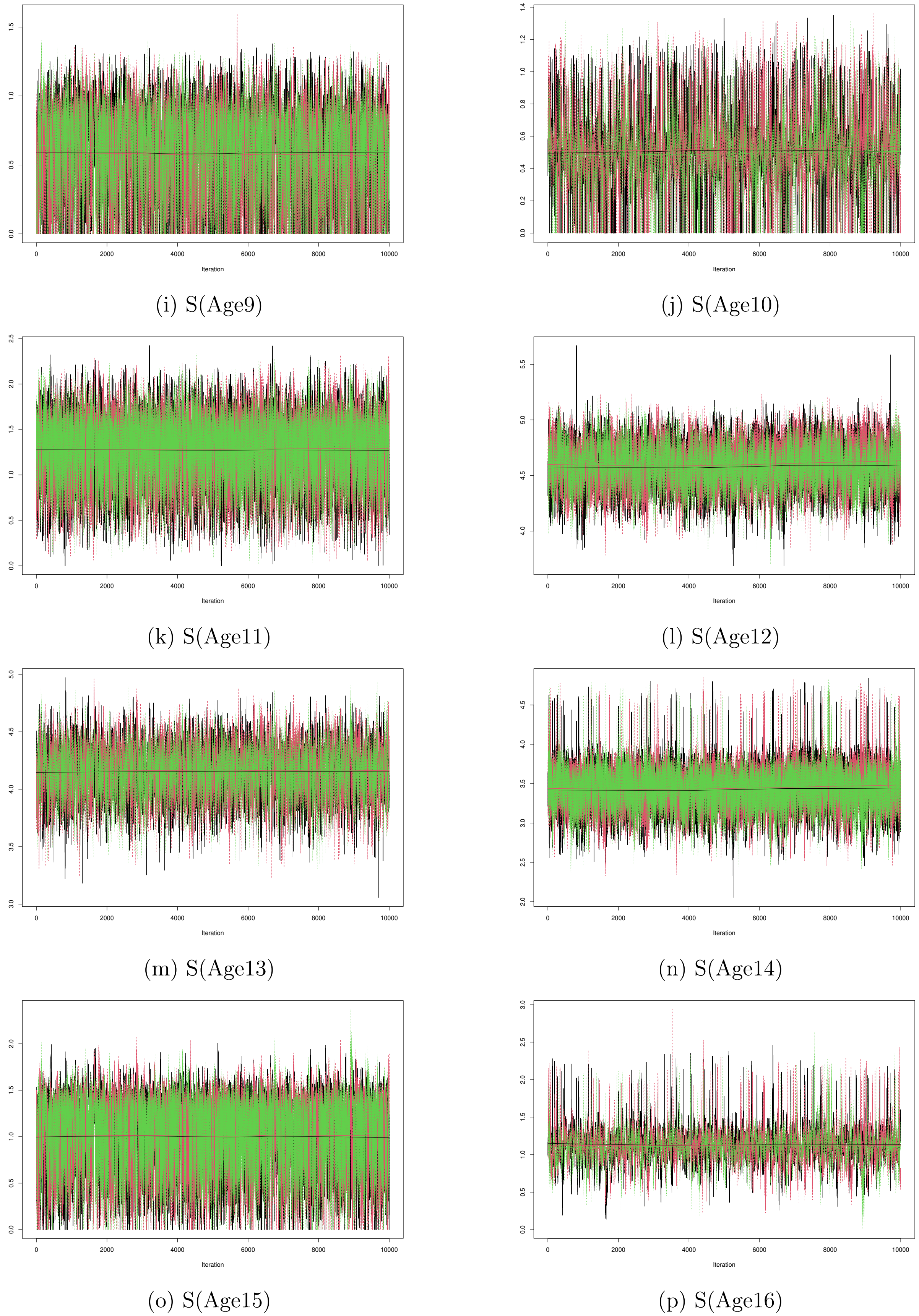
(e) S(Age5)

(f) S(Age6)

(g) S(Age7)

(h) S(Age8)

Figure D.7: Trace plots of MCMC samples for the basis function of age in the 0.10th quantile model for the centred raw weight



(i) S(Age9)

(j) S(Age10)

(k) S(Age11)

(l) S(Age12)

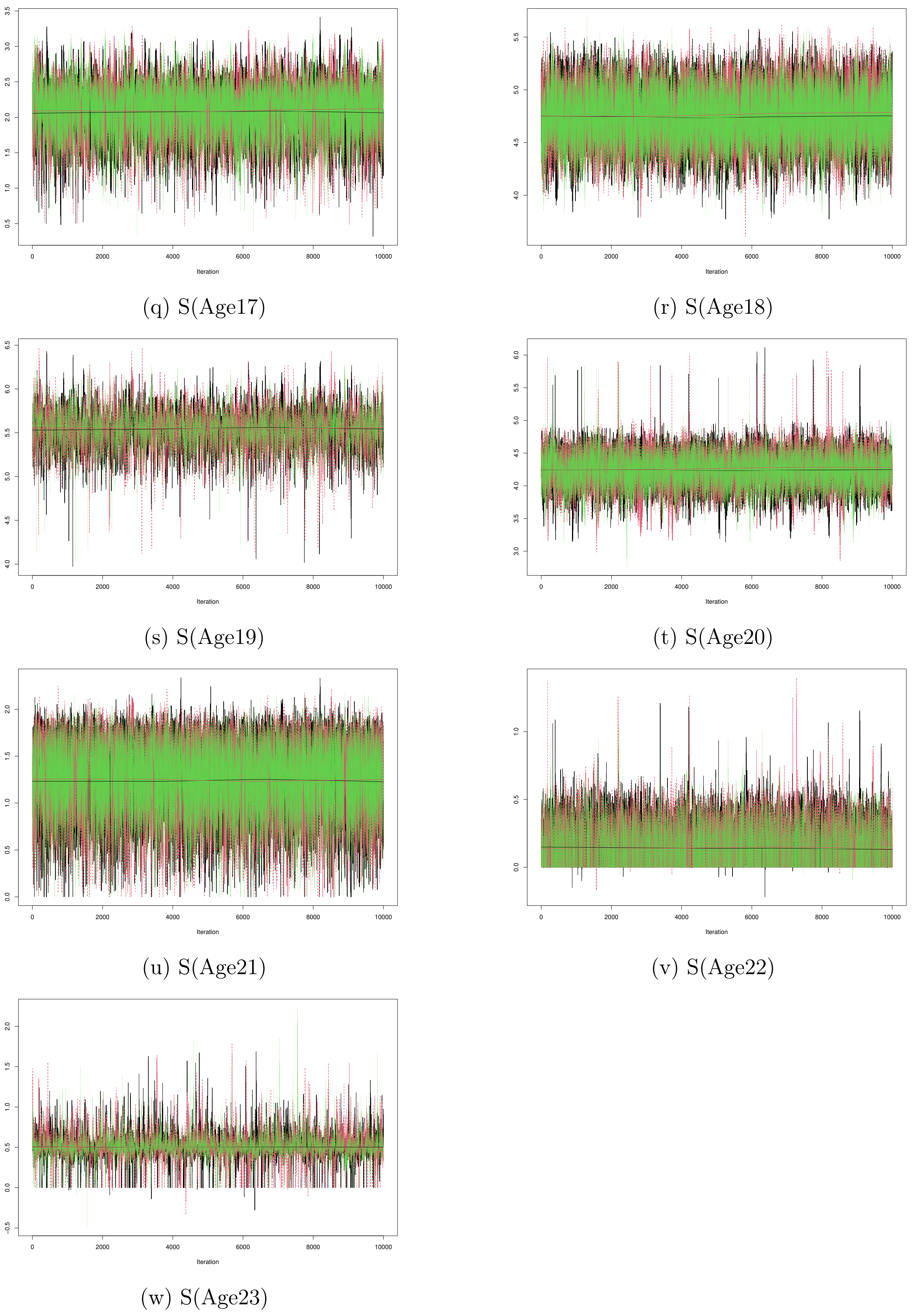
(m) S(Age13)

(n) S(Age14)

(o) S(Age15)

(p) S(Age16)

Figure D.7: Trace plots of MCMC samples for the basis functions of age in the 0.10th quantile model for the centred raw weight



(q) S(Age17)

(r) S(Age18)

(s) S(Age19)

(t) S(Age20)

(u) S(Age21)

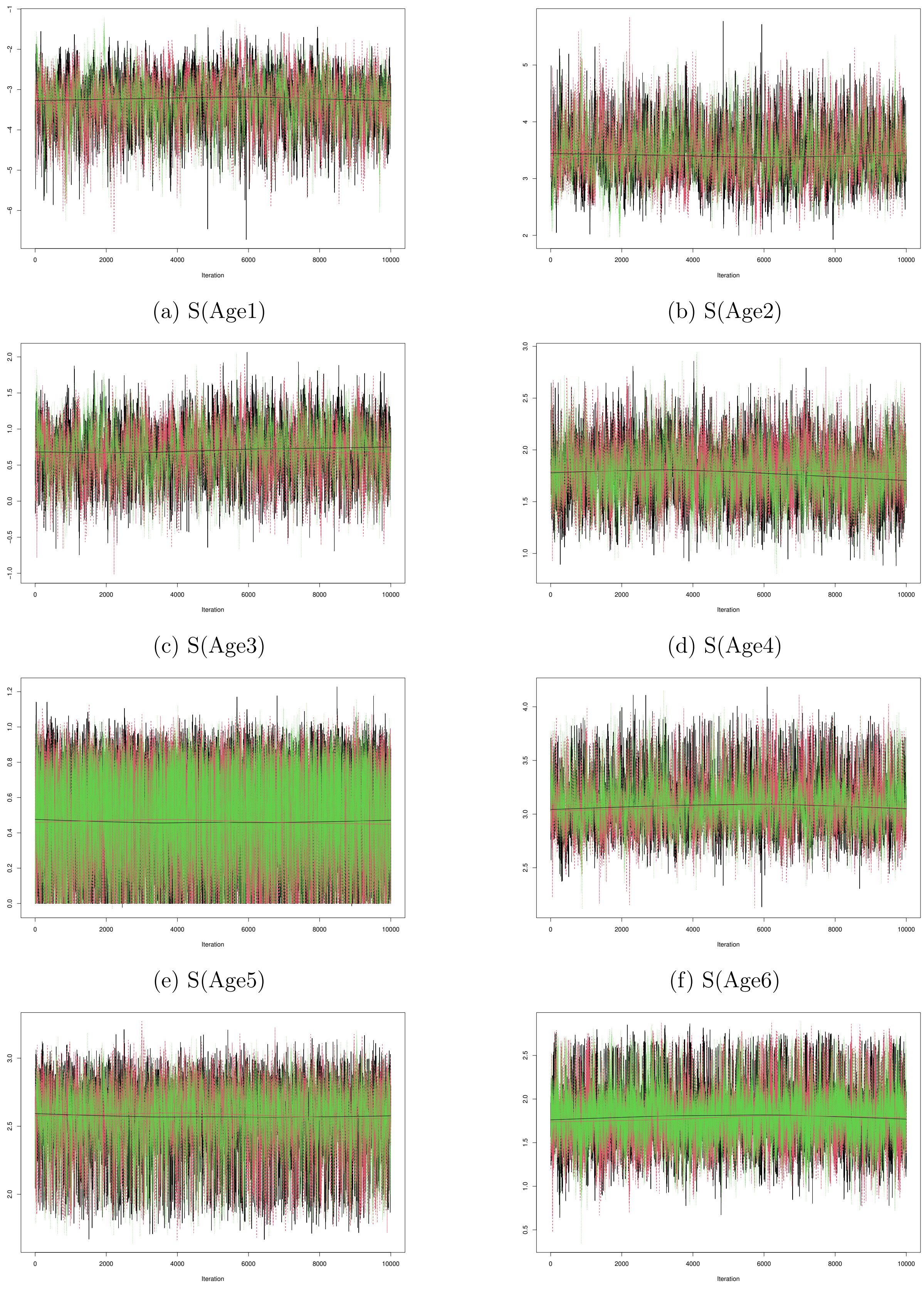
(v) S(Age22)

(w) S(Age23)

Figure D.7: Trace plots of MCMC samples for the basis functions of age in the 0.10th quantile model for the centred raw weight

Table D.6: Summary of the Gelman-Rubin diagnostic ($\hat{R}_{GR}$) and the effective sample size (n_{ESS}) for the basis functions of age in the 0.90th quantile model of centred raw weight

Basis functions	$\hat{R}_{GR}$	n_{ESS}	**Basis functions**	$\hat{R}_{GR}$	n_{ESS}
S(Age1)	1.00	1816.57	S(Age13)	1.00	7088.39
S(Age2)	1.00	1811.73	S(Age14)	1.00	4566.57
S(Age3)	1.00	2097.07	S(Age15)	1.00	7185.10
S(Age4)	1.01	2029.34	S(Age16)	1.00	2749.09
S(Age5)	1.00	11116.86	S(Age17)	1.00	8593.54
S(Age6)	1.00	2421.07	S(Age18)	1.00	10800.54
S(Age7)	1.00	6099.08	S(Age19)	1.00	4616.71
S(Age8)	1.00	4006.02	S(Age20)	1.00	9027.55
S(Age9)	1.00	4949.80	S(Age21)	1.00	19146.58
S(Age10)	1.00	2873.13	S(Age22)	1.00	7995.44
S(Age11)	1.00	16660.90	S(Age23)	1.00	3785.37
S(Age12)	1.00	3639.22			



(a) S(Age1)

(b) S(Age2)

(c) S(Age3)

(d) S(Age4)

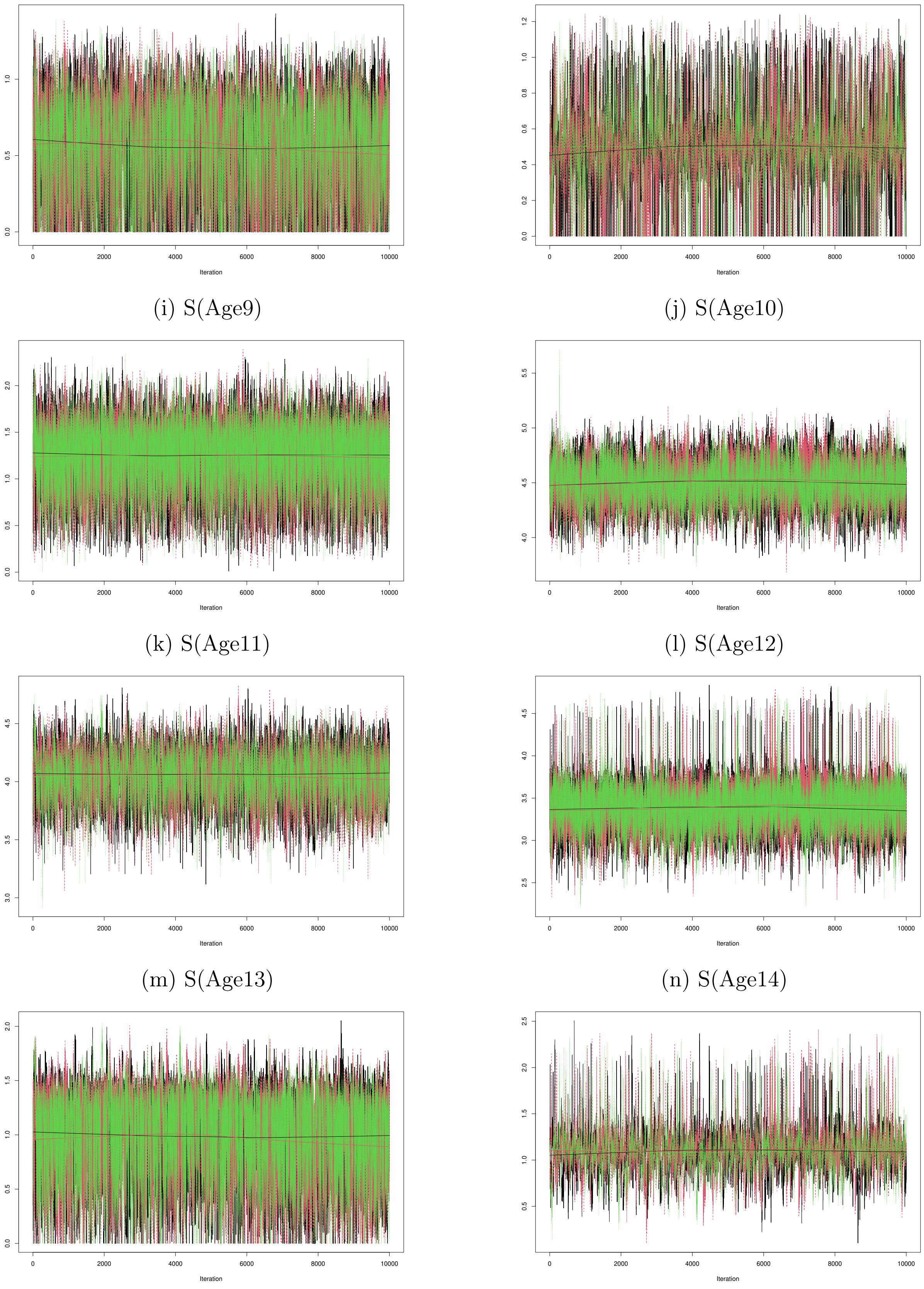
(e) S(Age5)

(f) S(Age6)

(g) S(Age7)

(h) S(Age8)

Figure D.8: Trace plots of MCMC samples for the basis functions of age in the 0.90th quantile model for the centred raw weight



(i) S(Age9)

(j) S(Age10)

(k) S(Age11)

(l) S(Age12)

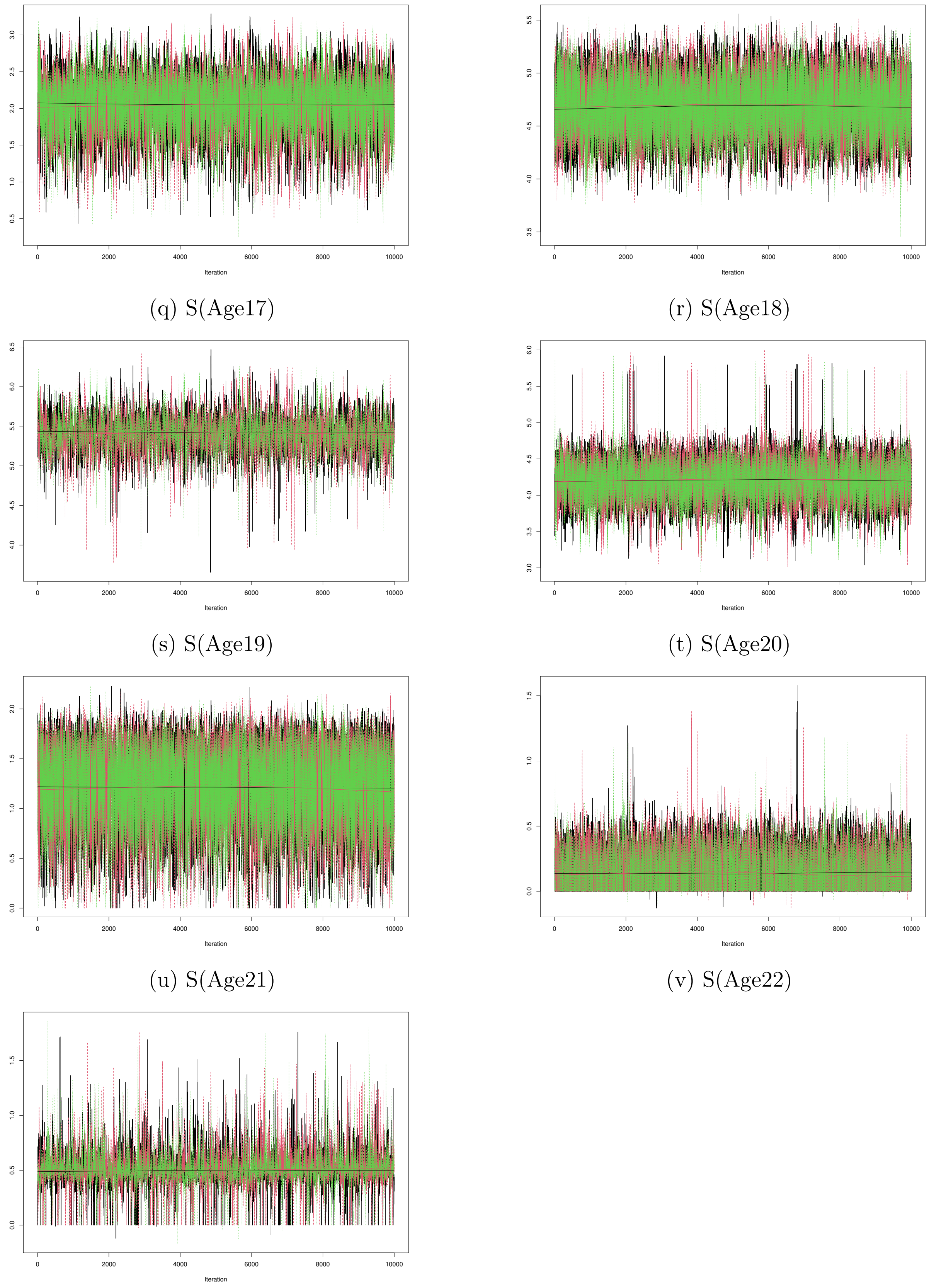
(m) S(Age13)

(n) S(Age14)

(o) S(Age15)

(p) S(Age16)

Figure D.8: Trace plots of MCMC samples for the basis functions of age in the 0.90th quantile model for the centred raw weight



(q) S(Age17)

(r) S(Age18)

(s) S(Age19)

(t) S(Age20)

(u) S(Age21)

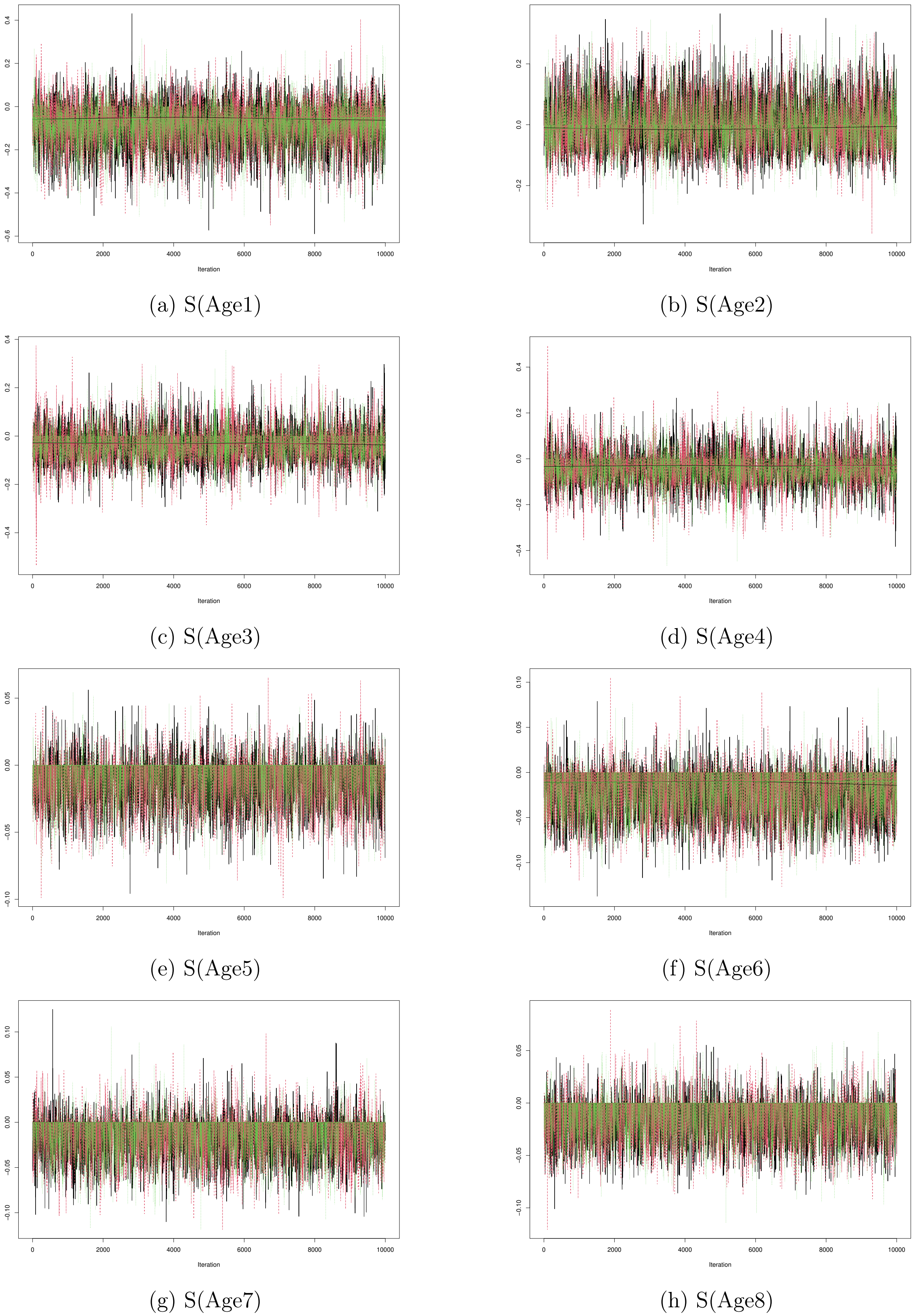
(v) S(Age22)

(w) S(Age23)

Figure D.8: Trace plots of MCMC samples for the basis functions of age in the 0.90th quantile model for the centred raw weight

Table D.7: Summary of the Gelman-Rubin diagnostic ($\hat{R}_{GR}$) and the effective sample size (n_{ESS}) for the basis functions of age in the 0.10th quantile model of centred WAZ

Basis functions	$\hat{R}_{GR}$	n_{ESS}	**Basis functions**	$\hat{R}_{GR}$	n_{ESS}
S(Age1)	1.00	5428.26	S(Age13)	1.00	9221.06
S(Age2)	1.00	7640.17	S(Age14)	1.00	9864.20
S(Age3)	1.00	8155.14	S(Age15)	1.00	15558.88
S(Age4)	1.00	6290.41	S(Age16)	1.01	16913.26
S(Age5)	1.00	12681.58	S(Age17)	1.00	12289.54
S(Age6)	1.00	5180.32	S(Age18)	1.00	13102.25
S(Age7)	1.00	8437.51	S(Age19)	1.00	12971.56
S(Age8)	1.00	8658.72	S(Age20)	1.00	13992.16
S(Age9)	1.00	14997.07	S(Age21)	1.00	15680.10
S(Age10)	1.00	17016.42	S(Age22)	1.00	19253.56
S(Age11)	1.00	12779.09	S(Age23)	1.00	16838.39
S(Age12)	1.00	6930.02			



(a) S(Age1)

(b) S(Age2)

(c) S(Age3)

(d) S(Age4)

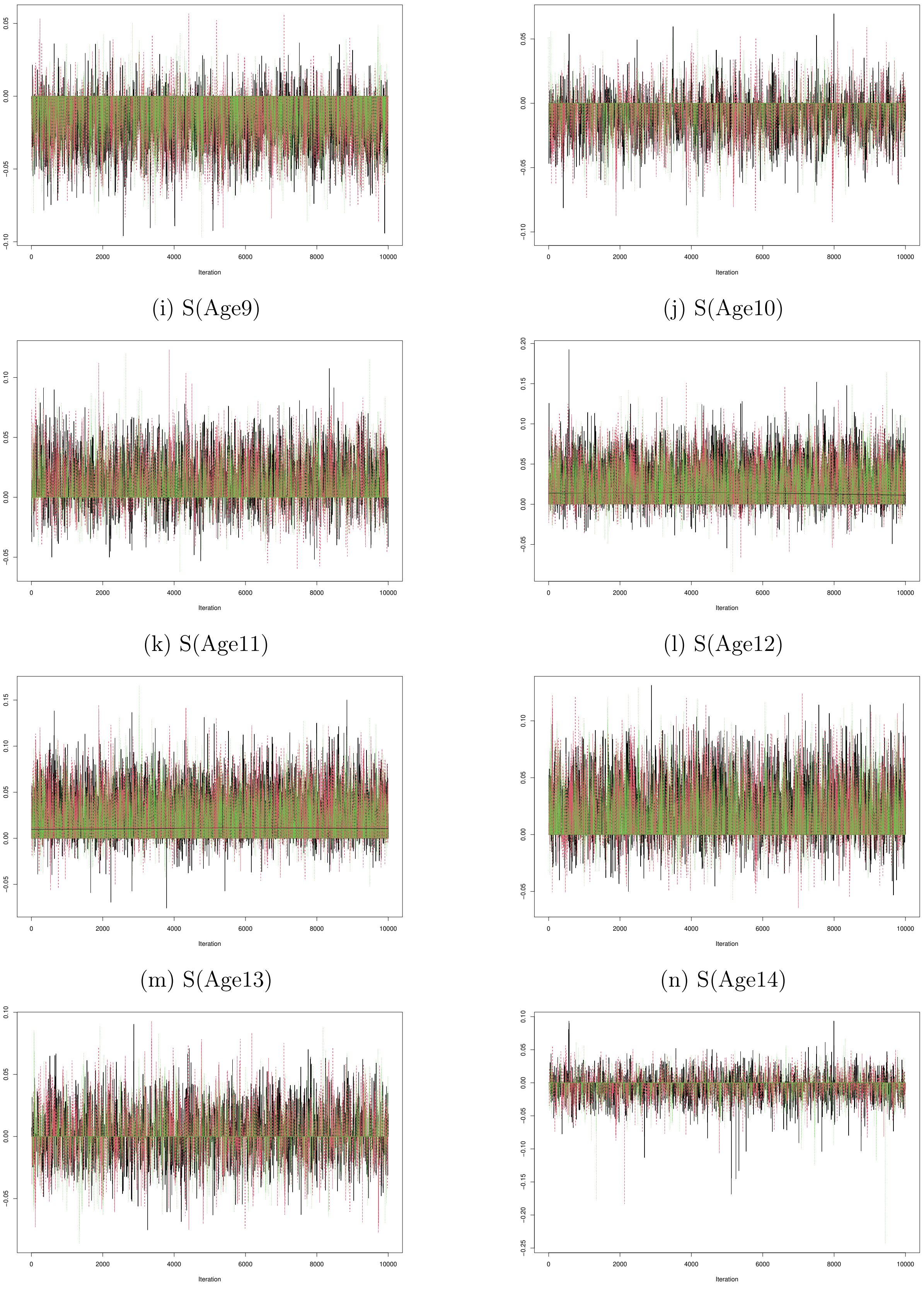
(e) S(Age5)

(f) S(Age6)

(g) S(Age7)

(h) S(Age8)

Figure D.9: Trace plots of MCMC samples for the basis functions of age in the 0.10th quantile model for the centred WAZ



(i) S(Age9)

(j) S(Age10)

(k) S(Age11)

(l) S(Age12)

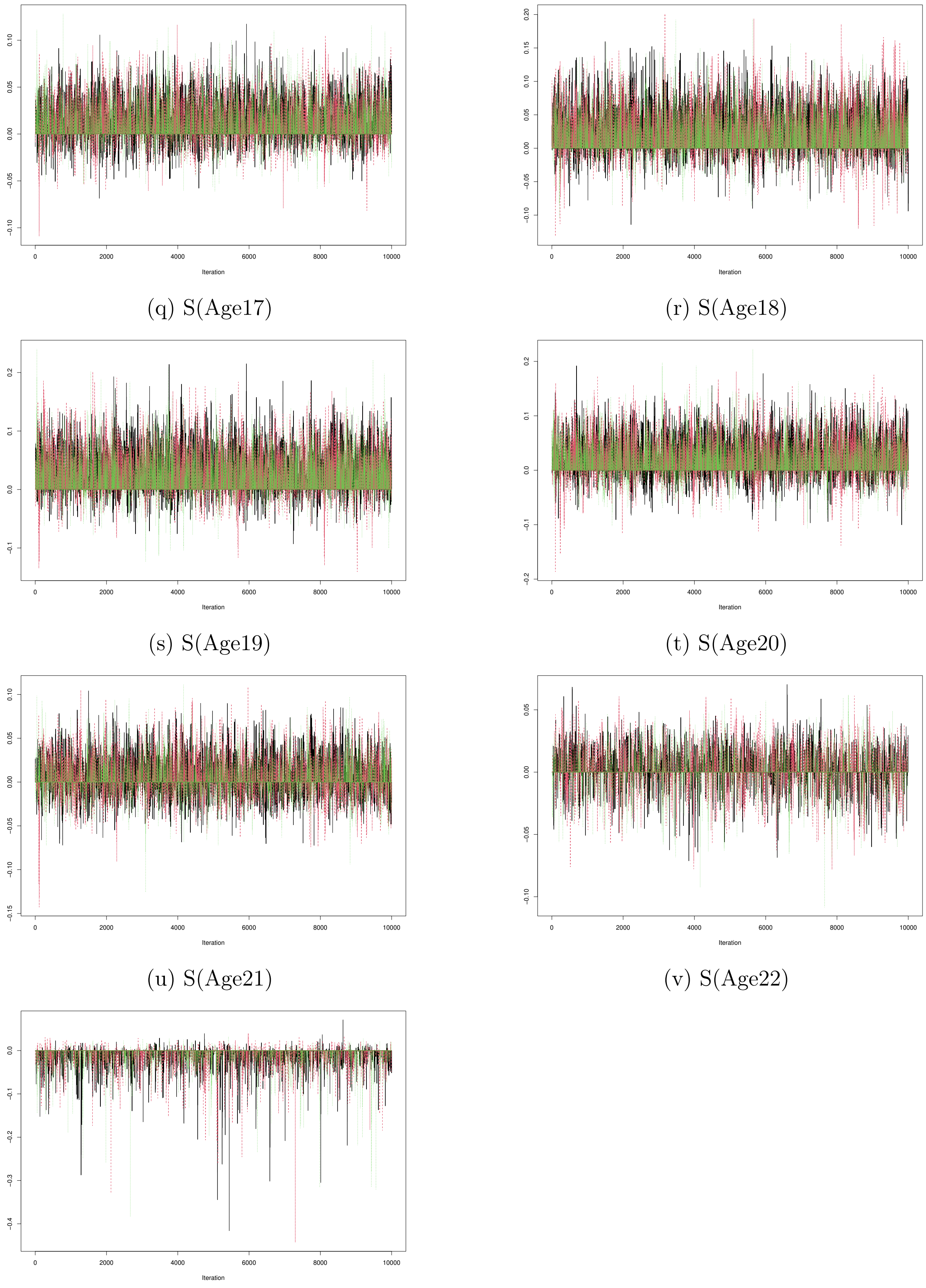
(m) S(Age13)

(n) S(Age14)

(o) S(Age15)

(p) S(Age16)

Figure D.9: Trace plots of MCMC samples for the basis functions of age in the 0.10th quantile model for the centred WAZ



(q) S(Age17)

(r) S(Age18)

(s) S(Age19)

(t) S(Age20)

(u) S(Age21)

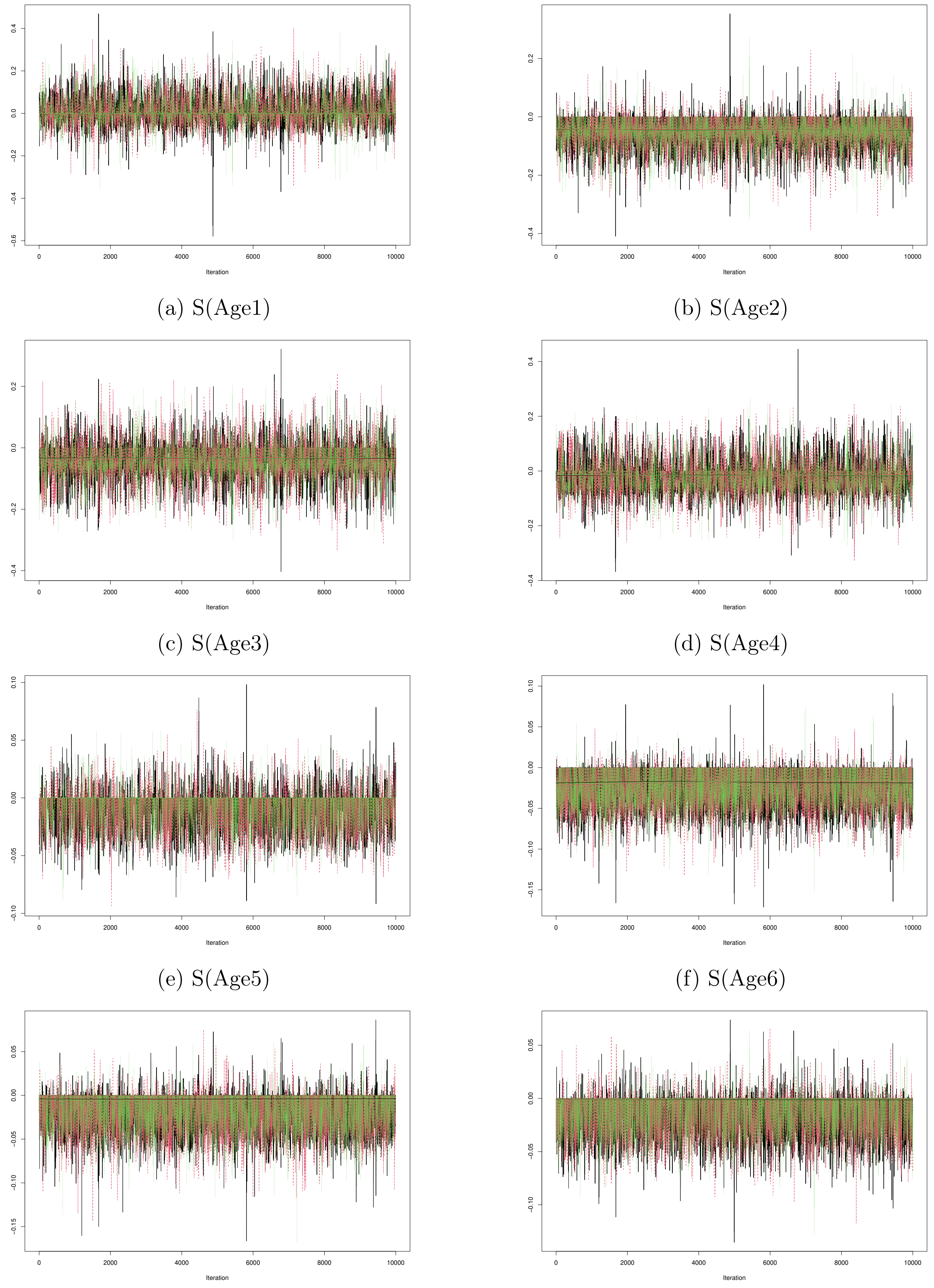
(v) S(Age22)

(w) S(Age23)

Figure D.9: Trace plots of MCMC samples for the basis functions of age in the 0.10th quantile model for the centred WAZ

Table D.8: Summary of the Gelman-Rubin diagnostic ($\hat{R}_{GR}$) and the effective sample size (n_{ESS}) for the basis functions of age in the 0.90th quantile model of centred WAZ

Basis functions	$\hat{R}_{GR}$	n_{ESS}	**Basis functions**	$\hat{R}_{GR}$	n_{ESS}
S(Age1)	1.00	7842.93	S(Age13)	1.00	13798.07
S(Age2)	1.00	6948.66	S(Age14)	1.00	14982.34
S(Age3)	1.00	6216.75	S(Age15)	1.00	16131.35
S(Age4)	1.00	6674.05	S(Age16)	1.00	16132.59
S(Age5)	1.00	11888.85	S(Age17)	1.00	18548.21
S(Age6)	1.00	6219.62	S(Age18)	1.01	16245.99
S(Age7)	1.00	8825.37	S(Age19)	1.00	14808.61
S(Age8)	1.00	10172.28	S(Age20)	1.00	15328.17
S(Age9)	1.00	15374.77	S(Age21)	1.00	15646.25
S(Age10)	1.00	20404.15	S(Age22)	1.00	15548.94
S(Age11)	1.00	18723.43	S(Age23)	1.00	16108.53
S(Age12)	1.00	13568.80			



(a) S(Age1)

(b) S(Age2)

(c) S(Age3)

(d) S(Age4)

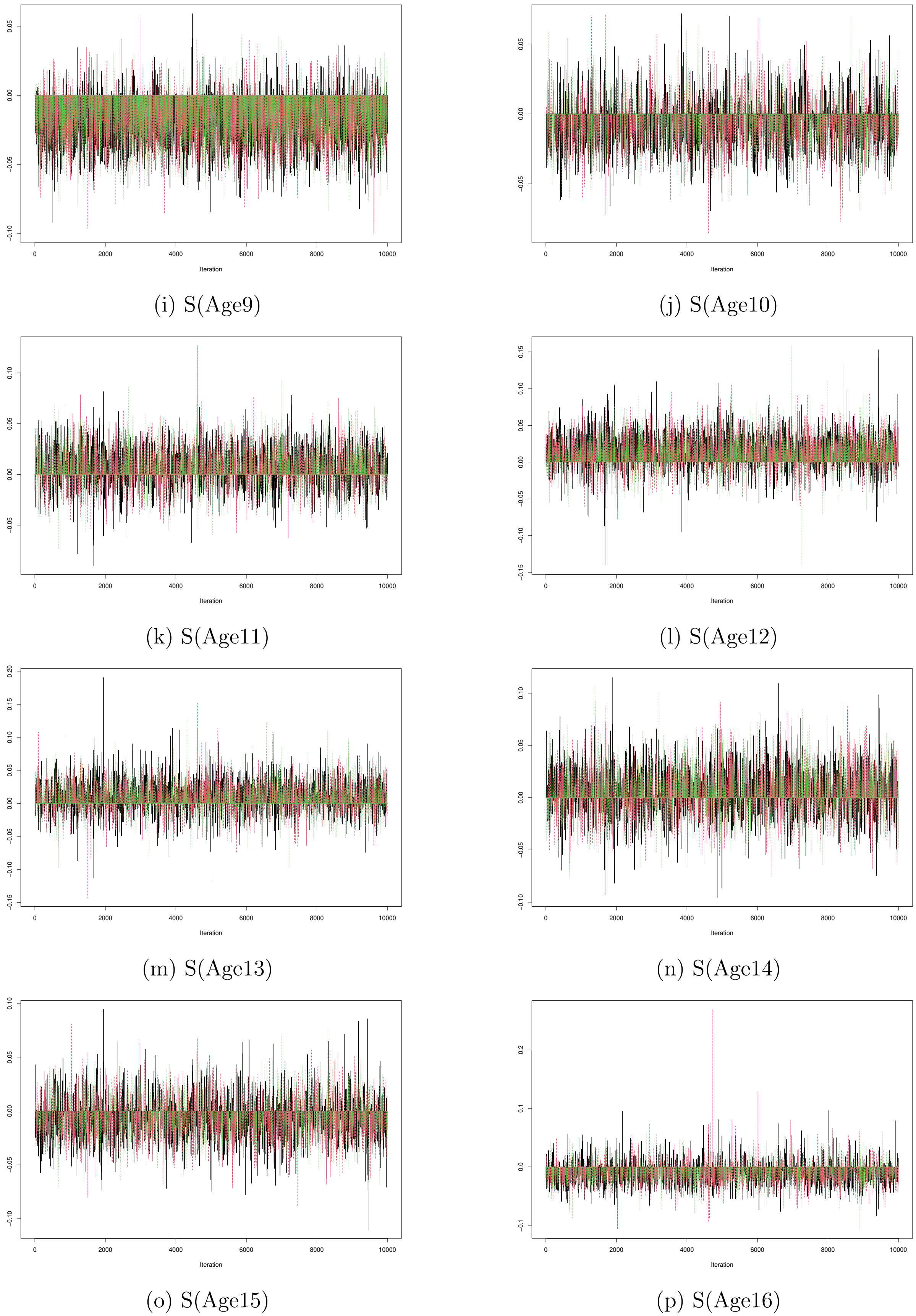
(e) S(Age5)

(f) S(Age6)

(g) S(Age7)

(h) S(Age8)

Figure D.10: Trace plots of MCMC samples for the basis functions of age in the 0.90th quantile model for the centred WAZ



(i) S(Age9)

(j) S(Age10)

(k) S(Age11)

(l) S(Age12)

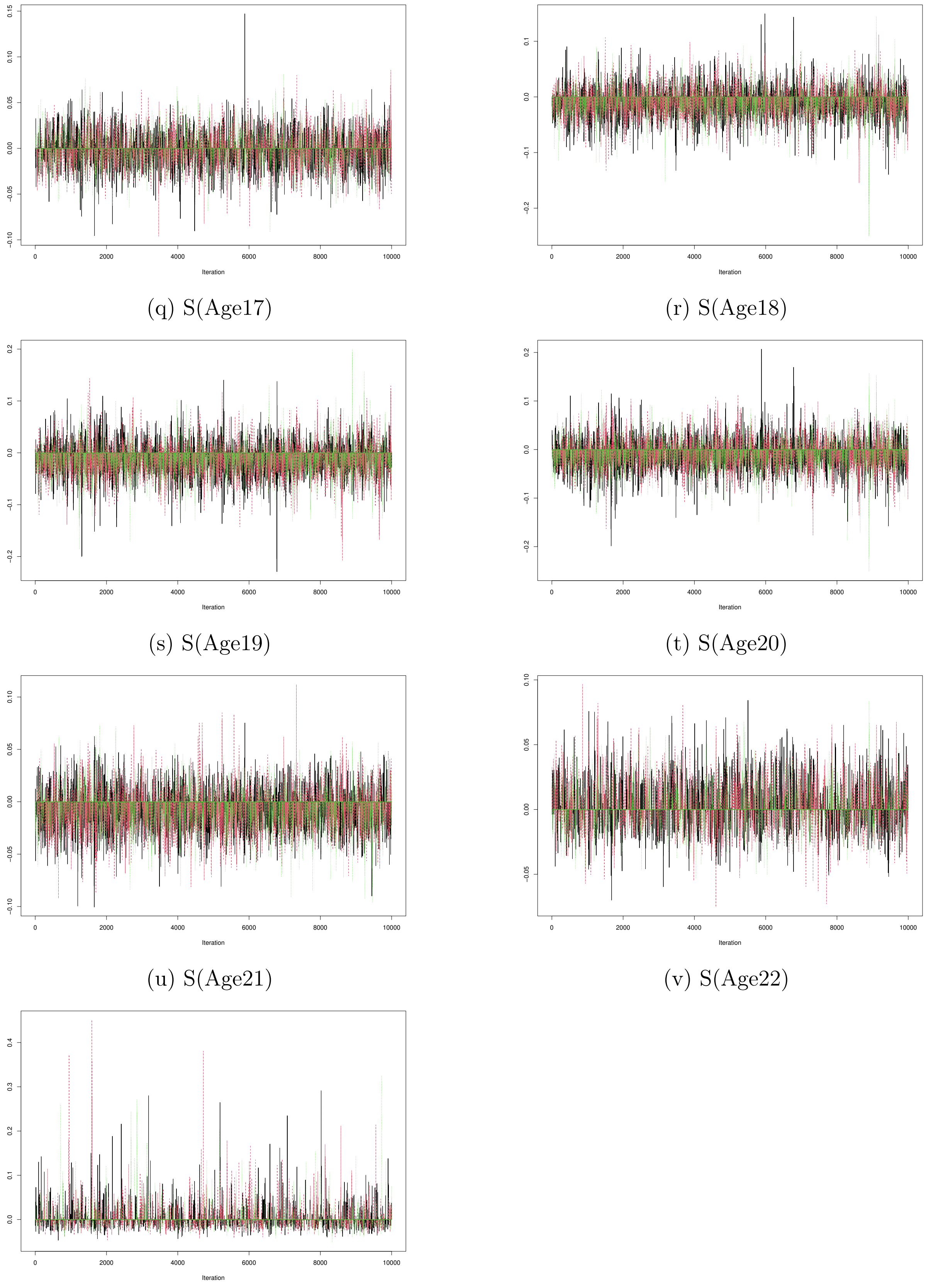
(m) S(Age13)

(n) S(Age14)

(o) S(Age15)

(p) S(Age16)

Figure D.10: Trace plots of MCMC samples for the basis functions of age in the 0.90th quantile model for the centred WAZ



(q) S(Age17)

(r) S(Age18)

(s) S(Age19)

(t) S(Age20)

(u) S(Age21)

(v) S(Age22)

(w) S(Age23)

Figure D.10: Trace plots of MCMC samples for the basis functions of age in the 0.90th quantile model for the centred WAZ

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petran & F. Csaaki (Eds.), *The second international symposium on information theory* (pp. 267–281).

Alhamzawi, R. (2020). *Brq: Bayesian analysis of quantile regression models* [R package version 3.0]. https://CRAN.R-project.org/package=Brq

Alhamzawi, R., & Ali, H. T. M. (2018). The Bayesian adaptive lasso regression. *Mathematical Biosciences*, *303*, 75–82. https://doi.org/10.1016/j.mbs.2018.06.004

Alhamzawi, R., & Ali, H. T. M. (2020). Brq: An R package for Bayesian quantile regression. *METRON*, *78*(3), 313–328. https://doi.org/10.1007/s40300-020-00190-6

Alhamzawi, R., & Yu, K. (2013). Conjugate priors and variable selection for Bayesian quantile regression. *Computational Statistics & Data Analysis*, *64*, 209–219. https://doi.org/10.1016/j.csda.2012.01.014

Alhamzawi, R., & Yu, K. (2014). Bayesian Lasso-mixed quantile regression. *Journal of Statistical Computation and Simulation*, *84*(4), 868–880. https://doi.org/10.1080/00949655.2012.731689

Alhamzawi, R., Yu, K., & Benoit, D. F. (2012). Bayesian adaptive Lasso quantile regression. *Statistical Modelling*, *12*(3), 279–297. https://doi.org/10.1177/1471082X1101200304

Alhamzawi, R. (2013). *Prior elicitation and variable selection for bayesian quantile regression* (Doctoral thesis). Brunel University. https://bura.brunel.ac.uk/handle/2438/7501

Ali, S. (2013). A brief review of risk-factors for growth and developmental delay among preschool children in developing countries. *Advanced Biomedical Research*, *2*(1), 91. https://doi.org/10.4103/2277-9175.122523

Allel, K., Abou Jaoude, G., Poupakis, S., Batura, N., Skordis, J., & Haghparast-Bidgoli, H. (2021). Exploring the associations between early childhood development outcomes and ecological country-level factors across low- and middle-income countries. *International Journal of Environmental Research and Public Health*, *18*(7), 3340. https://doi.org/10.3390/ijerph18073340

Allen, D. B., & Cuttler, L. (2013). Short stature in childhood — challenges and choices. *New England Journal of Medicine*, *368*(13), 1220–1228. https://doi.org/10.1056/NEJMcp1213178

Anderson, C., Hafen, R., Sofrygin, O., Ryan, L., & members of the HBGDki Community. (2019). Comparing predictive abilities of longitudinal child growth models. *Statistics in Medicine*, *38*(19), 3555–3570. https://doi.org/10.1002/sim.7693

Anderson, C., Xiao, L., & Checkley, W. (2019). Using data from multiple studies to develop a child growth correlation matrix. *Statistics in Medicine*, *38*(19), 3540–3554. https://doi.org/10.1002/sim.7696

Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(1), 99–102. https://doi.org/10.1111/j.2517-6161.1974.tb00989.x

Aniley, T. T., Debusho, L. K., Nigusie, Z. M., Yimer, W. K., & Yimer, B. B. (2019). A semi-parametric mixed models for longitudinally measured fasting blood sugar level of adult diabetic patients. *BMC Medical Research Methodology*, *19*(1), 13. https://doi.org/10.1186/s12874-018-0648-x

APS Group Scotland. (2022). *Best start, bright futures: Tackling child poverty. Delivery Plan, 2022-2026.* [OCLC: 1400078702]. The Scottish Government.

Argyle, J., Seheult, A. H., & Wooff, D. A. (2008). Correlation models for monitoring child growth. *Statistics in Medicine*, *27*(6), 888–904. https://doi.org/10.1002/sim.2973

Barker, D. J. (1990). The fetal and infant origins of adult disease. *BMJ*, *301*(6761), 1111–1111. https://doi.org/10.1136/bmj.301.6761.1111

Bayes, M., & Price, M. (1763). An essay towards solving a problem in the doctrine of chances. By the late rev. Mr. Bayes, f. R. S. Communicated by mr. Price, in a letter to john canton, a. M. F. R. S. *Philosophical Transactions (1683-1775)*, *53*, 370–418. Retrieved November 24, 2023, from https://www.jstor.org/stable/105741

Beath, K. J. (2007). Infant growth modelling using a shape invariant model with random effects. *Statistics in Medicine*, *26*(12), 2547–2564. https://doi.org/10.1002/sim.2718

Benoit, D. F., & Poel, D. V. d. (2017). Bayesqr: A bayesian approach to quantile regression. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i07

Berk, L. E. (2013). *Child development* (9th ed.) [OCLC: 774697228]. Pearson.

Berk, L. E., & Meyers, A. B. (2016). *Infants, children, and adolescents* (Eighth edition). Pearson.

Bertino, E., Milani, S., Fabris, C., & De Curtis, M. (2007). Neonatal anthropometric charts: What they are, what they are not. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, *92*(1), F7–F10. https://doi.org/10.1136/adc.2006.096214

Biljon, N. v., Lake, M. T., Goddard, L., Botha, M., Zar, H. J., & Little, F. (2023). Latent classes of anthropometric growth in early childhood using uni- and multivariate approaches in a south african birth cohort. https://doi.org/10.1101/2023.09.01.23294932

Birtwistle, S., Deakin, E., Whitford, R., Hinchliffe, S., Daniels-Creasey, A., & Rule, S. (2022). *The Scottish Health Survey* (Main Report Volume 1). The Scottish Government. Edinburgh.

Black, R. E., Allen, L. H., Bhutta, Z. A., Caulfield, L. E., De Onis, M., Ezzati, M., Mathers, C., & Rivera, J. (2008). Maternal and child undernutrition: Global and regional exposures and health consequences. *The Lancet*, *371*(9608), 243–260. https://doi.org/10.1016/S0140-6736(07)61690-0

Bornstein, M. H., & Lamb, M. E. (Eds.). (2015). *Developmental science: An advanced textbook* (Seventh edition). Psychology Press, Taylor & Francis Group.

Bowditch, H. (1891). The growth of children studied by galton's percentile grades in 22nd annual report of the state board of health of massachusetts (pp. 479–525). *Boston, MA: Wright and Potter*.

Bronfenbrenner, U. (1996). *The ecology of human development: Experiments by nature and design*. Harvard University Press.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 135–167. https://doi.org/10.1037/0033-2909.101.1.147

Bukatko, D., & Daehler, M. W. (1995). *Child development: A thematic approach* (2nd ed). Houghton Mifflin.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Cameron, N., & Bogin, B. (Eds.). (2012). *Human growth and development* (Second edition.) [OCLC: 795120449]. Academic Press.

Carel, J.-C., & Léger, J. (2008). Precocious puberty. *New England Journal of Medicine*, *358*(22), 2366–2377. https://doi.org/10.1056/NEJMcp0800459

Carey, V. J., Yong, F. H., Frenkel, L. M., & McKinney, R. M. (2004). Growth velocity assessment in paediatric AIDS: Smoothing, penalized quantile regression and the definition of growth failure: GROWTH FAILURE IN PAEDIATRIC AIDS. *Statistics in Medicine*, *23*(3), 509–526. https://doi.org/10.1002/sim.1578

Casella, G. (2001). Empirical bayes gibbs sampling. *Biostatistics*, *2*(4), 485–500. https://doi.org/10.1093/biostatistics/2.4.485

Cavanaugh, J. C., & Blanchard-Fields, F. (2015). *Adult development and aging* (7th Edition) [OCLC: ocn880336988]. Cengage Learning.

Chatterjee, A., & Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, *106*(494), 608–625. https://doi.org/10.1198/jasa.2011.tm10159

Chaudhuri, P. (1991). Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of Multivariate Analysis*, *39*(2), 246–269. https://doi.org/10.1016/0047-259X(91)90100-G

Cohen, P., Rogol, A. D., Deal, C. L., Saenger, P., Reiter, E. O., Ross, J. L., Chernausek, S. D., Savage, M. O., Wit, J. M., & on behalf of the 2007 ISS Consensus Workshop participants. (2008). Consensus statement on the diagnosis and treatment of children with idiopathic short stature: A summary of the growth hormone research society, the lawson wilkins pediatric endocrine society, and the european society for paediatric endocrinology workshop. *The Journal of Clinical Endocrinology & Metabolism*, *93*(11), 4210–4217. https://doi.org/10.1210/jc.2008-0509

Cole, T. J. (1988). Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *151*(3), 385–418. https://doi.org/10.2307/2982992

Cole, T. J. (1994). Growth charts for both cross-sectional and longitudinal data. *Statistics in Medicine*, *13*(23-24), 2477–2492. https://doi.org/10.1002/sim.4780132311

Cole, T. J. (1995). Conditional reference charts to assess weight gain in British infants. *Archives of Disease in Childhood*, *73*(1), 8–16. https://doi.org/10.1136/adc.73.1.8

Cole, T. J. (1997). Growth monitoring with the British 1990 growth reference. *Archives of Disease in Childhood*, *76*(1), 47–49. https://doi.org/10.1136/adc.76.1.47

Cole, T. J. (1998). Presenting information on growth distance and conditional velocity in one chart: Practical issues of chart design. *Statistics in Medicine*, *17*(23), 2697–2707. https://doi.org/10.1002/(SICI)1097-0258(19981215)17:23<2697::AID-SIM36>3.0.CO;2-O

Cole, T. J. (2012). The development of growth references and growth charts. *Annals of Human Biology*, *39*(5), 382–394. https://doi.org/10.3109/03014460.2012.694475

Cole, T. J. (2022). *Sitar: Super imposition by translation and rotation growth curve analysis* [R package version 1.3.0]. https://CRAN.R-project.org/package=sitar

Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: The lms method and penalized likelihood. *Statistics in Medicine*, *11*(10), 1305–1319. https://doi.org/10.1002/sim.4780111005

Cole, T. J., Donaldson, M. D. C., & Ben-Shlomo, Y. (2010). SITAR—a useful instrument for growth curve analysis. *International Journal of Epidemiology*, *39*(6), 1558–1566. https://doi.org/10.1093/ije/dyq115

Cole, T. J., Williams, A. F., & Wright, C. M. (2011). Revised birth centiles for weight, length and head circumference in the UK-WHO growth charts. *Annals of Human Biology*, *38*(1), 7–11. https://doi.org/10.3109/03014460.2011.544139

Congdon, P. (2006). *Bayesian statistical modelling* (2nd). John Wiley & Sons.

Connelly, R., & Platt, L. (2014). Cohort profile: Uk millennium cohort study(Mcs). *International Journal of Epidemiology*, *43*(6), 1719–1725. https://doi.org/10.1093/ije/dyu001

Corbeil, R. R., & Searle, S. R. (1976). Restricted maximum likelihood (Reml) estimation of variance components in the mixed model. *Technometrics*, *18*(1), 31. https://doi.org/10.2307/1267913

Craik, F. I. M., & Salthouse, T. A. (Eds.). (2015). *The handbook of aging and cognition* (Third edition.) [OCLC: 1101265439]. Routledge.

Craven, P., & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, *31*(4), 377–403. https://doi.org/10.1007/BF01404567

Currie, I. D., & Durban, M. (2002). Flexible smoothing with P-splines: A unified approach. *Statistical Modelling*, *2*(4), 333–349. https://doi.org/10.1191/1471082x02st039ob

De Sanctis, V., Soliman, A., Alaaraj, N., Ahmed, S., Alyafei, F., & Hamed, N. (2021). Early and long-term consequences of nutritional stunting: From childhood to adulthood: Early and long-term consequences of nutritional stunting. *Acta Bio Medica Atenei Parmensis*, *92*(1), 11346. https://doi.org/10.23750/abm.v92i1.11346

de Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, *6*(1), 50–62. https://doi.org/10.1016/0021-9045(72)90080-9

Demidenko, E. (2013). *Mixed models: Theory and applications with R* (Second [edition]). Wiley.

Denham, S. A. (2006). Social-emotional competence as support for school readiness: What is it and how do we assess it? *Early Education & Development*, *17*(1), 57–89. https://doi.org/10.1207/s15566935eed1701_4

Diggle, P. J. (2005). Variogram. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (b2a12078). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470011815.b2a12078

Durbán, M., Harezlak, J., Wand, M. P., & Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, *24*(8), 1153–1167. https://doi.org/10.1002/sim.1991

Edwards, M. (2017). The barker hypothesis. In V. Preedy & V. B. Patel (Eds.), *Handbook of Famine, Starvation, and Nutrient Deprivation: From Biology to Policy* (pp. 1–

21). Springer International Publishing. https://doi.org/10.1007/978-3-319-40007-5_71-1

Eilers, P. H. C. (1999). Comment on: The analysis of designed experiments and longitudinal data by using smoothing splines (by verbylaet al.) *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(3), 307–308.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. *Statistical Science*, *11*(2), 89–121. https://doi.org/10.1214/ss/1038425655

Eilers, P. H. C., & Marx, B. D. (2010). Splines, knots, and penalties. *WIREs Computational Statistics*, *2*(6), 637–653. https://doi.org/10.1002/wics.125

Eilers, P. H. C., Marx, B. D., & Durbán, M. (2015). Twenty years of p-splines. *Sort-statistics and Operations Research Transactions*, *39*, 149–186.

Eilers, P. H., & Marx, B. D. (2021). *Practical smoothing: The joys of p-splines* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108610247

Fang, X., Zuo, J., Zhou, J., Cai, J., Chen, C., Xiang, E., Li, H., Cheng, X., & Chen, P. (2019). Childhood obesity leads to adult type 2 diabetes and coronary artery diseases: A 2-sample mendelian randomization study. *Medicine*, *98*(32), e16825. https://doi.org/10.1097/MD.0000000000016825

Feng, Y., Xiao, L., Li, C., Chen, S. T., & Ohuma, E. O. (2020). Correlation models for monitoring fetal growth. *Statistical Methods in Medical Research*, *29*(10), 2795–2813. https://doi.org/10.1177/0962280220905623

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd) [OCLC: ocn708358043]. Wiley.

Fitzmaurice, G. M., & Ravichandran, C. (2008). A primer in longitudinal data analysis. *Circulation*, *118*(19), 2005–2010. https://doi.org/10.1161/CIRCULATIONAHA.107.714618

Fogel, A., King, B. J., & Shanker, S. G. (2008). *Human development in the twenty-first century: Visionary ideas from systems scientists* [OCLC: 776973941]. Cambridge University Press.

for Health Statistics, N. C. et al. (1977). Nchs growth curves for children 0-18 years (1977). *United States, Vital and Health Statistics*.

Fu, W., & Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*(5). https://doi.org/10.1214/aos/1015957397

Galton, F. (1883). *Inquiries into human faculty and its development*. Macmillan. https://archive.org/details/inquiriesintohum00galt

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on article by browne and draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A. (2014). *Bayesian data analysis* (3rd). CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4). https://doi.org/10.1214/ss/1177011136

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Geraci, M., & Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, *8*(1), 140–154. https://doi.org/10.1093/biostatistics/kxj039

Geraci, M. (2019). Additive quantile regression for clustered data with an application to children's physical activity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(4), 1071–1089. https://doi.org/10.1111/rssc.12333

Geraci, M., & Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, *24*(3), 461–479. https://doi.org/10.1007/s11222-013-9381-9

Goldstein, H. (1989). Models for multilevel response variables with an application to growth curves. *Multilevel Analysis of Educational Data* (pp. 107–125). Elsevier. https://doi.org/10.1016/B978-0-12-108840-8.50011-1

Grajeda, L. M., Ivanescu, A., Saito, M., Crainiceanu, C., Jaganath, D., Gilman, R. H., Crabtree, J. E., Kelleher, D., Cabrera, L., Cama, V., & Checkley, W. (2016). Modelling subject-specific childhood growth using linear mixed-effect models with cubic regression splines. *Emerging Themes in Epidemiology*, *13*(1), 1. https://doi.org/10.1186/s12982-015-0038-3

Grammer, J. K., Coffman, J. L., Ornstein, P. A., & Morrison, F. J. (2013). Change over time: Conducting longitudinal studies of children's cognitive development. *Journal of Cognition and Development*, *14*(4), 515–528. https://doi.org/10.1080/15248372.2013.833925

Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach* (1st). Chapman & Hall.

Guo, S. S., & Chumlea, W. C. (1999). Tracking of body mass index in children in relation to overweight in adulthood. *The American Journal of Clinical Nutrition*, *70*(1), 145S–148S. https://doi.org/10.1093/ajcn/70.1.145s

Hardin, J. W., & Hilbe, J. M. (2013). *Generalized estimating equations* (Second edition). CRC Press.

Harrell, J. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed. 2015). Springer International Publishing : Imprint: Springer.

Hastie, T., & Tibshirani, R. (1999). *Generalized additive models*. Chapman & Hall/CRC.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. CRC Press, Taylor & Francis Group.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

He, X., & Ng, P. (1999). COBS: Qualitatively constrained smoothing via linear programming. *Computational Statistics*, *14*(3), 315–337. https://doi.org/10.1007/s001800050019

He, X., & Shi, P. (1994). Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, *3*(3-4), 299–308. https://doi.org/10.1080/10485259408832589

Heckman, N., Lockhart, R., & Nielsen, J. D. (2013). Penalized regression, mixed effects models and appropriate modelling. *Electronic Journal of Statistics*, *7*(0), 1517–1552. https://doi.org/10.1214/13-EJS809

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (Ema) data. *Biometrics*, *64*(2), 627–634. https://doi.org/10.1111/j.1541-0420.2007.00924.x

Hendricks, W., & Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, *87*(417), 58–68. https://doi.org/10.1080/01621459.1992.10475175

Hermanussen, M., & Bogin, B. (2014). Auxology – an editorial. *Italian Journal of Pediatrics*, *40*(1), 8. https://doi.org/10.1186/1824-7288-40-8

Hurvich, C. M., Simonoff, J. S., & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *60*(2), 271–293. https://doi.org/10.1111/1467-9868.00125

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. https://doi.org/10.1093/biomet/76.2.297

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773. Retrieved July 3, 2023, from https://www.jstor.org/stable/3448605

Jacobs, P. A., & Lewis, P. A. W. (1978). Discrete time series generated by mixtures. I: Correlational and runs properties. *Journal of the Royal Statistical Society. Series B (Methodological)*, *40*(1), 94–105. Retrieved November 26, 2023, from https://www.jstor.org/stable/2984870

Jamrozik, J., Bohmanova, J., & Schaeffer, L. (2010). Selection of locations of knots for linear splines in random regression test-day models. *Journal of Animal Breeding and Genetics*, *127*(2), 87–92. https://doi.org/10.1111/j.1439-0388.2009.00829.x

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*(1007), 453–461. https://doi.org/10.1098/rspa.1946.0056

Jensen, L. A. (2016). *The Oxford handbook of human development and culture: An interdisciplinary perspective*. Oxford university press.

Ji, Y., & Shi, H. (2022). Shrinkage estimation of fixed and random effects in linear quantile mixed models. *Journal of Applied Statistics*, *49*(14), 3693–3716. https://doi.org/10.1080/02664763.2021.1962262

Johansen, A. (2010). Markov chain monte carlo. *International Encyclopedia of Education* (pp. 245–252). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.01347-6

Johnson, R. C., McClearn, G. E., Yuen, S., Nagoshi, C. T., Ahern, F. M., & Cole, R. E. (1985). Galton's data a century later. *American Psychologist*, *40*(8), 875.

Johnson, W. (2015). Analytical strategies in human growth research. *American Journal of Human Biology*, *27*(1), 69–83. https://doi.org/10.1002/ajhb.22589

Júlíusson, P., Roelants, M., Eide, G., Moster, D., Juul, A., Hauspie, R., Waaler, P., & Bjerknes, R. (2009). Vekstkurver for norske barn. *Tidsskrift for Den norske legeforening*, *129*(4), 281–286. https://doi.org/10.4045/tidsskr.09.32473

Kato, K., Galvao, F., & Montes-Rojas, G. (2012). Asymptotics for panel quantile regression models with individual effects. *Journal of Econometrics*, *170*(1), 76–91. https://doi.org/10.1016/j.jeconom.2012.02.007

Khadilkar, V., & Khadilkar, A. (2011). Growth charts: A diagnostic tool. *Indian Journal of Endocrinology and Metabolism*, *15*(7), 166. https://doi.org/10.4103/2230-8210.84854

Kinney, S. K., & Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, *63*(3), 690–698. https://doi.org/10.1111/j.1541-0420.2007.00771.x

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, *91*(1), 74–89. https://doi.org/10.1016/j.jmva.2004.05.006

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Koenker, R. (2019). *Quantreg: Quantile regression* [R package version 5.54]. https://CRAN.R-project.org/package=quantreg

Koenker, R. (2021). *Quantreg: Quantile regression* [R package version 5.86]. https://CRAN.R-project.org/package=quantreg

Koenker, R., & D'Orey, V. (1987). Algorithm as 229: Computing regression quantiles. *Applied Statistics*, *36*(3), 383. https://doi.org/10.2307/2347802

Koenker, R., & Machado, J. A. F. (1999a). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, *94*(448), 1296–1310. https://doi.org/10.1080/01621459.1999.10473882

Koenker, R., & Machado, J. A. F. (1999b). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, *94*(448), 1296–1310. https://doi.org/10.1080/01621459.1999.10473882

Koenker, R., & Ng, P. (2003). **sparsem** : A sparse matrix package for *r*. *Journal of Statistical Software*, *8*(6). https://doi.org/10.18637/jss.v008.i06

Koenker, R., Ng, P., & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, *81*(4), 673–680. https://doi.org/10.1093/biomet/81.4.673

Koenker, R., & Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, *71*(1-2), 265–283. https://doi.org/10.1016/0304-4076(96)84507-6

Kotelmann, L. W. J. (1879). Die körperverhältnisse der gelehrtenschüler des johanneums in hamburg: Ein statistischer beitrag zur schulhygiene. *Berlin: Zeitschrift des Königlich Preussischen statistischen Bureau's.*

Kotz, S., & Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge University Press.

Kozumi, H., & Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, *81*(11), 1565–1578. https://doi.org/10.1080/00949655.2010.496117

Krebs, N. F., Himes, J. H., Jacobson, D., Nicklas, T. A., Guilday, P., & Styne, D. (2007). Assessment of child and adolescent overweight and obesity. *Pediatrics*, *120*(Supplement_4), S193–S228. https://doi.org/10.1542/peds.2007-2329D

Kuczmarski, R. J. (2000). *Cdc growth charts: United states*. US Department of Health; Human Services, Centers for Disease Control and ...

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411. https://doi.org/10.1214/10-BA607

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963. https://doi.org/10.2307/2529876

Lamarche, C. (2010). Robust penalized quantile regression estimation for panel data. *Journal of Econometrics*, *157*(2), 396–408. https://doi.org/10.1016/j.jeconom.2010.03.042

Lawton, W. H., Sylvestre, E. A., & Maggio, M. S. (1972). Self modeling nonlinear regression. *Technometrics*, *14*(3), 513–532. https://doi.org/10.1080/00401706.1972.10488942

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*(1), 21–59. Retrieved November 1, 2023, from https://www.jstor.org/stable/3533623

Li, Q., Lin, N., & Xi, R. (2010). Bayesian regularized quantile regression. *Bayesian Analysis*, *5*(3). https://doi.org/10.1214/10-BA521

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, *3*(1), 112–115. https://doi.org/10.1111/j.2041-210X.2011.00131.x

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, *19*(13), 1793–1819. https://doi.org/10.1002/1097-0258(20000715)19:13<1793::AID-SIM482>3.0.CO;2-Q

Lu, C., Black, M. M., & Richter, L. M. (2016). Risk of poor development in young children in low-income and middle-income countries: An estimation and analysis at the global, regional, and country level. *The Lancet Global Health*, *4*(12), e916–e922. https://doi.org/10.1016/S2214-109X(16)30266-2

Maringwa, J. T., Geys, H., Shkedy, Z., Faes, C., Molenberghs, G., Aerts, M., Ammel, K. V., Teisman, A., & Bijnens, L. (2008). Application of semiparametric mixed models and simultaneous confidence bands in a cardiovascular safety experiment with longitudinal data. *Journal of Biopharmaceutical Statistics*, *18*(6), 1043–1062. https://doi.org/10.1080/10543400802368881

Marino, M. F., & Farcomeni, A. (2015). Linear quantile regression models for longitudinal experiments: An overview. *METRON*, *73*(2), 229–247. https://doi.org/10.1007/s40300-015-0072-5

Marshall, W. A. (1978). Puberty. In F. Falkner & J. M. Tanner (Eds.), *Human Growth: 2 Postnatal Growth* (pp. 141–181). Springer US. https://doi.org/10.1007/978-1-4684-2622-9_6

Maturana-Russel, P., & Meyer, R. (2021). Bayesian spectral density estimation using P-splines with quantile-based knot placement. *Computational Statistics*, *36*(3), 2055–2077. https://doi.org/10.1007/s00180-021-01066-7

Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(1), 53–71. https://doi.org/10.1111/j.1467-9868.2007.00627.x

Merhi Bleik, J. (2019). Fully bayesian estimation of simultaneous regression quantiles under asymmetric laplace distribution specification. *Journal of Probability and Statistics*, *2019*, 1–12. https://doi.org/10.1155/2019/8610723

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032. https://doi.org/10.1080/01621459.1988.10478694

Muggeo, V. M., Atkins, D. C., Gallop, R. J., & Dimidjian, S. (2014). Segmented mixed models with random changepoints: A maximum likelihood approach with application to treatment for depression study. *Statistical Modelling: An International Journal*, *14*(4), 293–313. https://doi.org/10.1177/1471082X13504721

Muggeo, V. M., Sciandra, M., Tomasello, A., & Calvo, S. (2013). Estimating growth charts via nonparametric quantile regression: A practical framework with application in ecology. *Environmental and Ecological Statistics*, *20*(4), 519–531. https://doi.org/10.1007/s10651-012-0232-1

Mwangome, M. K., & Berkley, J. A. (2014). The reliability of weight-for-length/height z scores in children. *Maternal & Child Nutrition*, *10*(4), 474–480. https://doi.org/10.1111/mcn.12124

Naumova, E. N., Must, A., & Laird, N. M. (2001). Tutorial in Biostatistics: Evaluating the impact of 'critical periods' in longitudinal studies of growth using piecewise mixed effects models. *International Journal of Epidemiology*, *30*(6), 1332–1341. https://doi.org/10.1093/ije/30.6.1332

Ng, P., & Maechler, M. (2007). A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling: An International Journal*, *7*(4), 315–328. https://doi.org/10.1177/1471082X0700700403

Ngo, L., & Wand, M. P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, *9*(1). https://doi.org/10.18637/jss.v009.i01

Papalia, D. E., Olds, S. W., & Feldman, R. D. (2004). *Human development* (9th ed). McGraw-Hill.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554. https://doi.org/10.1093/biomet/58.3.545

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology*, *19*(1), 46. https://doi.org/10.1186/s12874-019-0666-3

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-plus*. Springer.

Portnoy, S., & Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science*, *12*(4), 279–300. https://doi.org/10.1214/ss/1030037960

Pratesi, M., Ranalli, M. G., & Salvati, N. (2009). Nonparametric *M* -quantile regression using penalised splines. *Journal of Nonparametric Statistics*, *21*(3), 287–304. https://doi.org/10.1080/10485250802638290

Quetelet, A. (1832). Nouveaux memoire de l'academie royale des sciences et belles-lettres de bruxelles. *Recherches sur le poids de l'homme aux different ages*.

Revan Özkale, M., & Altuner, H. (2022). Bootstrap confidence interval of ridge regression in linear regression model: A comparative study via a simulation study. *Communications in Statistics - Theory and Methods*, 1–37. https://doi.org/10.1080/03610926.2022.2045024

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (With discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

Rigby, R. A., & Stasinopoulos, D. M. (2004). Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution. *Statistics in Medicine*, *23*(19), 3053–3076. https://doi.org/10.1002/sim.1861

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science*, *6*(1). https://doi.org/10.1214/ss/1177011926

Roche, A. F. (1992). *Growth, maturation, and body composition: The fels longitudinal study 1929-1991*. Cambridge University Press.

Roelants, M., Hauspie, R., & Hoppenbrouwers, K. (2009). References for growth and pubertal development from birth to 21 years in flanders, belgium. *Annals of human biology*, *36*(6), 680–694.

Rogoff, B. (2003). *The cultural nature of human development*. Oxford University Press.

Rosario, A. S., Schienkiewitz, A., & Neuhauser, H. (2014). German height references for children aged 0 to under 18 years compared to who and cdc growth charts (vol 38, pg 121, 2011). *ANNALS OF HUMAN BIOLOGY*, *41*(4), 381–381.

Royal College of Paediatrics and Child Health. (2013). *Boys uk-who growth chart 0-4 years*. Royal College of Paediatrics; Child Health. https://www.rcpch.ac.uk/sites/default/files/Boys%5C_0-4%5C_years%5C_growth%5C_chart.pdf

Royal College of Physicians and Surgeons of Glasgow. (2023). Obesity Action Scotland | Providing leadership and advocacy on preventing & reducing obesity & overweight in Scotland | 5 years on, are we on track to halve childhood obesity in Scotland by 2030? Retrieved November 29, 2023, from https://www.obesityactionscotland.org/blogs/5-years-on-are-we-on-track-to-halve-childhood-obesity-in-scotland-by-2030

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511755453

Sabates, R., & Dex, S. (2012). Multiple risk factors in young children's development. https://cls.ucl.ac.uk/wp-content/uploads/2017/04/CLS-WP-2012-1.pdf

Sabates, R., & Dex, S. (2015). The impact of multiple risk factors on young children's cognitive and behavioural development. *Children & Society*, *29*(2), 95–108. https://doi.org/10.1111/chso.12024

Sarton, G. (1935). Preface to volume xxiii of isis (quetelet).

Scammon, R. E. (1927). The first seriatim study of human growth. *American Journal of Physical Anthropology*, *10*(3), 329–336. https://doi.org/10.1002/ajpa.1330100303

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sherwood, B., Maidman, A., & Li, S. (2023). *Rqpen: Penalized quantile regression* [R package version 3.1.3]. https://CRAN.R-project.org/package=rqPen

Shonkoff, J. P., Phillips, D. A., & National Research Council (U.S.) (Eds.). (2000). *From neurons to neighborhoods: The science of early child development*. National Academy Press.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231–245. Retrieved June 24, 2023, from https://www.jstor.org/stable/43304828

Smith, P. K., Cowie, H., & Blades, M. (2015). *Understanding children's development* (6th). Wiley.

Starfield, B., Shi, L., & Macinko, J. (2005). Contribution of primary care to health systems and health. *The Milbank Quarterly*, *83*(3), 457–502. https://doi.org/10.1111/j.1468-0009.2005.00409.x

Steinberg, L. D. (2011). *Adolescence* (9th ed). McGraw-Hil.

Steinberg, L. D., & Lerner, R. M. (2009). *Handbook of adolescent psychology* (Third ed). J. Wiley & sons.

Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture. *The Annals of Statistics*, *25*(4). https://doi.org/10.1214/aos/1031594728

Stuart, H. C., & Meredith, H. V. (1946). Use of body measurements in the school health program. *American Journal of Public Health and the Nations Health*, *36*(12), 1365–1386.

Stützle, W., Gasser, T., Molinari, L., Largo, R., Prader, A., & Huber, P. (1980). Shape-invariant modelling of human growth. *Annals of Human Biology*, *7*(6), 507–528. https://doi.org/10.1080/03014468000004641

Sybert, V. P., & McCauley, E. (2004). Turner's Syndrome. *New England Journal of Medicine*, *351*(12), 1227–1238. https://doi.org/10.1056/NEJMra030360

Szczesniak, R. D., Li, D., & Amin, R. S. (2016). Semiparametric mixed models for nested repeated measures applied to ambulatory blood pressure monitoring data. *Journal of Modern Applied Statistical Methods*, *15*(1), 255–275. https://doi.org/10.22237/jmasm/1462075980

Tanner, J. M. (1970). Growth and development at adolescence. In J. Kracht (Ed.), *Endokrinologie der Entwicklung und Reifung* (pp. 117–130). Springer. https://doi.org/10.1007/978-3-642-80591-2_14

Tanner, J. M., Whitehouse, R., & Takaishi, M. (1966). Standards from birth to maturity for height, weight, height velocity, and weight velocity: British children, 1965. i. *Archives of disease in childhood*, *41*(219), 454.

Tanner, J. M. (1962). Growth at adolescence.

Tanner, J. M. (1981). A history of the study of human growth. https://api.semanticscholar.org/CorpusID:142088877

Taylor-Miller, T., & Simm, P. J. (2017). Growth disorders in adolescents. *Australian Family Physician*, *46*(12), 913–917.

Thilakarathne, P. J., Clement, L., Lin, D., Shkedy, Z., Kasim, A., Talloen, W., Versele, M., & Verbeke, G. (2011). The use of semiparametric mixed models to analyze PamChip® peptide array data: An application to an oncology experiment. *Bioinformatics*, *27*(20), 2859–2865. https://doi.org/10.1093/bioinformatics/btr475

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Toeplitz, O. (1911). Zur theorie der quadratischen und bilinearen formen von unendlichvielen veränderlichen. *Mathematische Annalen*, *70*, 351–376. https://api.semanticscholar.org/CorpusID:116745398

Tsionas, E. G. (2003). Bayesian quantile inference. *Journal of Statistical Computation and Simulation*, *73*(9), 659–674. https://doi.org/10.1080/0094965031000064463

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368–384. https://doi.org/10.1037/a0032222

Ülker, E., & Arslan, A. (2009). Automatic knot adjustment using an artificial immune system for B-spline curve approximation. *Information Sciences*, *179*(10), 1483–1494. https://doi.org/10.1016/j.ins.2008.11.037

Uribe, J. M., & Guillen, M. (2020). Goodness of fit in quantile regression models. *Quantile Regression for Cross-Sectional and Time Series Data* (pp. 45–48). Springer International Publishing. https://doi.org/10.1007/978-3-030-44504-1_6

Van Buuren, S. (2023). Evaluation and prediction of individual growth trajectories. *Annals of Human Biology*, *50*(1), 247–257. https://doi.org/10.1080/03014460.2023.2190619

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to markov chain monte–carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154. https://doi.org/10.3758/s13423-016-1015-8

Villar, J., Giuliani, F., Bhutta, Z. A., Bertino, E., Ohuma, E. O., Ismail, L. C., Barros, F. C., Altman, D. G., Victora, C., Noble, J. A., Gravett, M. G., Purwar, M., Pang, R., Lambert, A., Papageorghiou, A. T., Ochieng, R., Jaffer, Y. A., & Kennedy, S. H. (2015). Postnatal growth standards for preterm infants: The Preterm Postnatal

Follow-up Study of the INTERGROWTH-21 st Project. *The Lancet Global Health*, *3*(11), e681–e691. https://doi.org/10.1016/S2214-109X(15)00163-1

Villar, J., Ismail, L. C., Victora, C. G., Ohuma, E. O., Bertino, E., Altman, D. G., Lambert, A., Papageorghiou, A. T., Carvalho, M., Jaffer, Y. A., Gravett, M. G., Purwar, M., Frederick, I. O., Noble, A. J., Pang, R., Barros, F. C., Chumlea, C., Bhutta, Z. A., & Kennedy, S. H. (2014). International standards for newborn weight, length, and head circumference by gestational age and sex: The Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. *The Lancet*, *384*(9946), 857–868. https://doi.org/10.1016/S0140-6736(14)60932-6

Villermé, L. R. (1835). *Tableau de l'état physique et moral des ouvriers employés dans les manufactures de coton, de laine et de soie* (tech. rep.) [Published as part of the work of the French Royal Academy of Medicine]. Imprimerie Royale. Paris, France.

Vinod, H. (1995). Double bootstrap for shrinkage estimators. *Journal of Econometrics*, *68*(2), 287–302. https://doi.org/10.1016/0304-4076(94)01639-H

Vosnaki, K., Bradshaw, P., Scholes, A., Scottish Centre for Social Research., & APS Group Scotland. (2019). *Life at age 12: Initial findings from the growing up in Scotland study* [OCLC: 1181146196]. The Scottish Government.

Votaw, D. F. (1948). Testing compound symmetry in a normal multivariate distribution. *The Annals of Mathematical Statistics*, *19*(4), 447–473. https://doi.org/10.1214/aoms/1177730145

Wadsworth, M., Kuh, D., Richards, M., & Hardy, R. (2006). Cohort profile: The 1946 national birth cohort(Mrc national survey of health and development). *International Journal of Epidemiology*, *35*(1), 49–54. https://doi.org/10.1093/ije/dyi201

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, *18*(2), 223–249. https://doi.org/10.1007/s001800300142

Wang, M. (2014). Generalized estimating equations in longitudinal data analysis: A review and recent developments. *Advances in Statistics*, *2014*, 1–11. https://doi.org/10.1155/2014/303728

Wang, Y. (2017). Potential mechanisms in childhood obesity: Causes and prevention. In I. Romieu, L. Dossus, & W. C. Willett (Eds.), *Energy Balance and Obesity*. International Agency for Research on Cancer. Retrieved August 5, 2024, from http://www.ncbi.nlm.nih.gov/books/NBK565802/

Wei, Y., & He, X. (2006). Conditional growth charts. *The Annals of Statistics*, *34*(5). https://doi.org/10.1214/009053606000000623

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, *25*(8), 1369–1382. https://doi.org/10.1002/sim.2271

Wei, Y., Richardson, T. G., Zhan, Y., & Carlsson, S. (2023). Childhood adiposity and novel subtypes of adult-onset diabetes: A Mendelian randomisation and genome-wide genetic correlation study. *Diabetologia*, *66*(6), 1052–1056. https://doi.org/10.1007/s00125-023-05883-x

Weiss, S. T., & Ware, J. H. (1996). Overview of issues in the longitudinal analysis of respiratory data. *American Journal of Respiratory and Critical Care Medicine*, *154*(6_pt_2), S208–S211. https://doi.org/10.1164/ajrccm/154.6_Pt_2.S208

WHO. (2021). Obesity and overweight. Retrieved November 29, 2023, from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

WHO. (2024). Malnutrition. Retrieved March 1, 2024, from https://www.who.int/news-room/fact-sheets/detail/malnutrition

WHO Multicentre Growth Reference Study Group, & de Onis, M. (2006). Who child growth standards based on length/height, weight and age: Who child growth standards. *Acta Paediatrica*, *95*, 76–85. https://doi.org/10.1111/j.1651-2227.2006.tb02378.x

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Chapman & Hall/CRC.

Wood, S. N. (2017a). *Generalized additive models: An introduction with R* (2nd). CRC Press/Taylor & Francis Group.

Wood, S. N. (2017b). P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. *Statistics and Computing*, *27*(4), 985–989. https://doi.org/10.1007/s11222-016-9666-x

World Health Organization. (n.d.). [Child] - Risk factors. Retrieved October 30, 2023, from https://www.who.int/data/gho/data/themes/topics/topic-details/mca/child---risk-factors

Wright, C. M., Williams, A. F., Elliman, D., Bedford, H., Birks, E., Butler, G., Sachs, M., Moy, R. J., & Cole, T. J. (2010). Using the new UK-WHO growth charts. *BMJ*, *340*(mar15 1), c1140–c1140. https://doi.org/10.1136/bmj.c1140

Wu, H., & Zhang, J.-T. (2006). *Nonparametric regression methods for longitudinal data analysis: Mixed-effects modeling approaches* [OCLC: ocm62525265]. Wiley-Interscience.

Xu, X., & Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, *10*(4). https://doi.org/10.1214/14-BA929

Yu, K., & Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, *54*(4), 437–447. https://doi.org/10.1016/S0167-7152(01)00124-9

Yu, K., & Zhang, J. (2005). A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics - Theory and Methods*, *34*(9-10), 1867–1879. https://doi.org/10.1080/03610920500199018

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Yue, Y. R., & Rue, H. (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis*, *55*(1), 84–96. https://doi.org/10.1016/j.csda.2010.05.006

Zeger, S. L., & Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, *50*(3), 689. https://doi.org/10.2307/2532783

Zhang, D., Lin, X., Raz, J., & Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, *93*(442), 710–719. https://doi.org/10.1080/01621459.1998.10473723

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., & Do, K.-A. (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *63*(4), 595–620. https://doi.org/10.1111/rssc.12053

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x