COMPARISON OF METHODS FOR GROWTH CHART CONSTRUCTION IN THE CANADIAN HEALTH MEASURES SURVEY

by

Bryan Maguire

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
June 2016

Dedicated to Sam and Max, without whom I could not have completed this project

Table of Contents

List of Tables

List of Figures

Abstract

Quantile and GAMLSS methods for growth curves are described and applied to data from the Canadian Health Measures Survey for child Body Mass Index (BMI) and triceps skinfold thickness. Both methods use cubic splines and GAIC is used to determine the extent of smoothness. Diagnostic worm plots were used to refine the models. A measure of smoothness of the final quantiles was developed and applied to the curves.

List of Abbreviations and Symbols Used

μ	Mean
ν	Skewness
σ	Variance
τ	Kurtosis
ARDC	Atlantic Research Data Centre
BMI	Body Mass Index
CHMS	Canadian Health Measures Survey
GAMLSS	Generalized Additive Models for Location Scale and Shape
QR	Quantile Regression

Acknowledgements

Thank you to Dr. Hamilton and Dr. Kuhle for all patience and insight, and to PERU, Dalhousie Statistics Department, and the ARDC for their support.

Chapter 1

Introduction

Monitoring the growth of a child is one of the key ways a healthcare provider can assess if that child is developing normally or if there is a cause for concern. What is considered the normal range of growth is highly dependent on that child's age. As a result, growth charts that are conditioned upon age are a useful tool that allow healthcare providers to quickly compare the development of a child to that of their peers. If a child is found to be at the extreme of the normal range for their age, intervention can begin immediately.

Age conditional growth charts are constructed from a cross-sectional representative sample of the target population (in rare cases there may be longitudinal data available). A fixed set of centile curves (in this thesis the 5th, 10th, 25th, 50th, 75th, 90th and 95th) show how we expect the distribution at each age to behave. If the lines are smooth, reading the chart is easier for the health care provider. Additionally smooth centile curves make much more sense biologically, because growth takes time and we would not expect any dramatic or abrupt changes.

There are many methods available for the construction of smooth centile curves. This work will focus on variations of the LMS method first proposed by Cole and Green (1992) and on quantile regression (Koenker and Bassett, 1978) for conditional centile curve estimation. Both methods have attractive attributes: The LMS method gives a parametric distribution allowing the calculation of an explicit conditional distribution at any given age, while the nonparametric quantile regression method requires no underlying assumption about the distribution of the data and allows for more flexibility in the shape of the fitted curves.

1.1 Canadian Health Measures Survey

The data used in this thesis comes from the Canadian Health Measures Survey (CHMS) cycles 1, 2 and 3, a cross-sectional survey of Canadians between 3 and

79 years. The CHMS consists of an at-home interview to collect demographic and lifestyle data as well as a follow-up examination consisting of anthropometric measurements, blood work, blood pressure, and other physical health tests at a mobile examination clinic. The clinic locations are chosen based on the Labor Force Survey, and households in those locations are chosen using data from the 2006 census.The collection of data happened at locations in seven provinces, Nova Scotia, Ontario, Quebec, British Columbia, Manitoba, Alberta and Newfoundland and Labrador.

Data collection occurred during 2007-2009 for cycle 1, 2009-2011 for cycle 2 and 2011-2013 for cycle 3. The cycles had a response rate of 51.7%, 55.7% and %52 respectively. Data from the two cycles was combined as per Statistics Canada guidelines (Statistics Canada, 2013) and weighted to account for the design effect and non-response bias (Statistics Canada, 2013). This thesis uses children that were between the age of 6 and 19 at the time of examination. Cycles 2 and 3 collected more information during the home interview portion of the study than cycle 1 but the measurements used for this work were done in the same way.

The anthropomorphic measurements captured by the CMHS included height, weight, body bass index (BMI), hip circumference, chest circumference, and thigh circumstance as well as five skinfold thickness measurements. Two of these measurements were used to compare and contrast the methods of centile curve creation: BMI, because of its common usage in identifying overweight and obese individuals, and triceps skinfold thickness, because the dramatic changes in the distribution of this variable as a function of age provide challenges during the fitting process. The population was divided by sex for analysis because males and females show very different growth curves for anthropomorphic measures. A total of 2965 males and 2868 females had their BMIs measured in cycles 1, 2, and 3. Skinfold thickness measurements were only available for cycle 1 and 2. A total of 1996 males and 1942 females had triceps skinfold thickness measurements in the target age range. Tables 1.1 and 1.2 show mean and interquartile range by age for BMI and triceps skinfold thickness.

Age	25th centile	50th centile	75th centile
6	14.6	15.6	16.3
7	15.2	16.2	18.4
8	15.8	16.5	18.6
9	15.9	17.3	19.4
10	16.3	17.4	20.7
11	16.6	18.4	20.6
12	17.0	18.6	21.7
13	18.2	20.9	22.8
14	18.6	20.9	24.5
15	19.3	21.8	26.3
16	19.5	21.3	25.0
17	20.4	23.0	25.3
18	21.6	24.0	28.0

(a) Male BMI

Age	25th centile	50th centile	75th centile
6	14.5	15.8	16.8
7	14.8	16.4	17.5
8	15.2	16.6	19.2
9	15.9	16.7	19.0
10	15.9	17.6	19.5
11	16.5	18.6	21.1
12	17.6	19.7	22.0
13	18.0	19.7	23.4
14	18.6	20.2	23.2
15	19.6	21.8	26.7
16	20.0	21.3	24.4
17	20.1	22.4	25.0
18	19.6	21.8	24.8

(b) Female BMI

Table 1.1: Male and female Body Mass Index by age

Age	25th centile	50th centile	75th centile
6	8.1	9.0	11.7
7	7.9	10.2	15.0
8	8.2	10.4	14.1
9	9.0	11.1	16.1
10	9.0	12.6	17.8
11	8.3	11.1	17.1
12	8.2	12.1	16.6
13	7.7	10.8	16.0
14	8.2	9.0	11.3
15	6.1	8.1	11.1
16	6.2	8.2	11.2
17	6.5	8.4	11.1
18	6.5	8.8	11.9

(a) Male Triceps SF

Age	25th centile	50th centile	75th centile
6	9.1	10.5	12.2
7	8.4	11.0	13.5
8	9.8	11.9	16.6
9	9.9	13.0	17.6
10	10.1	13.1	16.5
11	10.4	12.4	16.7
12	10.5	13.9	16.9
13	11.2	14.0	19.1
14	12.1	16.1	20.9
15	12.7	16.2	20.7
16	14.0	16.9	18.5
17	13.5	16.8	19.7
18	15.4	17.3	20.5

(b) Female Triceps SF

Table 1.2: Male and female triceps skinfold thickness by age

1.2 Growth Charts

The large range of information gathered during the CHMS allows researchers to explore the link between obesity and other health indicators such as cardiovascular disease, infectious diseases, and exposure to environmental contaminants. Additionally the survey data collected during interviews could reveal links between obesity and lifestyle choices such as nutrition, smoking habits, alcohol use, sexual behavior, and physical activity as well as demographic and socioeconomic variables. The distribution of anthropometric measures used to assess body composition vary dramatically as a function of age. This thesis aims to present models of anthropomorphic measurements that are conditional on a subject's age through the use of centile charts.

Childhood obesity is a risk factor for many diseases, including diabetes, heart disease and adverse psychological outcomes in childhood and adulthood (Reilly et al., 2003). BMI is often used as a measure of obesity because it is convenient and non-invasive to measure; however, it can not differentiate between muscle and fat and does not indicate where fat is stored on the body. Cardiovascular disease risk is an example of a health outcome associated with how fat is distributed around the body (Daniels et al., 1999). Skinfold measurements provide a clearer picture of how fat is stored on a child's body, giving healthcare providers a better understanding of the risk factors each child faces.

BMI is calculated using height and weight as $BMI = weight/height^2$ $[kg/m^2]$. Weight was measured using a digital scale and height using a stadiometer with movable head mount. Skinfold measurements were conducted according to the CPAFLA protocol (Canadian Society for Exercise Physiology, 2003) using a Harpenden skinfold caliper. Each skinfold was measured twice, the subscapular skinfold used for this thesis was measured below the inferior angle of the scapula at an angle of 45 degrees to the spine.

1.3 Thesis Overview

In this thesis, we compare and contrast various methods for the construction of growth charts for growth measurements from the CHMS data. Chapter 2 describes the method of quantile regression and fits models using an automatic model selection method. Chapter 3 details the LMS method and its generalization the GAMLSS method, then both model types are fitted to the CMHS data. In Chapter 4 the models produced in Chapter 2 and 3 are assessed and refined using the diagnostic tool of worm plots. Finally in Chapter 5 the refined models of each method are examined for overall smoothness.

Chapter 2

Quantile Regression

Parametric modeling methods for growth curves rely on the assumption that the data, after transformation, follows some known distribution conditional on age, and that its distribution is the same for all ages (or at least from the same family of distributions). Quantile regression is an attractive alternative because it requires no assumption about the underlying distribution of the data. It also does not impose any global constraints, meaning one age range can behave quite differently from another in the model.

2.1 Quantile Regression

Given data (t_i, y_i), $i = 1, \ldots, n$ quantile regression estimates a conditional quantile function $g(t_i)$ by minimizing an objective function given by Koenker and Bassett (1978)

$$\sum_{i=1}^{n} \rho_\tau \left(y_i - g(t_i)\right),$$

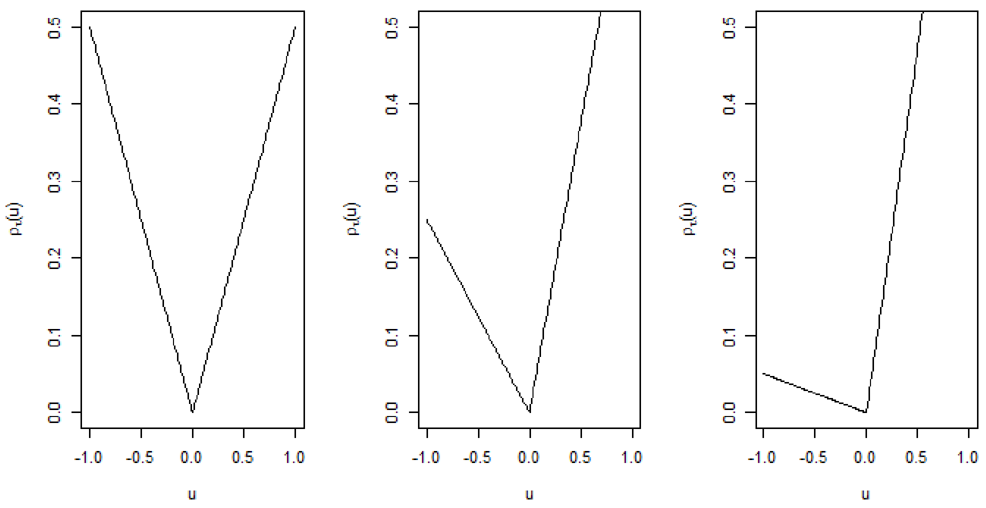
where τ is the desired quantile and g belongs to a family of smooth functions G. The function ρ_τ is a simple piecewise loss function

$$\rho_\tau(u) = u(\tau - I(u < 0)) = \begin{cases} \tau u & u \geq 0 \\ (\tau - 1)u & u < 0. \end{cases}$$

This loss function weights the errors above and below the fitted curve differently as seen in Figure 2.1.

If G is the set of constant functions, the estimated function $\hat{g}$ will be the τ^{th} sample quantile of Y. For example choosing $\tau = 0.5$ will result in $\hat{g}$ equal to the

Figure 2.1: Quantile regression loss function, $\tau = 0.5,\ 0.75,\ 0.95$



unconditional sample median which minimizes

$$\sum_{i=1}^{n} |y_i - g(t_i)|.$$

The growth of children, however, is known to be nonlinear so it is convenient to choose the family of functions G to be cubic splines. Cubic splines are flexible functions consisting of piecewise cubic polynomials over intervals given by a sequence of knots $\mathbf{u} = (u_0, \ldots, u_{N+1})^\top$. These polynomials have equal values, and equal first and second derivatives at the interior knots. Any cubic spline for a set of knots $\mathbf{u}$ can be written as a linear combination of basis splines, $h_j,\ j = 1, \ldots, J$ i.e

$$g(t_i) = \sum_{j=1}^{J} \beta_j h_j(t_i),$$

so conditional quantile estimation involves finding $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^\top$ which minimizes

$$\sum_{i=1}^{n} \rho_\tau \left(y_i - \sum_{j=1}^{J} \beta_j h_j(t_i) \right).$$

One advantage of quantile regression is the speed at which the models can be fitted.

Quantile regression can be formulated as a linear programming problem and solved using the simplex algorithm (Roberts and Barrodale, 1973). The error function ρ_τ is linear, so given p parameters $\boldsymbol{\beta}$ there are p possible directional derivatives which are positive or negative constants, and $2p$ possible directions. The algorithm shrinks the objective function by measuring the derivative of the objective function with respect to each parameter and changing the parameter corresponding to the largest negative derivative. This step is repeated until there are no negative directional derivatives remaining, indicating that a solution has been achieved (Koenker and Hallock, 2001).

2.2 B-splines

B-splines are a set of recursively defined basis splines with domain $[u_0, u_{N+1}]$. Let $u_0, \ldots, u_{N+1}$ be a set of nondecreasing numbers. Call u_j the j^{th} knot of the knot vector $\mathbf{u}$, with u_0 and u_{N+1} referred to as the end knots. All the B-splines developed from the set of knots $\mathbf{u}$ will have domain $[u_0, u_{N+1}]$. Knots u_0 and u_{N+1} are end knots and there are N interior knots. Any cubic splines with these knots can be written as a linear combination of B-splines of the degree three developed from this knot sequence.

The splines are defined recursively and depend on the choice of knot sequence $\mathbf{u}$ and the degree of the splines n ($n = 3$ in this case). For a given set of knots $\mathbf{u}$ and degree n, to define the set of basis function we first define a new vector of knots $\mathbf{v}$ with the end knots repeated m times where m is the order of the basis functions, $m = n + 1$ ($m = 4$ in this case)

$$\mathbf{v} = (v_0, \ldots, v_{N+2m})^\top = (u_{0_1}, \ldots, u_{0_m}, u_1, \ldots, u_{(N+1)_1}, \ldots, u_{(N+1)_m})^\top.$$

The set of B-splines of degree $n = 0$ is the set of "top-hat" functions spanning the space between each sequential knot.

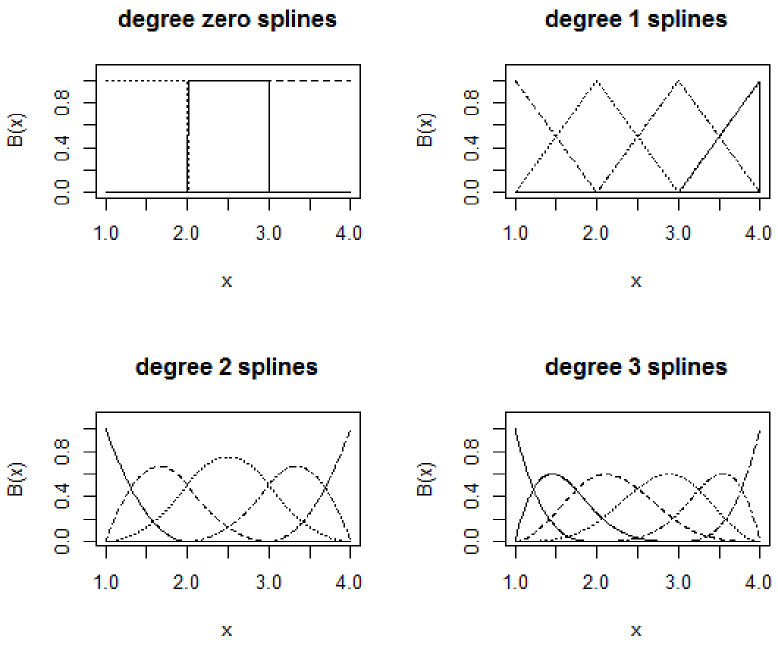
$$B_{j,0}(t) = \begin{cases} 1 & for\ v_j \leq t < v_{j+1} \\ 0 & otherwise \end{cases}$$

for $j = 1, \ldots, N+2m-1$. For $n > 0$ the B-splines are created by combining B-splines of lower order, using

$$B_{j,n}(t) = \frac{t - v_j}{v_{j+n} - v_j} B_{j,n-1}(t) + \frac{v_{j+n+1} - t}{v_{j+n+1} - v_{j+1}} B_{j+1,n-1}(t)$$

where $j = 0, \ldots, N+2m-1-n$. As a example, a spline function with a set of interior knots (2,3), end knots (1,4) and degree 3 can be constructed from $N+m = 2+4$ basis functions $B_{0,3}(t), \ldots, B_{5,3}(t)$. These basis functions and the lower order functions used to create them are shown in Figure 2.2. $B_{j,n}(t)$ is a weighted combination of two other basis functions so it is only non zero where they are non zero, thus $B_{j,n}(t)$ is non zero on at most the interval $[u_j, u_{j+n+1})$. This means that more densely placed knots result on B-splines that are non-zero on smaller intervals. Because the terms multiplying $B_{j,p-1}(t)$ and $B_{j+1,p-1}(t)$ are linear in t, $B_{j,n}(t)$ is indeed of order $n-1+1 = n$.

Figure 2.2: Basis splines for interior knots 2,3



We can use the set of functions $B_{j,n}(t)$ for $j = 1, \ldots, J = N+2m-1-n$ as a basis to approximate a function $f(t)$ by using a weighted sum of the basis functions. The first basis function is omitted $B_{0,n}(t)$ to avoid collinearity because $\sum_0^J B_{j,n}(t) = 1$. The function $g(t) = \sum_{j=1}^J \alpha_j B_{j,p}(t)$ is used to approximate the true function $f(t)$. This set of basis splines converts a covariate t into a flexible polynomial, that can be made more flexible by increasing the number of knots.

2.3 Model Construction

Quantile regression requires selection of a set of knots $\boldsymbol{u}$ from which the basis splines are constructed. Wei et al. (2004) suggests using the same set of knots for each centile and choosing a relatively even spacing of knots with additional knots during times of rapid change (younger ages). There is no universally accepted method of choosing the best set of knots in the literature.

In an effort to reduce the effect of the arbitrary choice of knots on the final model, the Generalized Akaike Information Criterion GAIC was used. The GAIC calculation normally requires the log-likelihood to be calculated. However, since quantile regression has no explicit distribution from which to calculate a likelihood, the minimized objective function is used instead, penalized by a factor proportional to the number of splines used J, and the number of parameters in those splines m (the order of the splines)(Koenker et al., 2016). For a single centile τ

$$GAIC_\tau = \sum_{i=1}^{n} \rho_\tau \left(Y_i - \sum_{j=1}^{J} \beta_j B_{j,p}(x_i) \right) + kJm.$$

The choice of k is left to the user, for BMI and triceps skinfold thickness the value $k = 3$ was used.

To created the set of candidate knot vectors we either included or excluded a knot at each integer value in the data range and enumerated all possible interior knot vectors for a total of 2^Ncandidate vectors. The end knots remain fixed and are included in each model. By calculating the GAIC of the model constructed from each possible knot vector and choosing the one with the lowest value, a set of knots that balances fidelity to the data and smoothness is found. This method is possible through brute force optimization. Since quantile regression models are so fast to fit, a model for each possible knot vector can be calculated and compared. For a set of centiles, the sum of the GAIC$_\tau$

$$GAIC = \sum GAIC_\tau$$

is used. This approach selects the same set of knots for every centile in the set.

2.4 Quantile Regression Results

To illustrate the effect of the number and placement of knots, growth charts are compared using various numbers of knots with an approximately even spacing across the domain of the data. For all models, fixed end knots at 6 and 19 are used. Internal knot sequences of (7,11,14), (7,9,14,16,17), and (7,9,11,13,14,15,16,17,18) are used to show how the graphs compare using a low, medium and high number of knots respectively. Each graph shows the 5th, 10th, 25th, 50th, 75th, 90th and 95th centiles. Figures 2.3, 2.4, and 2.5 show the B-splines constructed for each set of knots. A function that is the sum of the curves depicted Figure 2.5 will be very flexible but may overfit the noise in the data.

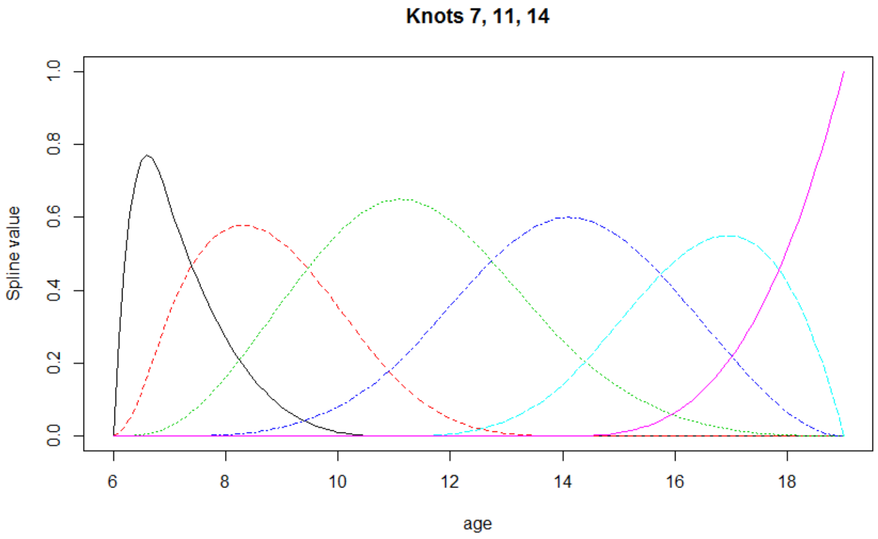


Figure 2.3: B splines for knots 7, 11, 14

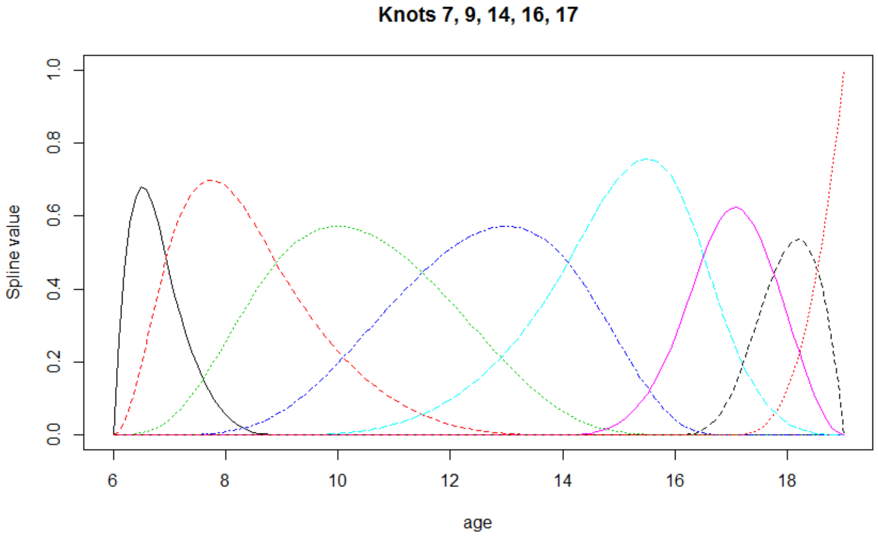


Figure 2.4: B splines for knots 7, 9, 14, 16, 17

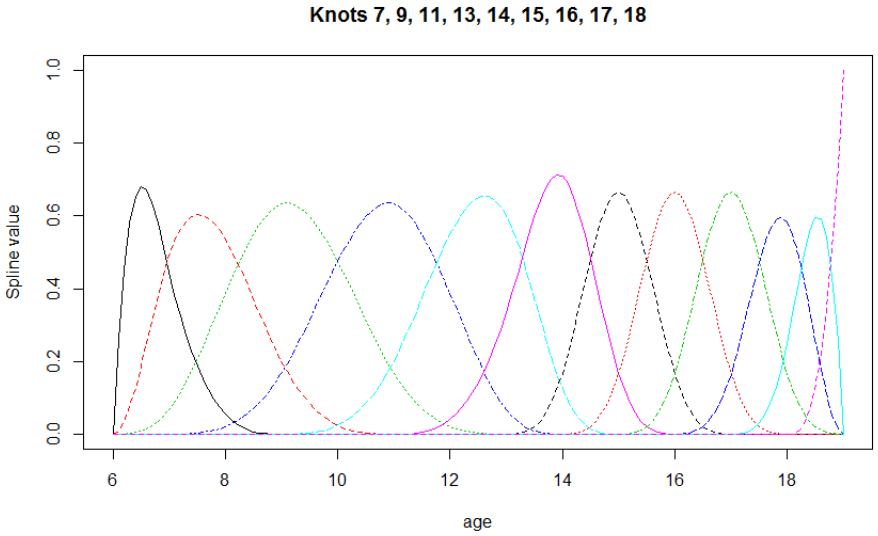


Figure 2.5: B splines for knots 7, 9, 11, 13, 14, 15, 16, 17, 18

The growth charts of BMI produced for males and females in Figures 2.6 and 2.7 respectively, show that a small and moderate number of knots produce smooth and plausible results. Using a high number of knots, however, causes rather extreme fluctuations to appear in the highest centiles. Interestingly, for males these fluctuations also appear in the lowest centiles as well. Figure 2.7 also illustrates another hazard of knot selection: The 95th percentile curve for female BMI constructed using many knots crosses the 90th percentile at the beginning of the range, indicating that there is too much flexibility in the curve there.

Constructing the growth curves of triceps skin fold thickness with the same set of knots yields results similar to BMI. Curves created with 3-5 knots have a well defined structure but curves constructed with more knots suffer from sporadic variations and the problem of crossing near the ends of the age range. Females (Figure 2.9) show more variation than males (Figure 2.8).

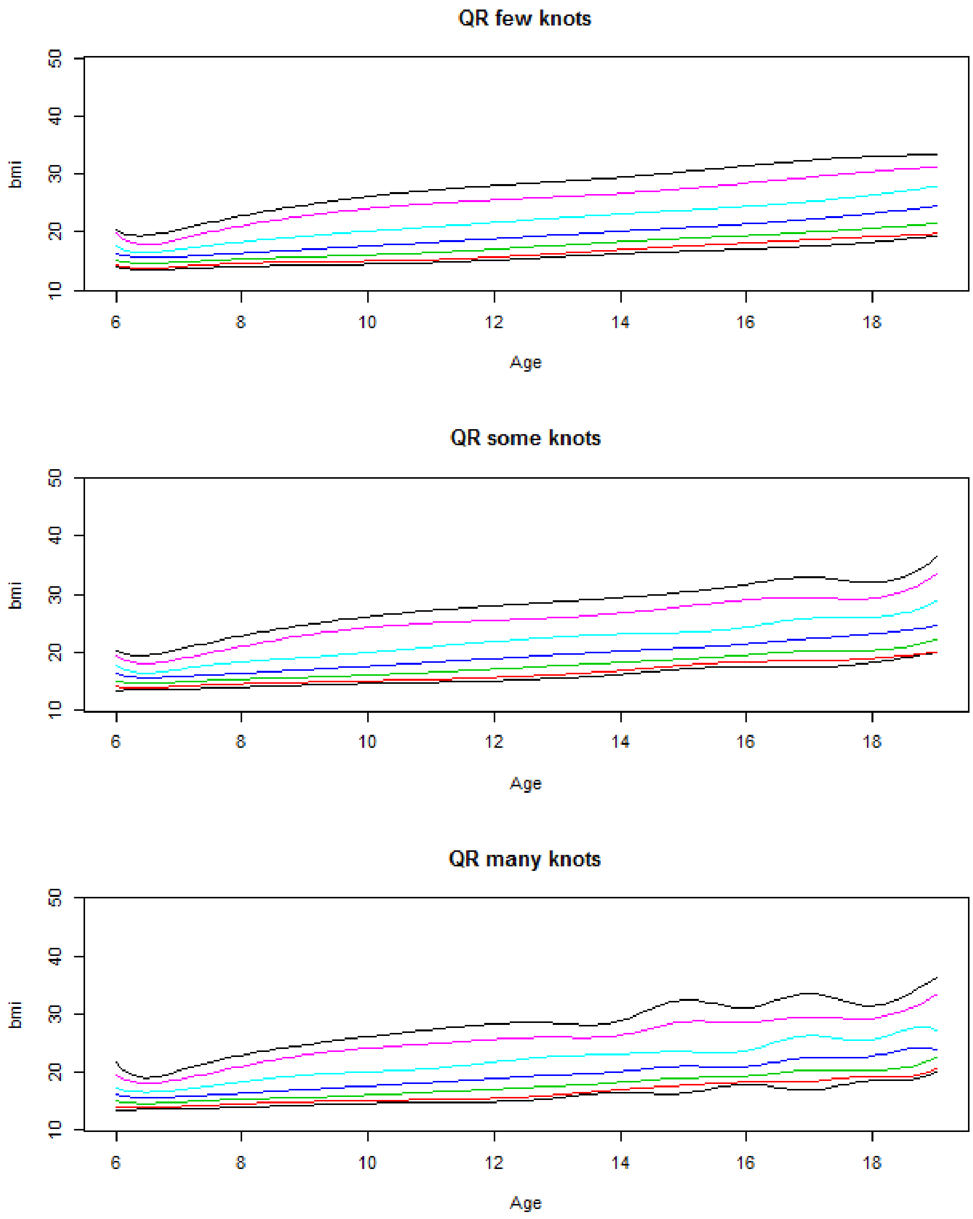


Figure 2.6: QR growth curves for male BMI using 3, 5, and 9 interior knots

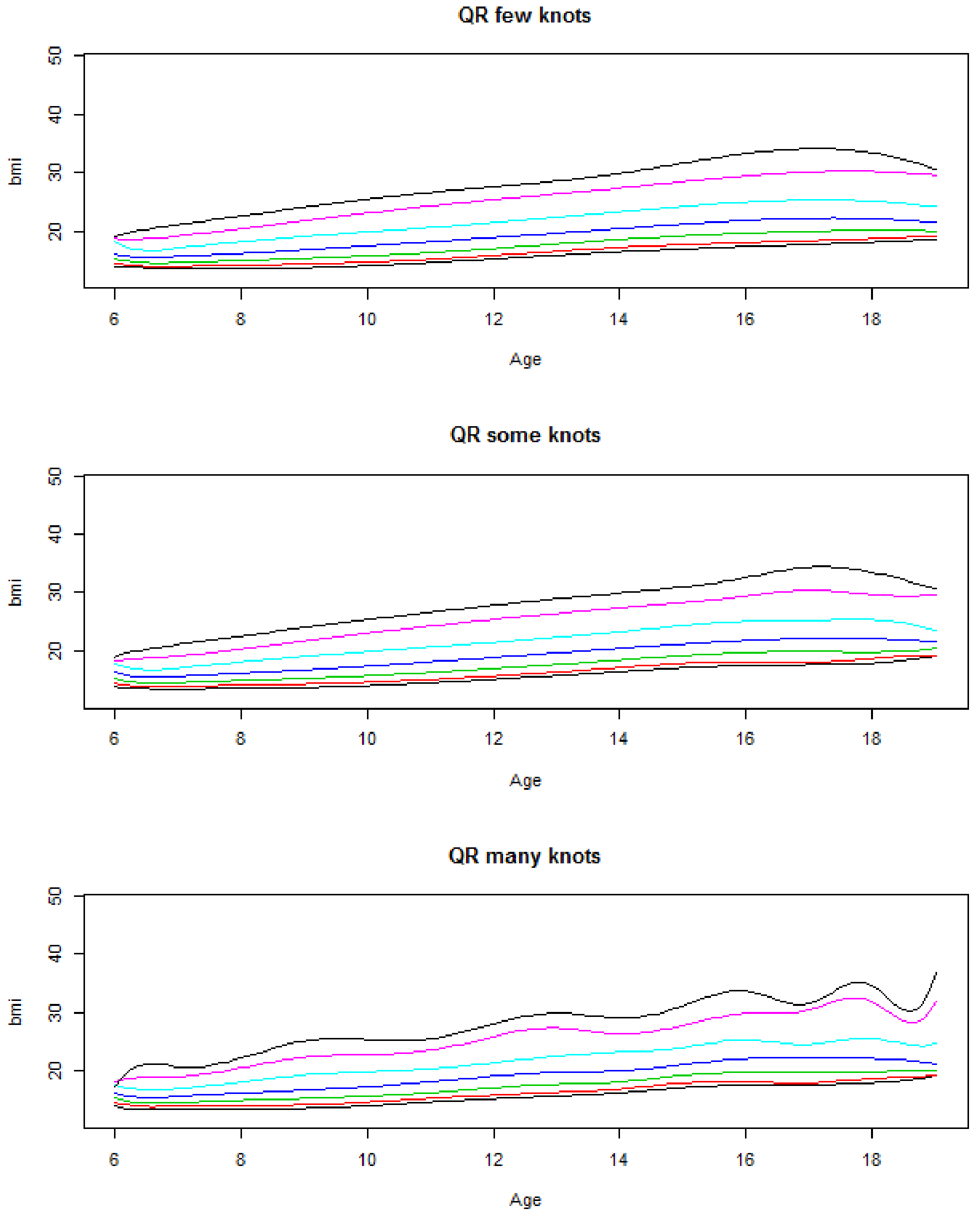
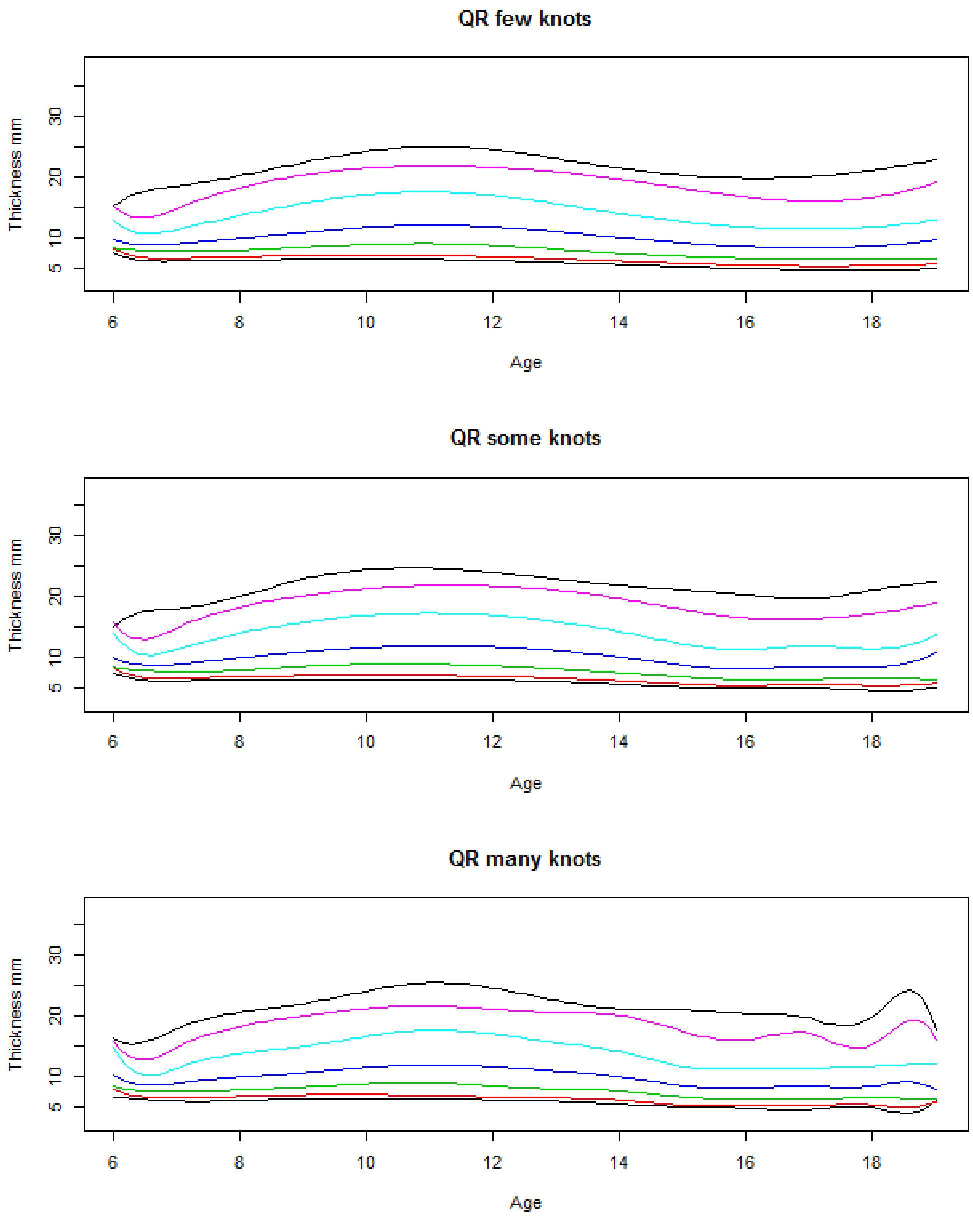


Figure 2.7: QR growth curves for female BMI using 3, 5, and 9 interior knots

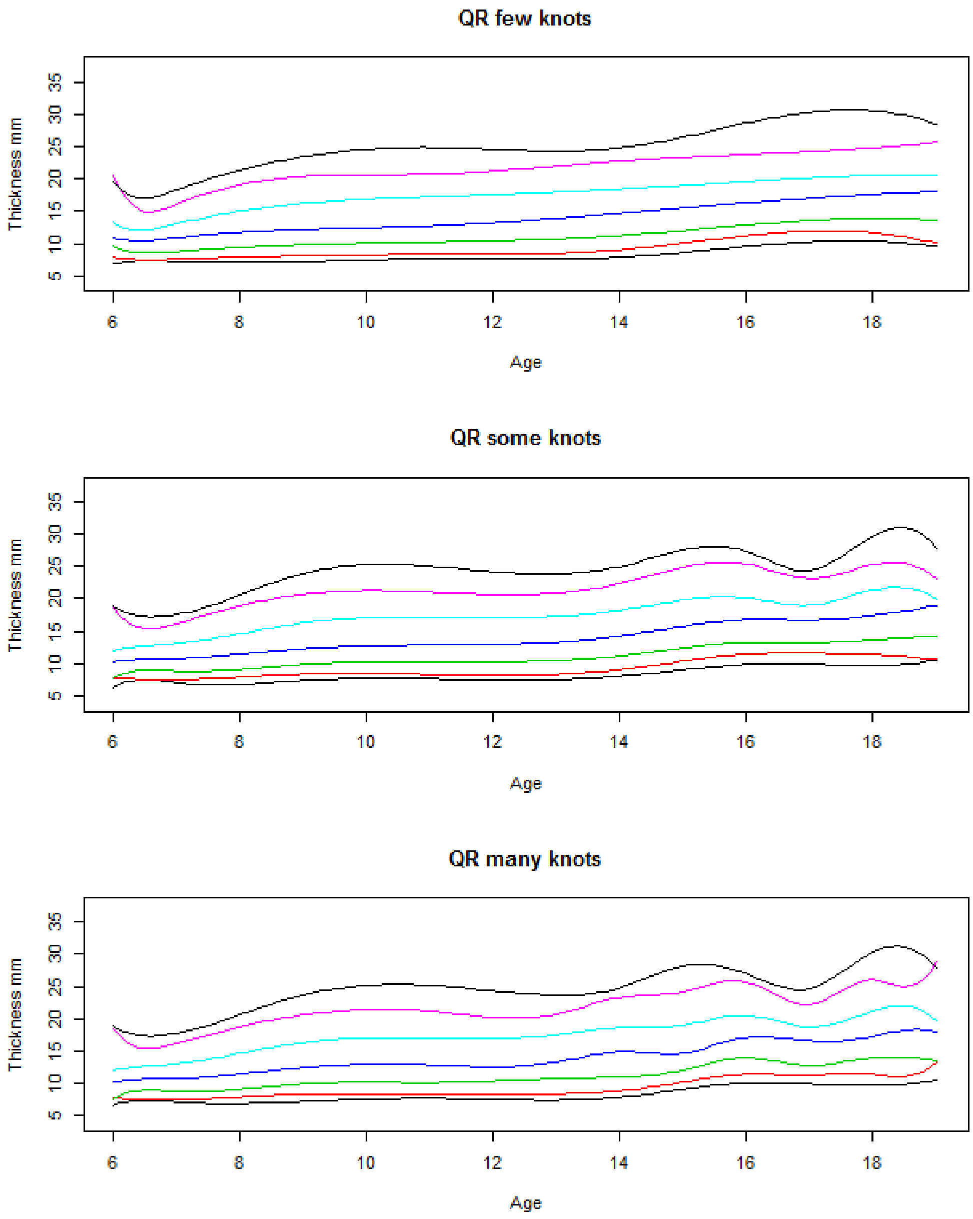


QR few knots

QR some knots

QR many knots

Figure 2.8: QR growth curves for male triceps using 3, 5, and 9 interior knots


QR few knots

QR some knots

QR many knots

Figure 2.9: QR growth curves for female triceps using 3, 5, and 9 interior knots

Using the GAIC method, the interior knot vector selected for each variable and sex is listed in Table 2.1. The models created for male BMI and female triceps skinfold both use 3 interior knots along with end knots at 6 and 19. Both feature smooth curves, with no undesirable rapid changes (Figures 2.10 and 2.13). Female BMI and male triceps skinfold have 7 and 6 interior knots, respectively. They show much more rapid variation in their growth curves (Figures 2.11 and 2.12). They both also suffer from crossing centiles at the endpoints of the graphs. Male triceps skinfold in particular suffers from the problem of having very wiggly centiles near the end points. This is likely the result of the knots that were selected being clustered around the maximum and minimum ages (7 and 8 on the low end, 16, 17 and 18 on the high end).

Measure	Gender	Knots selected
BMI	Male	10 11 17
BMI	Female	10 11 13 14 16 17 18
TRIC	Male	7 8 10 16 17 18
TRIC	Female	9 14 15

Table 2.1: Knots selected through GAIC for quantile regression

These models selected by automatic model selection will be evaluated and improved if needed in Chapter 4.

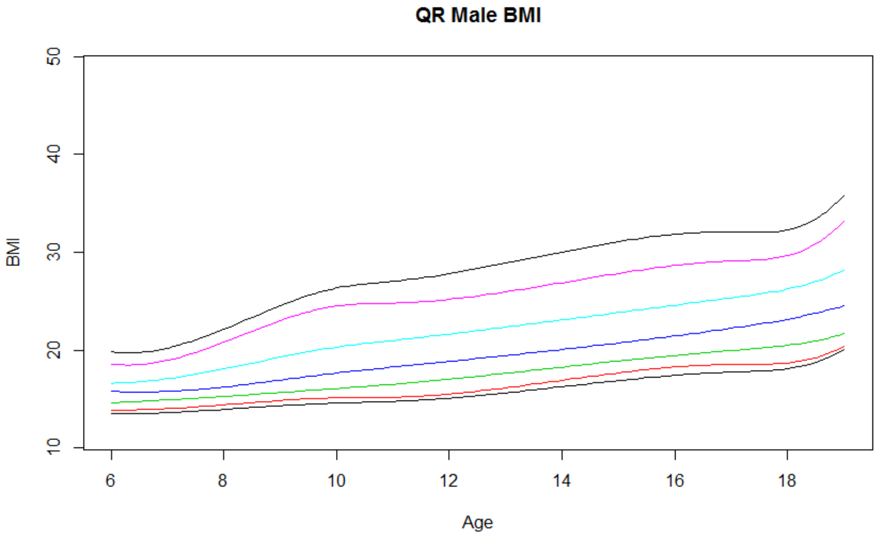


Figure 2.10: QR growth curves for male BMI with knots at 10, 11, 17 years

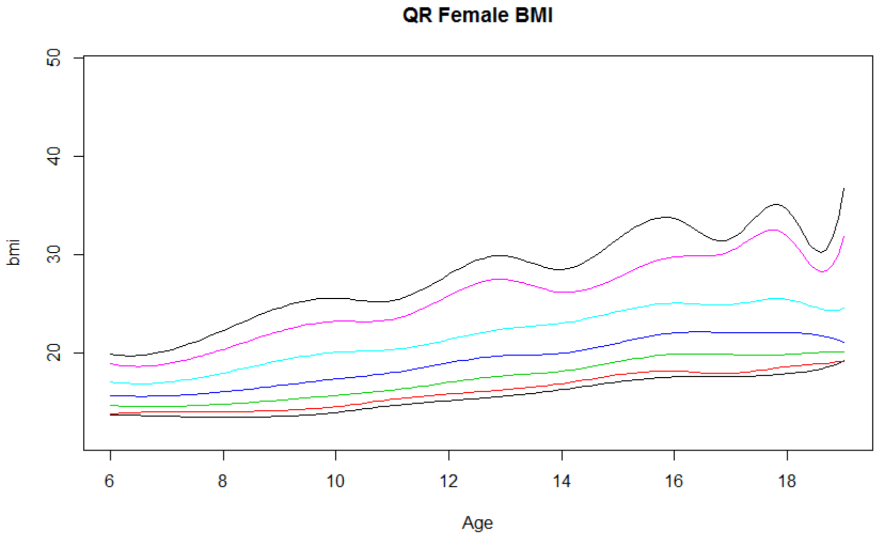


Figure 2.11: QR growth curves for female BMI with knots at 10, 11, 13, 14, 16, 17, 18 years

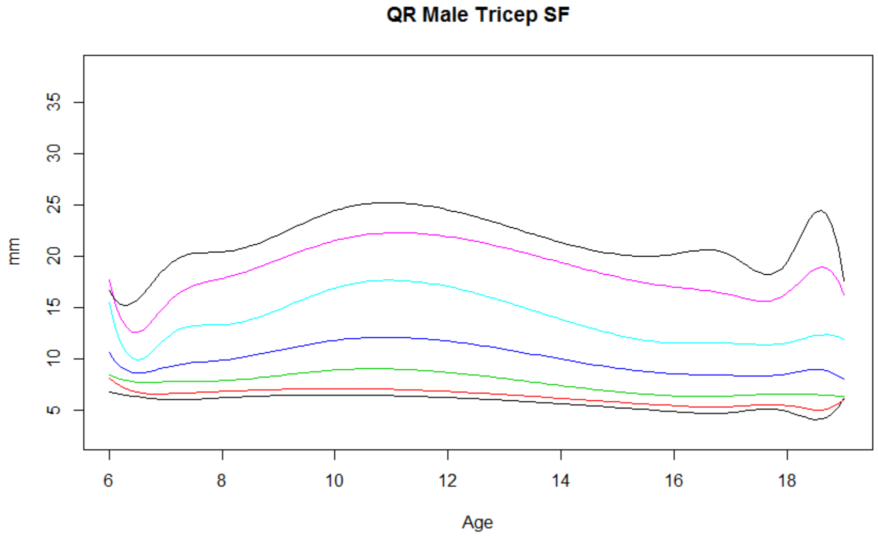


Figure 2.12: QR growth curves for male triceps skinfold thickness with knots at 7, 8, 10, 16, 17, 18 years

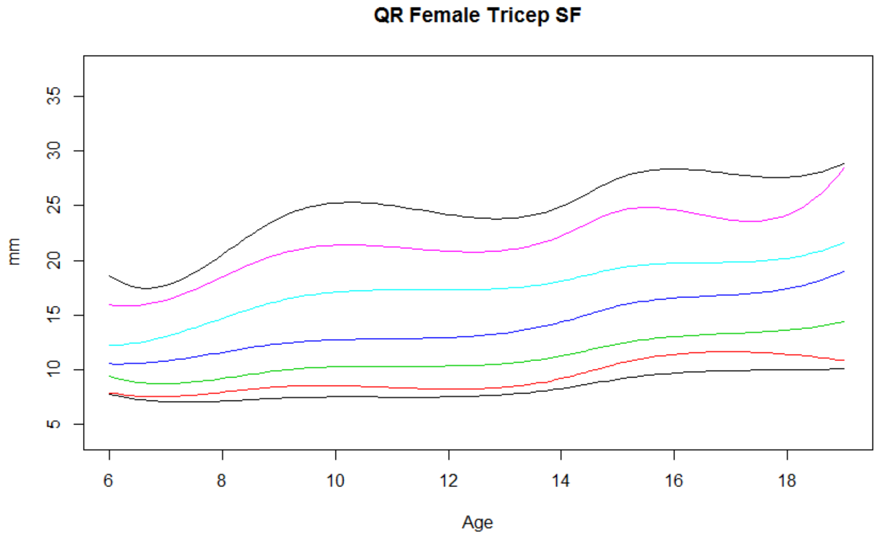


Figure 2.13: QR growth curves for female triceps skinfold thickness with knots at 9, 14, 15 years

Chapter 3

Generalized Additive Models for Location Scale and Shape

In contrast to the nonparametric method of quantile regression, parametric methods offer the advantage of creating a full conditional likelihood function that allows for the creation of centiles at any desired value as well as producing a likelihood of any observed value. The LMS and GAMLSS methods outlined in this section transform the data to follow a known distribution and allow the parameters of the distribution to change smoothly with covariates.

3.1 Generalized Additive Models For Location Scale and Shape

The widely used general linear models and general additive models describe a response y as a random variable with a distribution in the exponential family. The mean of y is modeled as function of explanatory variables using a monotonic link function. The wide range of distributions included in the exponential family give these models a good deal of flexibility. However while these models allow for the mean of the distribution to depend on various covariates, the variance of y is a function of the mean and a dispersion parameter through the variance function, $V(y) = \phi v(\mu)$.

If the variance of y changes as a function of the covariates in a different way than implied by the variance function then our model may not be accurate. Additionally, for members of the exponential family the skewness and kurtosis of y are generally functions of of the mean μ as well.

The GAMLSS method developed by Rigby and Stasinopoulos (Rigby and Stasinopoulos, 2005) generalizes these methods by relaxing the requirement that the distribution be a member of the exponential family and by allowing any of the parameters of the distribution to be a function of covariates. This allows the different moments to have different relationships with the covariates than the usual dependence on the mean of the distribution.

In general it is assumed each observation y_i is independent of the other and has probability density function $f(y_i, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ is a vector of parameters. Each parameter is modeled using its own monotonic link function and linear predictor

$$g_k(\theta_k) = \eta_j.$$

For most distributions the first parameter θ_1 is the mean, or location parameter μ, and θ_2 is the variance parameter σ. Some distributions have one or two shape parameters customarily denoted $\theta_3 = \nu$ and $\theta_4 = \tau$. Using this parameterization and a distribution with four parameters $f(y|\mu, \sigma, \nu, \tau)$, the GAMLSS model has the form

$$\begin{aligned} g_1(\mu) &= \eta_1 \\ g_2(\sigma) &= \eta_2 \\ g_3(\nu) &= \eta_3 \\ g_4(\tau) &= \eta_4. \end{aligned} \tag{3.1}$$

These models have been shown to be very flexible and have been applied to a wide range of topics. Rigby and Stasinopoulos (Rigby and Stasinopoulos, 2004) showed that the GAMLSS model could be used to to create centile curves that vary as a function of an explanatory variable (age).

3.1.1 Parametric Centile Curve Estimation with GAMLSS

To model the distribution of y as a smooth function of a single explanatory variable x (age) we can use the GAMLSS model. Given $X = x, Y$ is modeled as a random variable with a density function $f_Y(y|\mu, \sigma, \nu, \tau)$ where the parameters μ, σ, ν, τ are modeled as

$$\begin{aligned} g_1(\mu) &= h_1(x) \\ g_2(\sigma) &= h_2(x) \\ g_3(\nu) &= h_3(x) \\ g_4(\tau) &= h_4(x) \end{aligned} \tag{3.2}$$

where h_i are cubic smoothing splines in x.

The choice of monotonic link function g_k, for $k = 1, \ldots, 4$, can be changed to suit the needs of the problem under consideration, but g_1 is conventionally chosen to be the identity link so that it will have an additive effect and ease in interpretability. For the second link function, g_2 is in general chosen to be the log link to ensure that the scale parameter is always non-negative. Finally g_3 and g_4 are also conventionally chosen to be the identity and log link respectively for similar reasons but these can vary based on the requirements of the density function f chosen for the model.

The functions $h_k, \; k = 1, \ldots, 4$, are estimated by maximizing the penalized log likelihood defined as

$$l_p(\mu, \sigma, \nu, \tau) = l_d(\mu, \sigma, \nu, \tau) - \frac{1}{2}\sum_{k=1}^{4} \lambda_k \int (h_k^{''}(x))^2 dx \tag{3.3}$$

where l_d is the log-likelihood with $l_d = \sum_{i=1}^{n} l_i$ due to independence, where l_i is the log-likelihood of a single observation from the distribution $f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$. The constants λ_k are chosen by the user and determine the amount of penalty that is applied to each parameter. A large λ_k results in a smoother curve $h_k(x)$ being fitted at the expense of fidelity to the data. Rigby and Stasinopoulos (2005) show that l_p depends only on the heights of the spline at the observed ages and that the penalty can be written as a quadratic form in these heights with the matrix dependent on the second derivatives of the splines. Appendix A contains a description of the smoothing splines and of the penalty matrix.

3.1.2 LMS Method

Before the GAMLSS framework was developed by Rigby and Stasinopoulos, the first use of a model from this family for centile curves was in a 1992 paper by Cole and Green (Cole and Green, 1992). Most growth data exhibits some degree of skewness so the method assumed that the data could be normalized using a Box-Cox power transformation. For the positive random variable Y the transformed variable Z is defined as

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right] & if\ \nu \neq 0 \\ \frac{1}{\sigma} log\left(\frac{Y}{\mu}\right) & if\ \nu = 0 \end{cases} \tag{3.4}$$

where $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$ and Z is standard normal. When $\nu = 1$, this gives a normal distribution, $\nu < 1$ gives a skewed right distribution and $\nu > 1$ a skewed left distribution. If we allow the values of the parameters to change as a smooth function of time t, we can replace ν, μ, and σ with the smooth curves $L(t)$, $M(t)$, and $S(t)$ respectively giving the form

$$Z = \begin{cases} \frac{1}{S(t)L(t)} \left[\left(\frac{Y}{M(t)} \right)^{L(t)} - 1 \right] & if\ L(t) \neq 0 \\ \frac{1}{S(t)} log \left(\frac{Y}{M(t)} \right) & if\ L(t) = 0. \end{cases} \tag{3.5}$$

The loglikelihood of this distribution is given by

$$l(L, M, S) = \sum_{i=1}^{n} \left(L(t_i) log \frac{y_i}{M(t_i)} - log(S(t_i)) - \frac{z_i^2}{2} \right), \tag{3.6}$$

where z_i is the standardized score of observation y_i found in equation (3.5). To estimate the three curves of the LMS method the penalized log-likelihood

$$l(L, M, S) - \frac{\lambda_\mu}{2} \int (M''(t))^2 dt - \frac{\lambda_\sigma}{2} \int (S''(t))^2 dt - \frac{\lambda_\nu}{2} \int (L''(t))^2 dt \tag{3.7}$$

is maximized, which has the same form as equation (3.3) showing that the LMS method is a special case of the more general GAMLSS method. Cole and Green (1992) provide an algorithm for maximizing this penalized likelihood. Once the model has been fitted we can explicitly calculate any centile $100(1-\alpha)$ by rearranging equation (3.5) as

$$\begin{cases} C_{100(1-\alpha)} = M(t) \left(1 + L(t)S(t)z_\alpha\right)^{1/L(t)} & if\ L(t) \neq 0 \\ C_{100(1-\alpha)} = M(t) exp(S(t)z_\alpha) & if\ L(t) = 0 \end{cases} \tag{3.8}$$

where z_α is the upper α quantile of a standard normal distribution. Since the lines $L(t)$, $M(t)$, and $S(t)$ are smooth functions, $C_{100(1-\alpha)}$ will be a smooth function (Cole and Green, 1992).

3.1.3 Box-Cox Power Exponential

In an effort to deal with the presence of kurtosis seen in some growth data (Van Buuren and Fredriks, 2001) which the LMS method had no mechanism for modeling, Rigby and Stasinopoulos described a generalization of the LMS method (Rigby and Stasinopoulos, 2004) called the Box-Cox Power Exponential method (BCPE). The BCPE method works similarly to the LMS method with the exception that the transformed variable Z is now assumed to follow a standard power exponential distribution with power parameter τ, rather than a standard normal distribution. As with the other parameters, τ is modeled as a smooth continuous function of time. The probability density function of a power exponential distribution is

$$f_Z(z) = \frac{\tau}{c2^{(1+\tau^{-1})}\Gamma(\tau^{-1})} exp\left(-\frac{1}{2}\left|\frac{z}{c}\right|^{\tau}\right) \tag{3.9}$$

where $\tau > 0$ and $c^2 = 2^{-2/\tau}\Gamma\left(\tau^{-1}\right)\left(\Gamma(3/\tau)\right)^{-1}$. If τ is $\tau = 2$ then $f_Z(z)$ is the standard normal distribution, showing that the LMS method is indeed a special case of the BCPE method. Other distributions are also special cases of the BCPE parameterization, with $\tau = 1$ corresponding to a double exponential distribution, and the limiting case $\tau \to \infty$ corresponding to the uniform distribution (Rigby and Stasinopoulos, 2004).

To compute the centiles of a BCPE function a formula similar to the LMS method is used

$$\begin{cases} C_{100\alpha} = M(t)\left(1 + L(t)S(t)z_\alpha\right)^{1/L(t)} & if\ L(t) \neq 0 \\ C_{100\alpha} = M(t)exp(S(t)z_\alpha) & if\ L(t) = 0, \end{cases} \tag{3.10}$$

however now z_α is the $100(1-\alpha)$ centile of a power exponential distribution with power parameter $K(t)$, given by

$$z_\alpha = \begin{cases} -c\left[2F_s^{-1}\left(1-2\alpha\right)\right]^{1/K(t)} & if\ L(t) \neq 0 \\ c\left[2F_s^{-1}\left(1-2\alpha\right)\right]^{1/K(t)} & if\ L(t) = 0 \end{cases} \tag{3.11}$$

where F_S^{-1} is the inverse cumulative probability distribution of a gamma distribution with shape parameter equal to the inverse of $K(t)$. Rigby and Stasinopoulos (2004) showed that this distribution offered a significant improvement over the LMS method in a study of the growth of Dutch boys. This improvement was due to the presence of leptokurtosis (fatter tails) at young ages that went unmodeled using the LMS method.

3.1.4 Model Selection

As mentioned above, the smoothing penalty corresponding to the cubic spline for each parameter can be expressed in a quadratic form. The trace of the corresponding matrix is called the effective degrees of freedom for variable k (edf_k), and is inversely related to λ_k. Models are specified by the *edf* of each parameter, for example $BCPE(5,4,3,3)$ represents a Box-Cox Power Exponential distribution with 5 effective degrees of freedom on the location parameter μ, 4 effective degrees of freedom on the scale parameter σ, etc. The total effective degrees of freedom for a given model is simply the sum of the parameter *edf*'s.

In order to choose how much smoothing to apply to each parameter (what each edf_k should be) the log-likelihood penalized by the total effective degrees of freedom is maximized using the Generalized Akaike Information Criterion (GAIC). The GAIC for a given model is

$$GAIC(b) = -2l(\mu, \sigma, \nu, \tau) + b \bullet edf \tag{3.12}$$

with $b > 0$. A choice of $b = 2$ is equivalent to using AIC, and $b = log(n)$ with n equal to the number of observations is equivalent to the Bayesian Information Criterion (BIC).

To find the optimal choice of smoothing parameters for a chosen model and information criterion penalty we start with a base model with $edf_k = 1$ for all k, ie: $BCPE(1,1,1,1)$, and optimize the GAIC of the model over the parameter space of possible smoothing parameter combinations. The R function **gamlss** from the GAMLSS package (Stasinopoulos et al., 2015) is used to fit the models for a given set of smoothing parameters and to calculate the GAIC. In conjunction with this the R function **optim** from the base STATS package (R Core Team, 2016) is used to determine the best values for the smoothing coefficients.

3.2 LMS and GAMLSS results

The LMS and BCPE methods described in the previous section were applied to the CHMS data to construct growth charts for BMI and triceps skinfold thickness. To illustrate the effect of the choice of smoothing parameters on these results, the graphs were created by applying a large amount of smoothing and a small amount of smoothing with the analysis stratified by sex. The constructed charts for both anthropometric measures using automatic model selection using GAIC, first with the LMS method, and then with the BCPE model are shown.

Figure 3.1 shows male BMI fitted using the LMS method with 1 effective degrees of freedom for each of the 3 parameters. This model is notated as lms(1,1,1). The centile curves are almost straight lines and increase in the spacing of the centiles until approximately age 13 when they become relatively parallel. This behavior is reflected in Figure 3.2 which shows plots of the three smoothed parameters L, M, and S as a function of age. The conditional mean μ is nearly linear while the variance increases at a relatively constant rate until about 13 years of age when it becomes constant. The power parameter ν also increases rapidly until age 13 when its increase becomes more gradual. An increasing power parameter indicates that the distribution of BMI in the sample becomes more similar to a normal distribution as age increases.

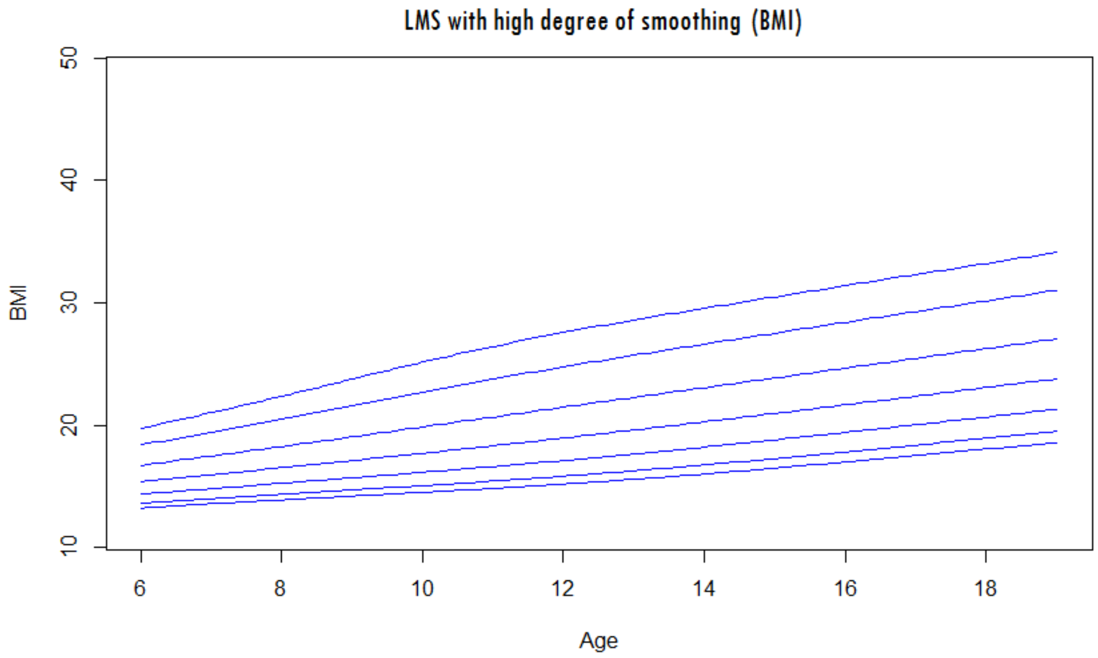


Figure 3.1: LMS(1,1,1) growth curves for male BMI

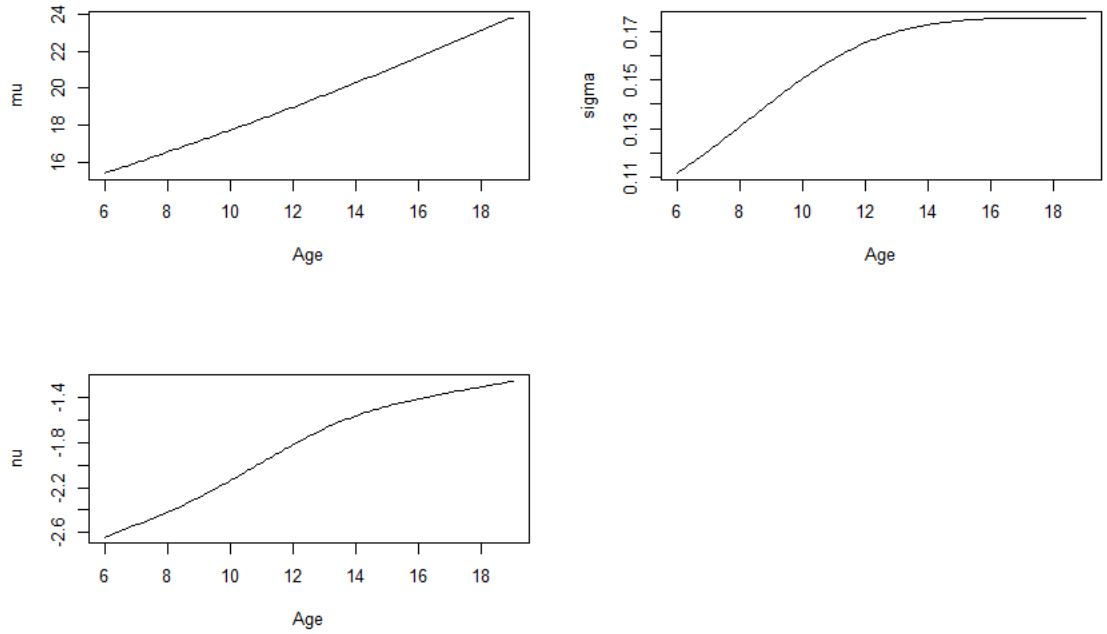


Figure 3.2: LMS(1,1,1) parameter curves for male BMI

Female BMI follows a similar pattern to male BMI (Figure 3.3) with centiles which are nearly straight lines that diverge slightly until about age 13 when the spacing becomes more constant. Figure 3.4 shows this reflected in the smoothed age conditional parameters, L, M, and S with nearly linearly increasing mean and increasing standard deviation until age 13. The notable contrast with male BMI is that the power parameter ν begins to decrease at higher ages after peaking in the 12-14 years of age range. This indicates that the distribution of BMI is more skewed for older females.

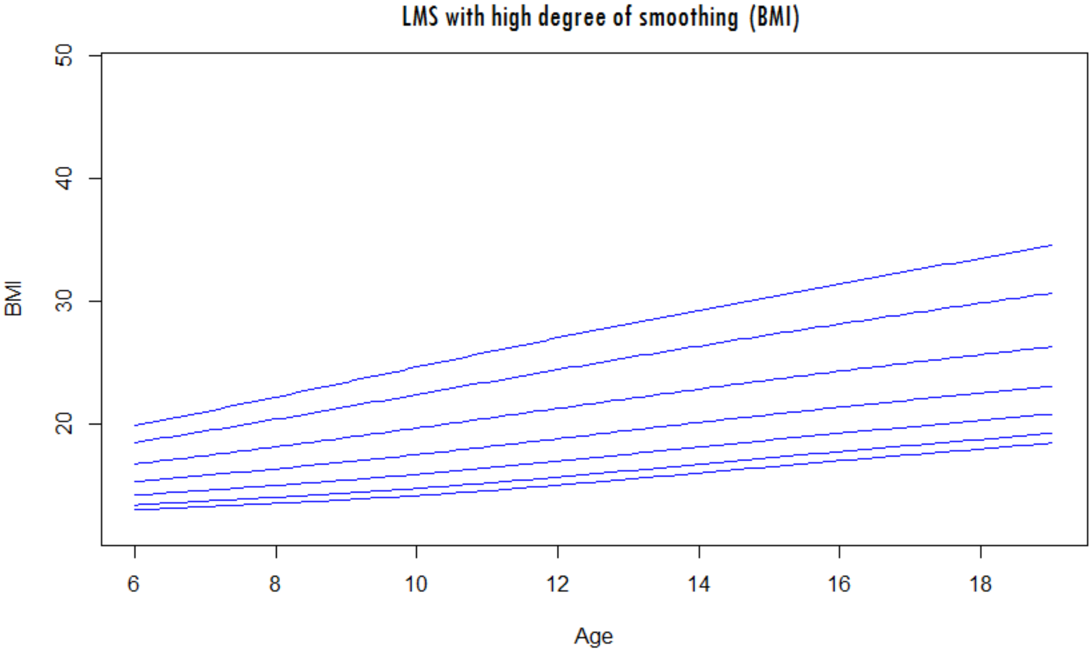


Figure 3.3: LMS(1,1,1) growth curves for female BMI

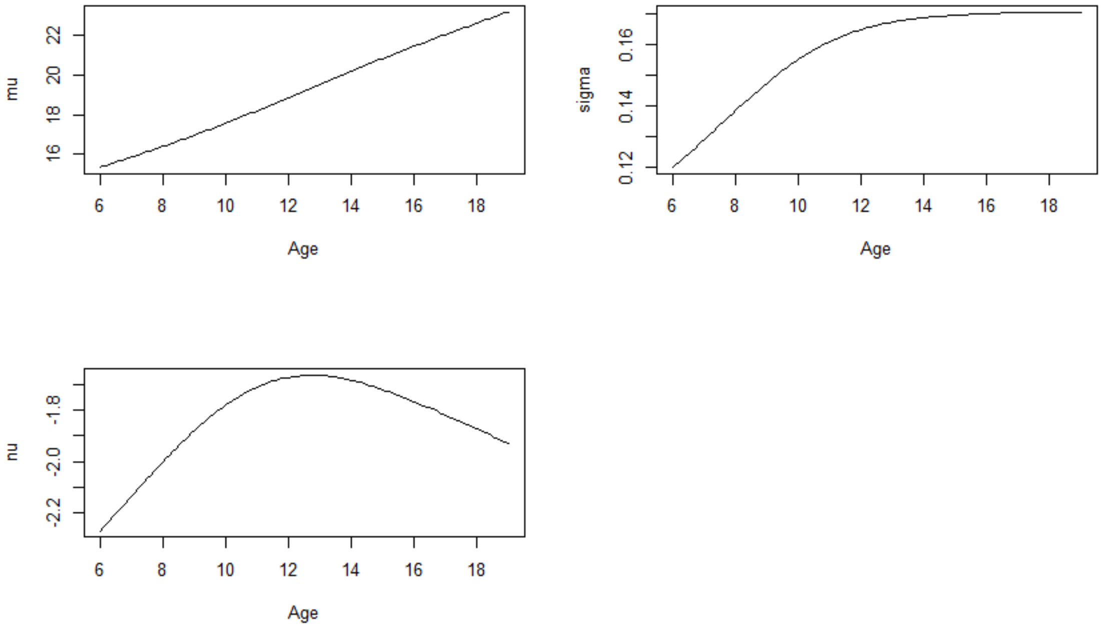


Figure 3.4: LMS(1,1,1) parameter curves for female BMI

When an edf of 12 is used for each parameter (model lms(12,12,12)) much less smoothing is applied and the centile curves are much more responsive to minor fluctuations in the data. Figures 3.5 and 3.6 show the fitted male and female BMI centiles using this small amount of smoothing. The differences compared to the heavily smoothed centile are immediately obvious, most notably in the highest and lowest centiles.

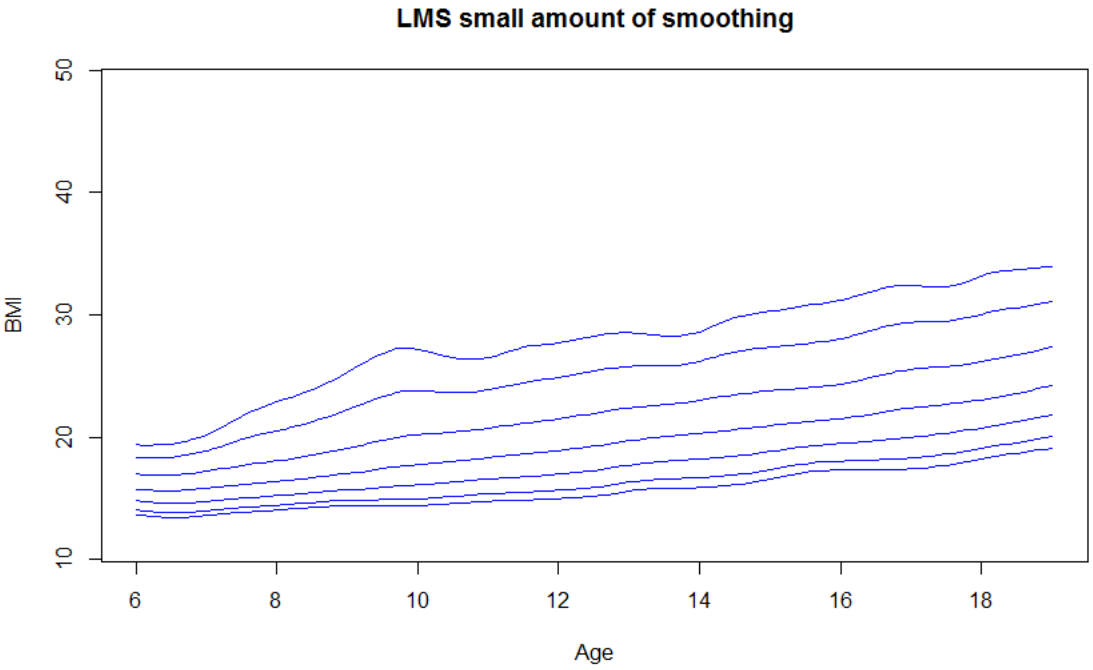


Figure 3.5: LMS(12,12,12) growth curves for male BMI

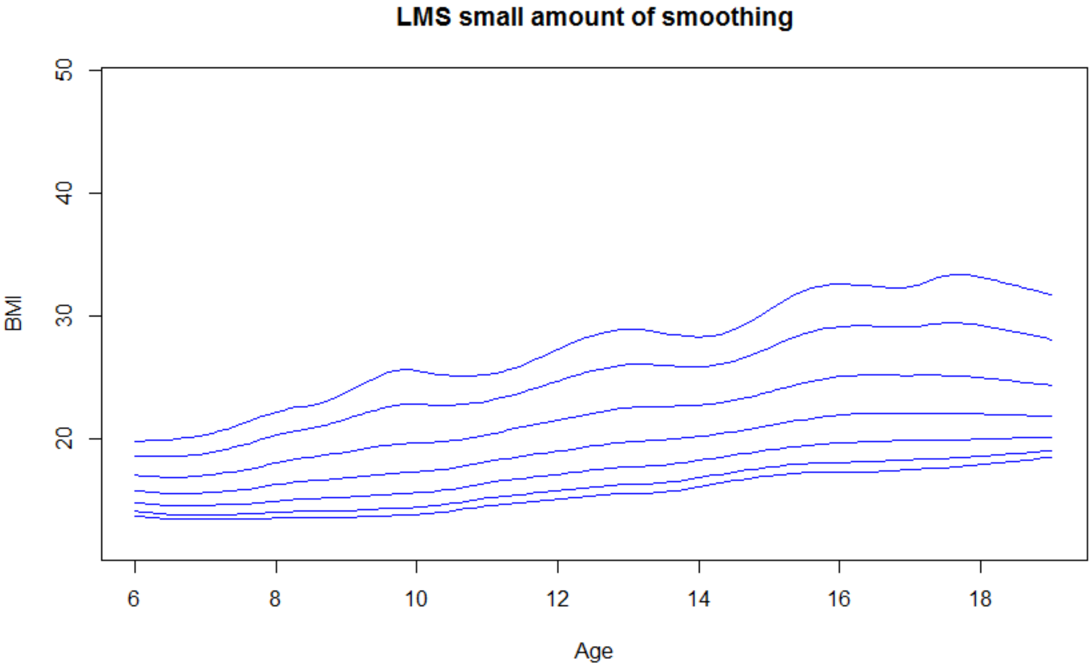


Figure 3.6: LMS(12,12,12) growth curves for female BMI

The small fluctuations that appear in the 95th centile of the male BMI curves are unlikely to be the result of an underling biological process but rather of the curves being fitted to the noise of the data. However some new features are visible in this graph that possibly do represent an underlying biological process. Rather than increasing steadily from age 6 to age 13 the centiles in Figure 3.5 show that BMI is constant until 7 years of age when it begins a more rapid change than depicted in the heavily smoothed centiles. The increasing variance and diverging centile lines are only observed until approximately age 10 compared to the higher degree of smoothing where they exhibited this behaviour until age 13.

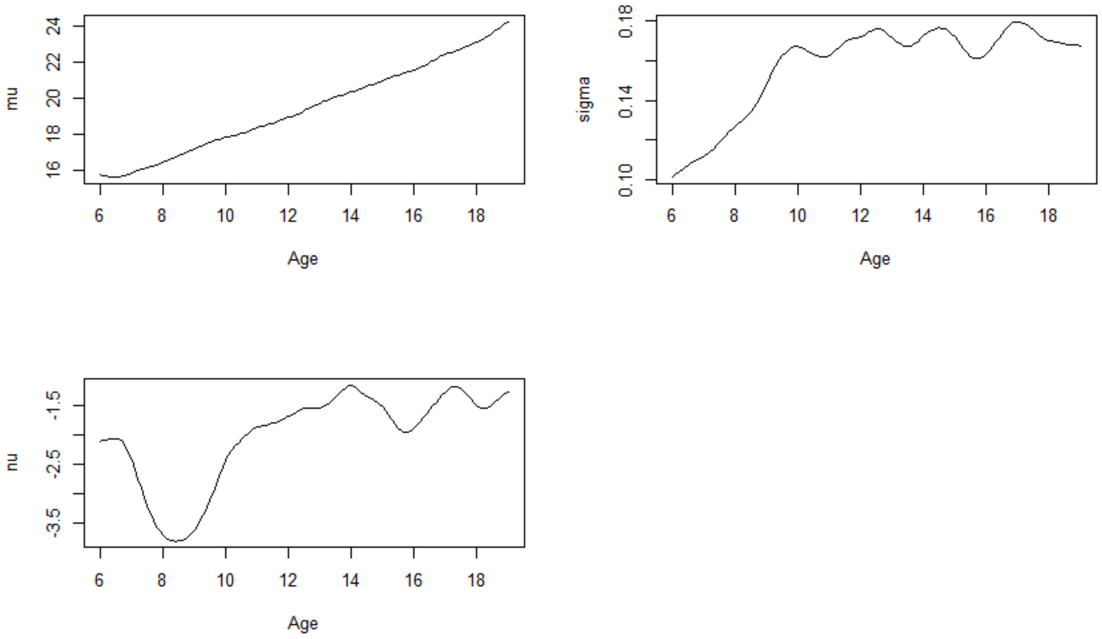


Figure 3.7: LMS(12,12,12) parameter curves for male BMI

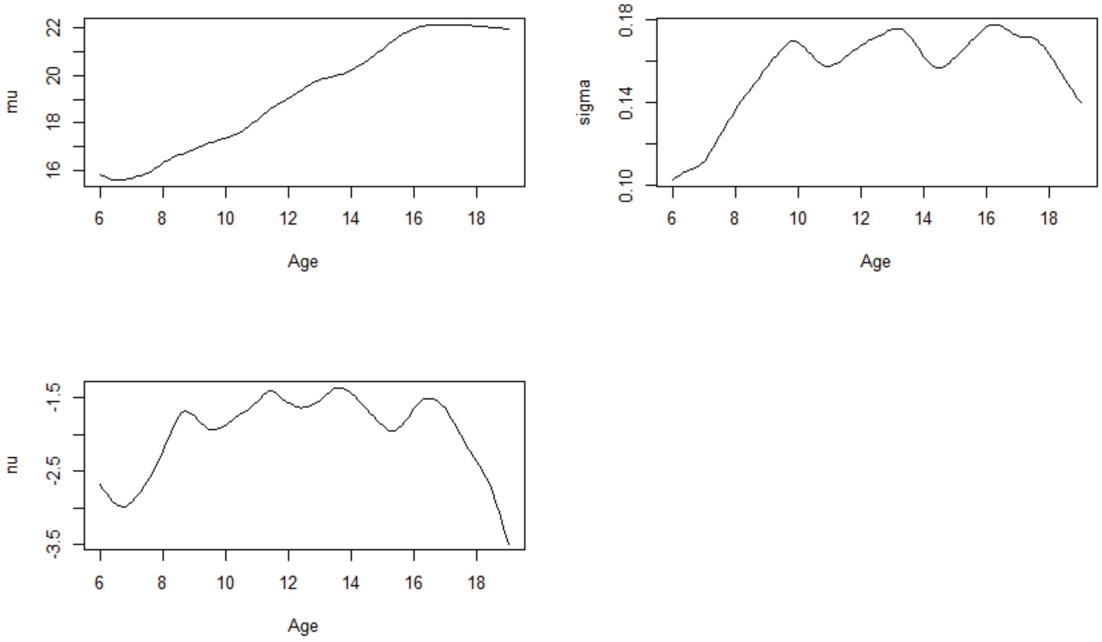


Figure 3.8: LMS(12,12,12) parameter curves for female BMI

The results are also reflected in the parameter plots seen in Figure 3.7. The σ parameter increases until age 10 where it reaches a noisy but relatively constant value. Comparing the parameter plots for low and high levels of smoothing indicates that the high level of smoothing forces the period during which the centile curves start to diverge to occur later. This is a result of the smoothing of the σ parameter into a smooth curve that changes values over a larger range than it should have. Comparing the power parameter ν between both models also shows that higher skewness associated with lower age ranges may actually only occur during this period of rapid change in the 7 to 10 year age range.

The female BMI growth charts with low amount of smoothing also show some new features, most notably that the mean parameter μ becomes constant after approximately age 16 (Figure 3.8). This indicates that the population mean of the BMI distribution stops increasing after this point. The female low smoothing fit also exhibits more and larger fluctuations in the highest centiles, suggesting that there is more noise and that female centile curves might require more smoothing when the final model is fitted.

Performing the same types of comparisons using the triceps skinfold thickness measurements produces results similar to those for the BMI measurements. Figure 3.9 and Figure 3.10 show that the high smoothing model for the males is possibly over-smoothing a bump structure in the 10-13 age range. Once again it is the σ and ν parameters where this over smoothing seems to be most noticeable, as seen in Figure 3.11 and Figure 3.12. The age of the peak mean measurements is approximately 2 years earlier in the lower smoothed graphs, akin to the BMI graphs.

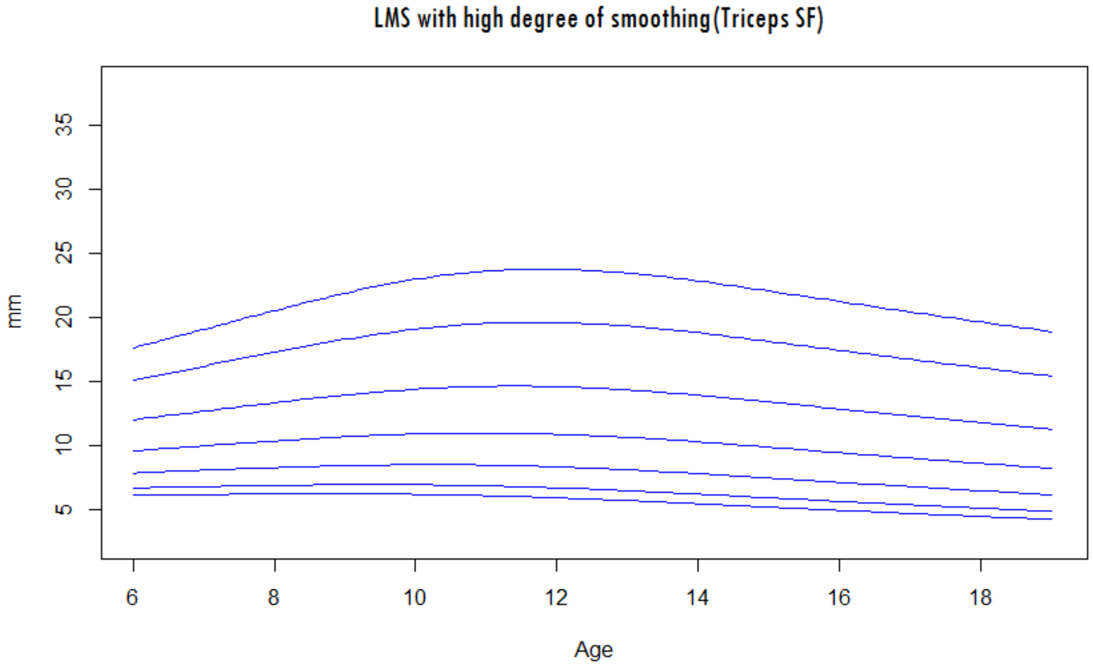


Figure 3.9: LMS(1,1,1) growth curves for male triceps

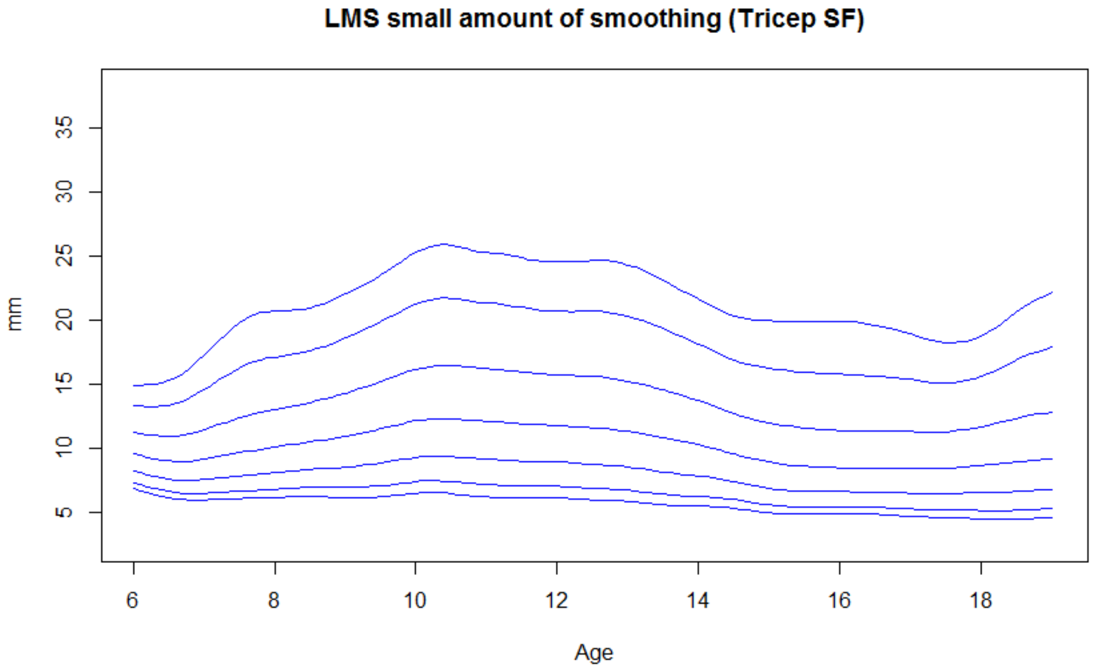


Figure 3.10: LMS(12,12,12) growth curves for male triceps

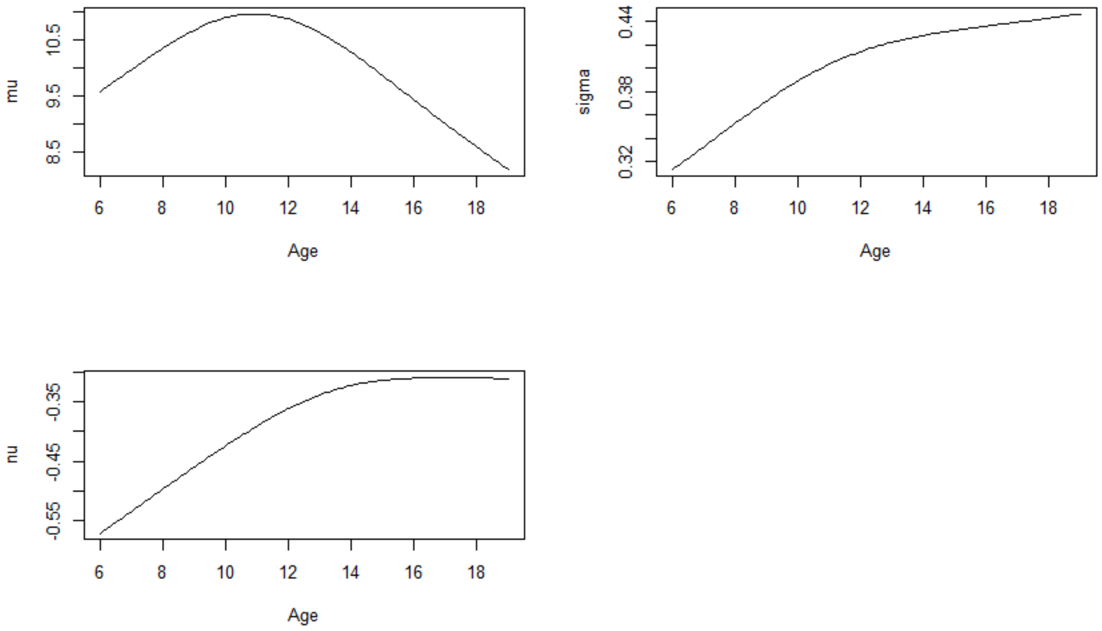


Figure 3.11: LMS(1,1,1) parameter curves for male triceps

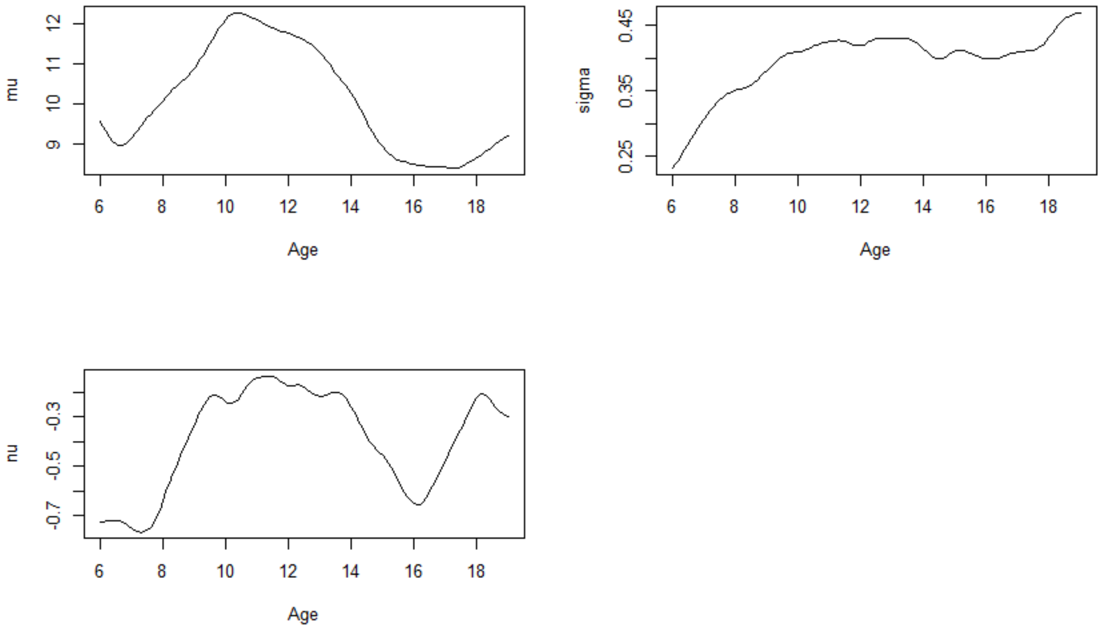


Figure 3.12: LMS(12,12,12) parameter curves for male triceps

The female triceps skinfold thickness measurements are shown in Figure 3.13 and Figure 3.14. Unlike BMI, the distribution of triceps skinfold thickness has a completely different shape between men and women. This supports the choice to create different growth charts for males and females. The mean female triceps skinfold thickness increases at a fairly constant rate as a function of age as seen in the μ parameter in both Figure 3.15 and Figure 3.16. Apart from adding some noise to the line, decreasing the amount of smoothing does not change the shape. The σ and ν parameters are also better approximated by the high smoothing model than with the previous three sex-variable combinations, suggesting that the final model for female triceps skinfold thickness will incorporate a higher amount of smoothing.

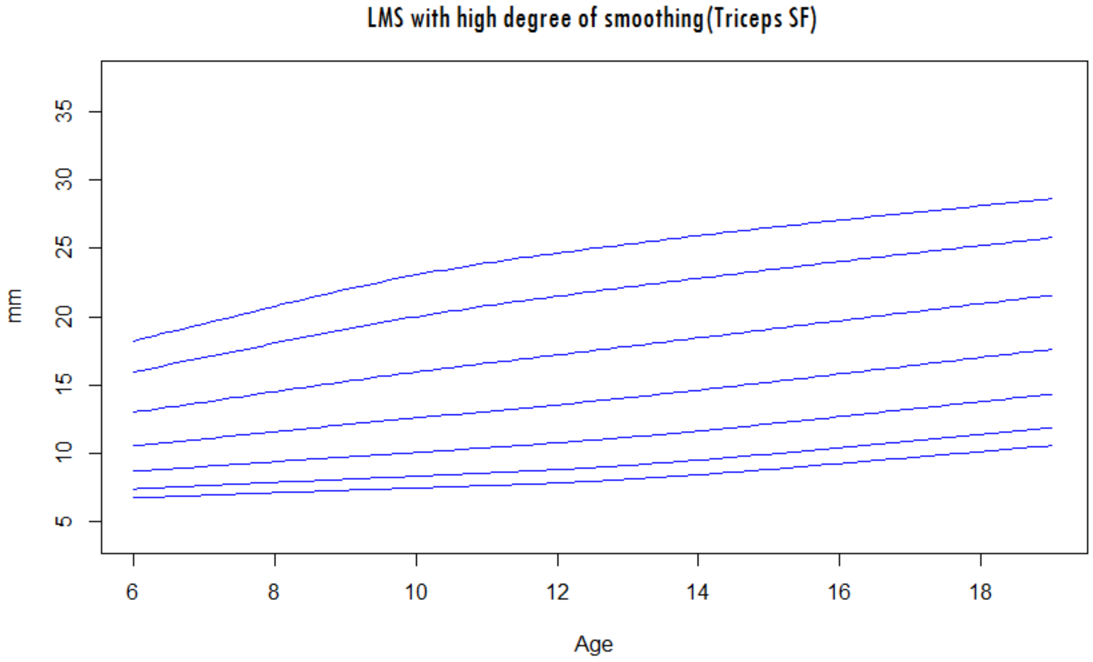


Figure 3.13: LMS(1,1,1) growth curves for female triceps

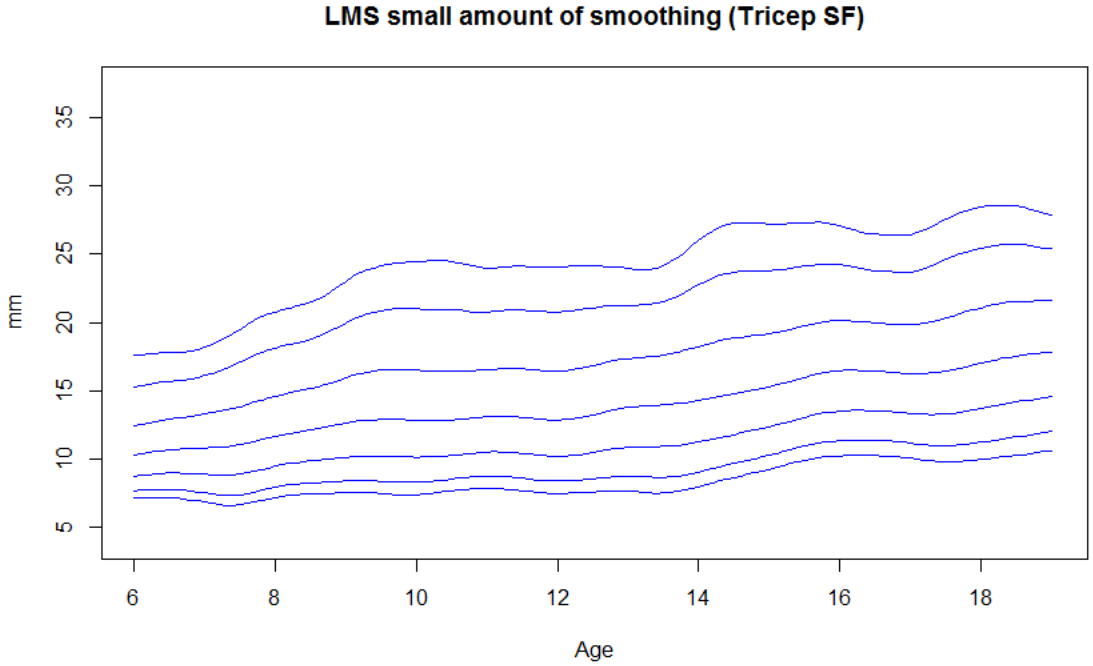


Figure 3.14: LMS(12,12,12) growth curves for female triceps

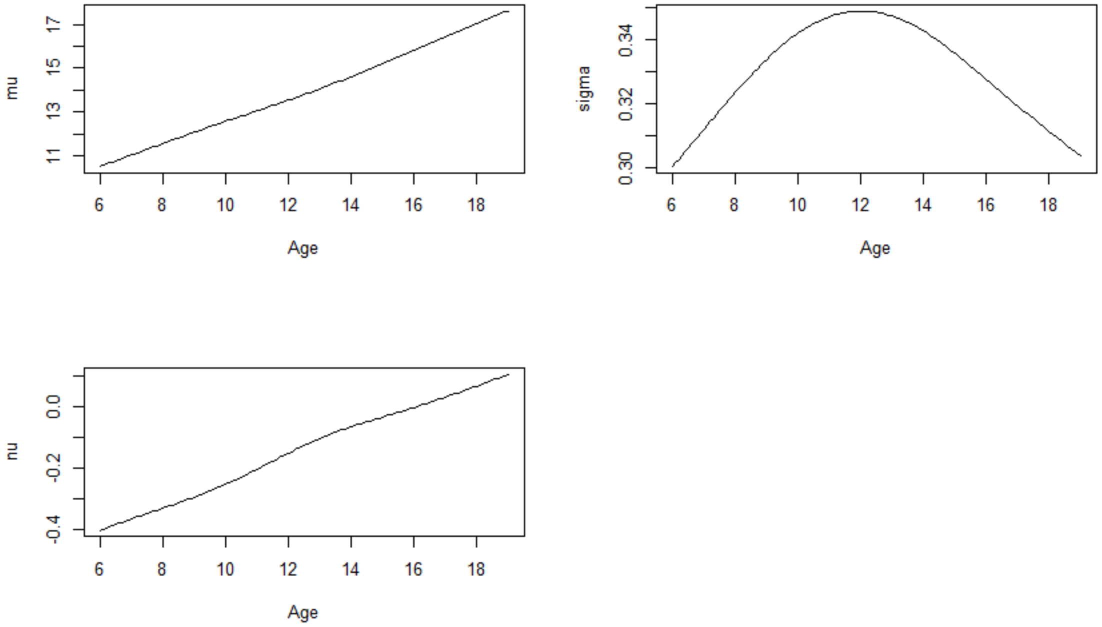


Figure 3.15: LMS(1,1,1) parameter curves for female triceps

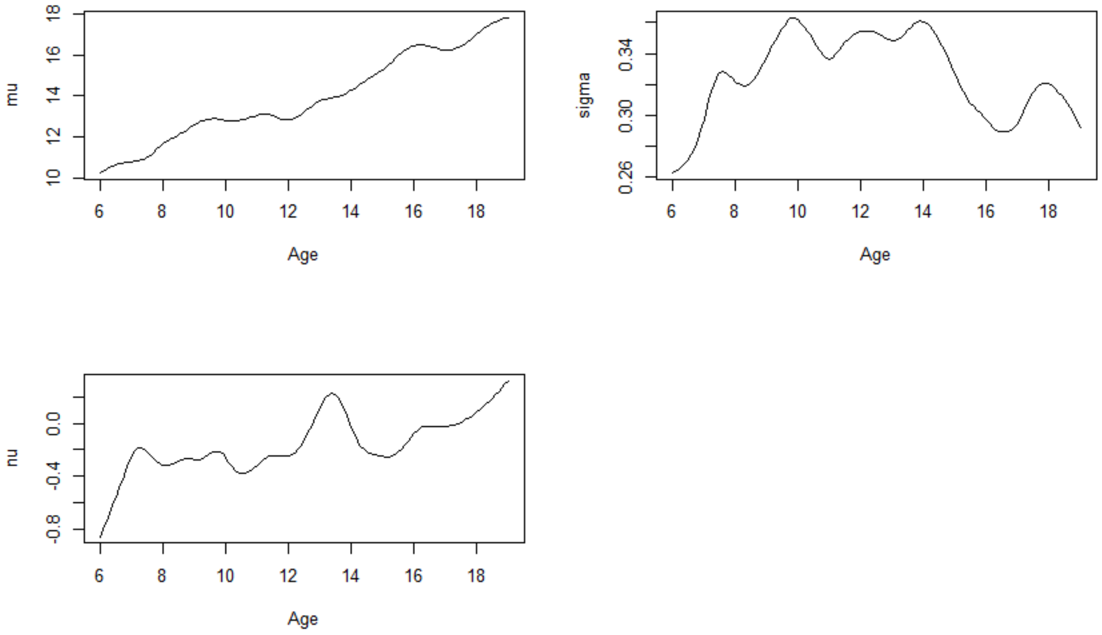


Figure 3.16: LMS(12,12,12) parameter curves for female triceps

Using the GAIC model selection method discussed in the previous section, a model was developed for each measure and gender. Each of the models fell somewhere between the two extremes of the models discussed above. During the selection process only integer *efds* were considered and constrained to be greater than or equal to 1 and less than or equal to 20. The edfs selected for each model are shown in table 3.1.

Measure	Gender	μ edf	σ edf	ν edf	GAIC $k = 3$
BMI	Male	3	3	5	15623
BMI	Female	6	10	2	14963
TRIC	Male	4	4	3	10768
TRIC	Female	4	3	1	10801

Table 3.1: LMS models selected though GAIC

Figure 3.17 shows the model selected for male BMI and Figure 3.18 shows its parameter plots. The model allows the power parameter enough flexibility to incorporate the increased skewness observed in the 7 to 10 years of age range. The σ parameter incorporates a more sudden bend, but not quite to the extent that was observed in the low smoothing model discussed above. This model avoids the undesirable wiggles and bumps observed with the low smoothing model.

The model selected for female BMI has a higher total effective degrees of freedom resulting in curves which show more fluctuations than those of their male counterpart. Its power component ν is nearly identical to the heavily smoothed model and μ has a slight S shape, but most of the rapid variation seen in Figure 19 comes from the σ parameter shown in Figure 20 that has comparatively little smoothing. It is unlikely that these rapid variations is σ reflect a biological process.

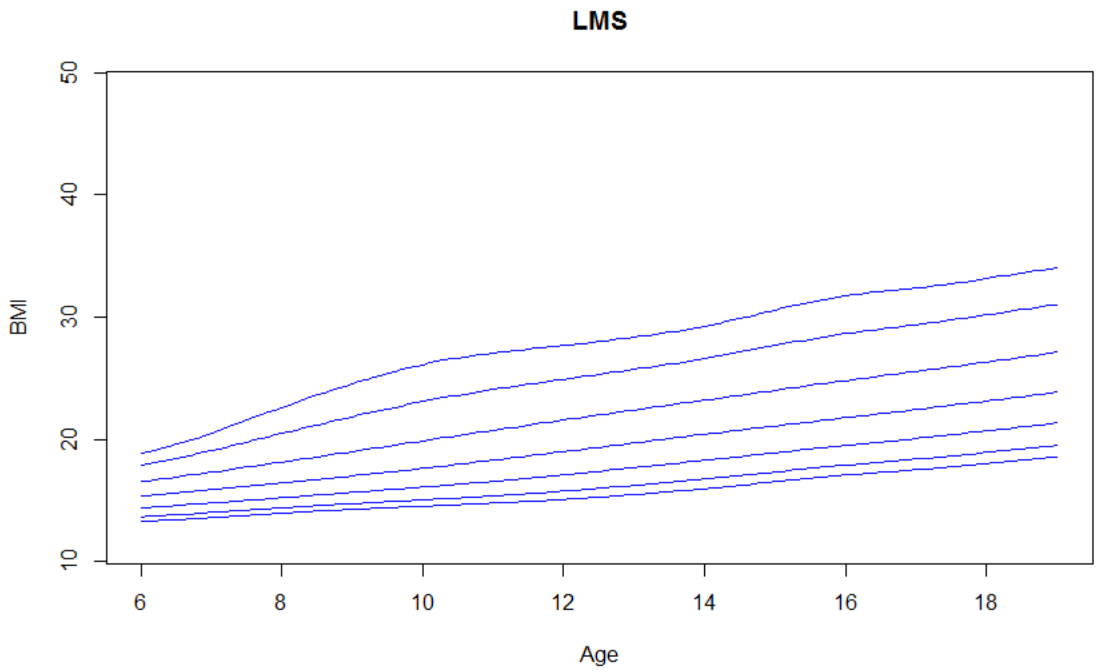


Figure 3.17: LMS(3,3,5) growth curves for male BMI

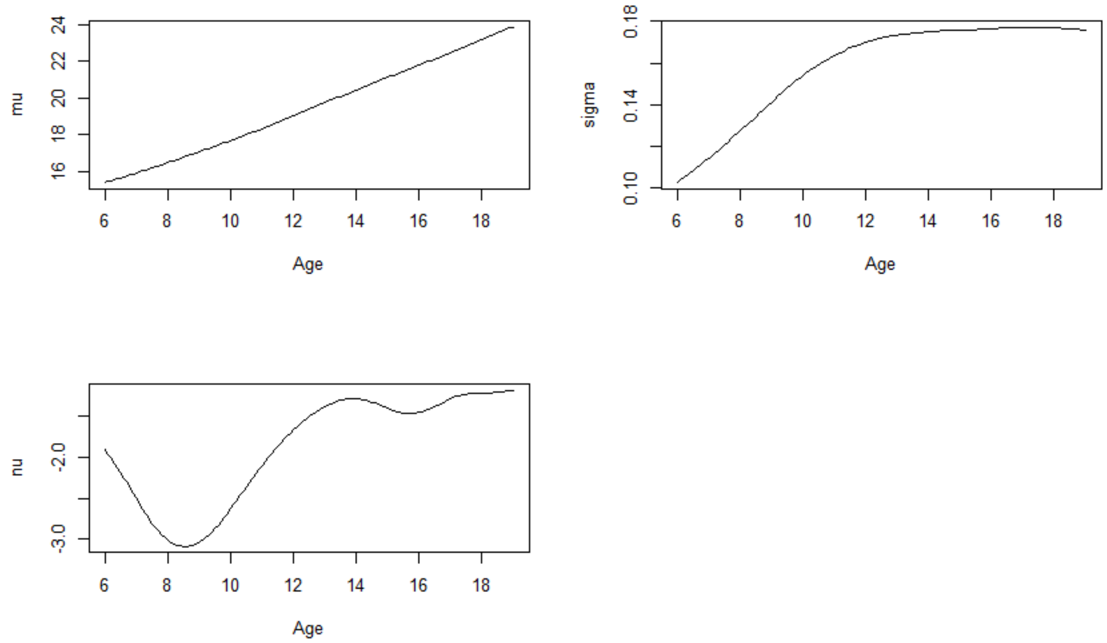


Figure 3.18: LMS(3,3,5) parameter curves for male BMI

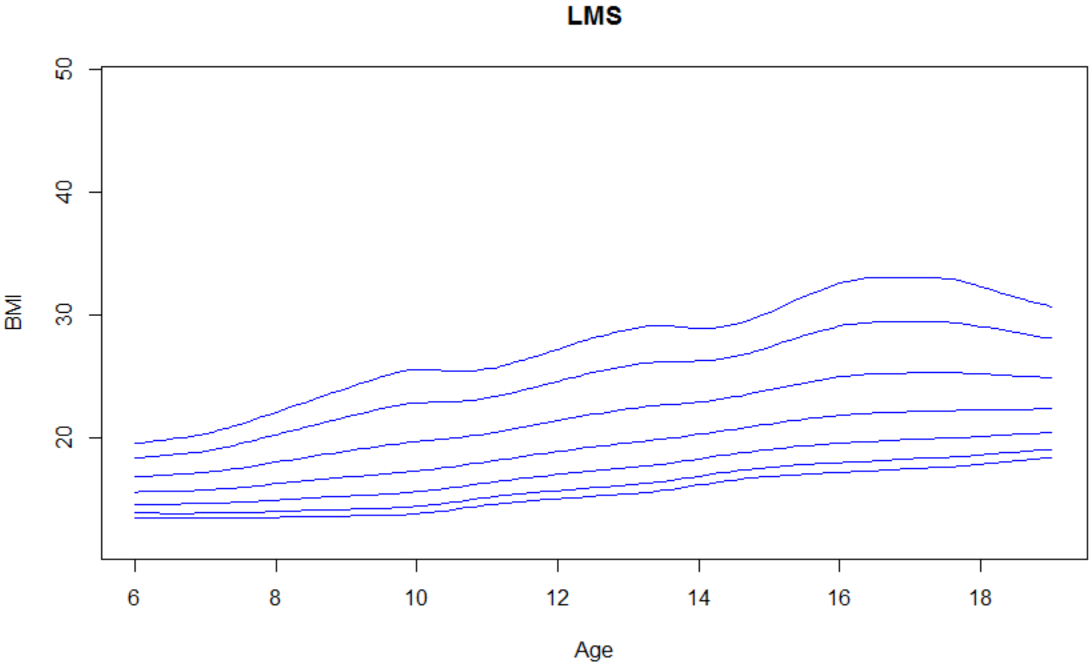


Figure 3.19: LMS(6,10,2) growth curves for female BMI

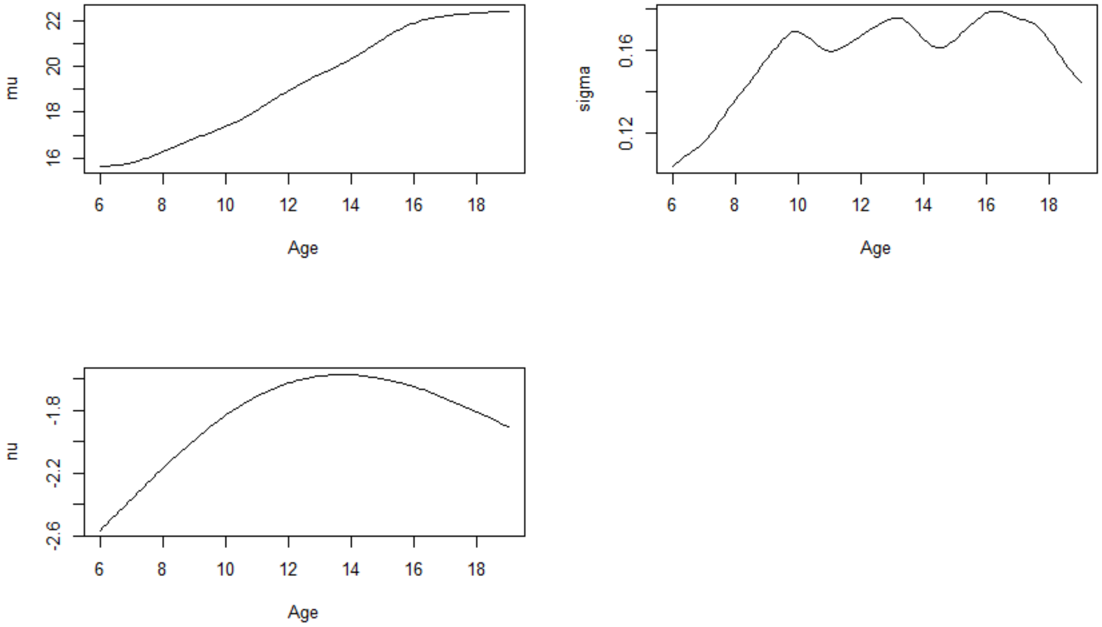


Figure 3.20: LMS(6,10,2) parameter curves for female BMI

The models selected for male and female triceps skinfold thickness have a similar amount of smoothing in both the μ and σ parameters. The mean for females increases with age, males increase to age 12 and then decreases. Females, however, have an almost linear ν function that steadily increases toward zero (Figure 3.24), which is reflected in Figure 3.23 which shows that the spacing between the bottom centiles is almost the same as the spacing between the highest centiles. The males have a comparatively much more skewed distribution which appears to have a biologically plausible lower limit that many points are clustered against for all age ranges seen in Figure 3.21.

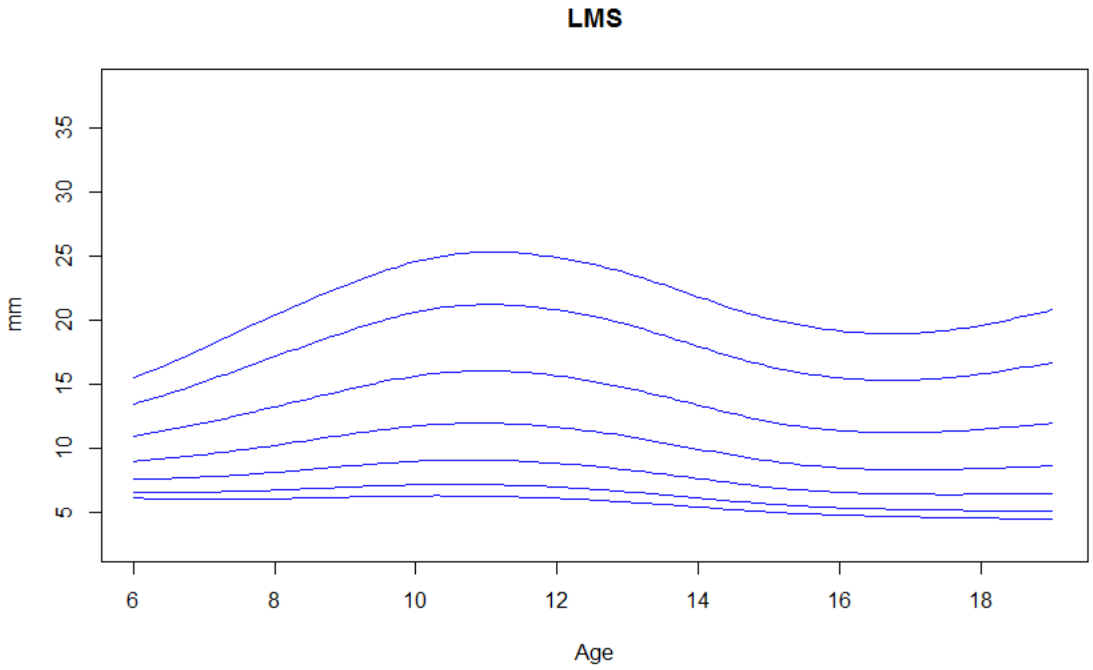


Figure 3.21: LMS(4,4,3) growth curves for male triceps

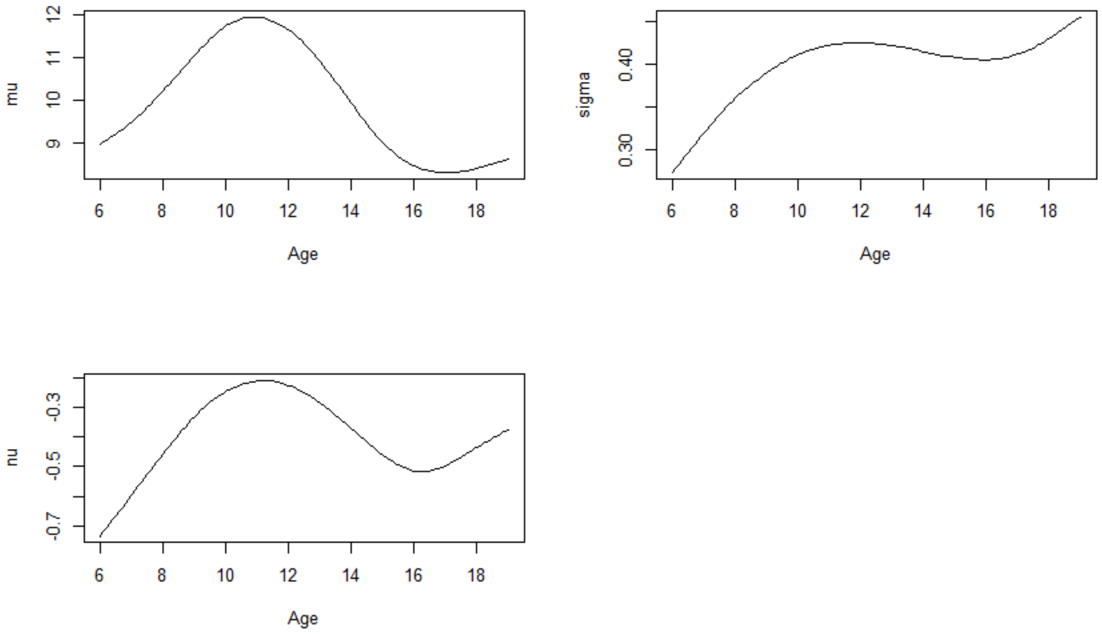


Figure 3.22: LMS(4,4,3) parameter curves for male triceps

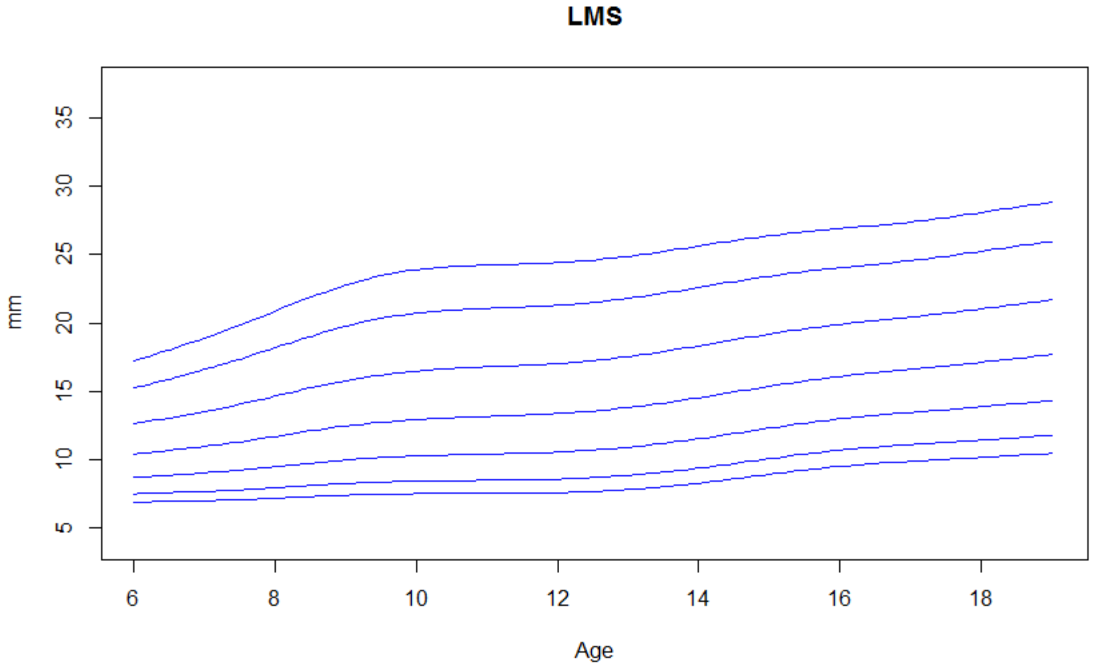


Figure 3.23: LMS(4,3,1) growth curves for female triceps

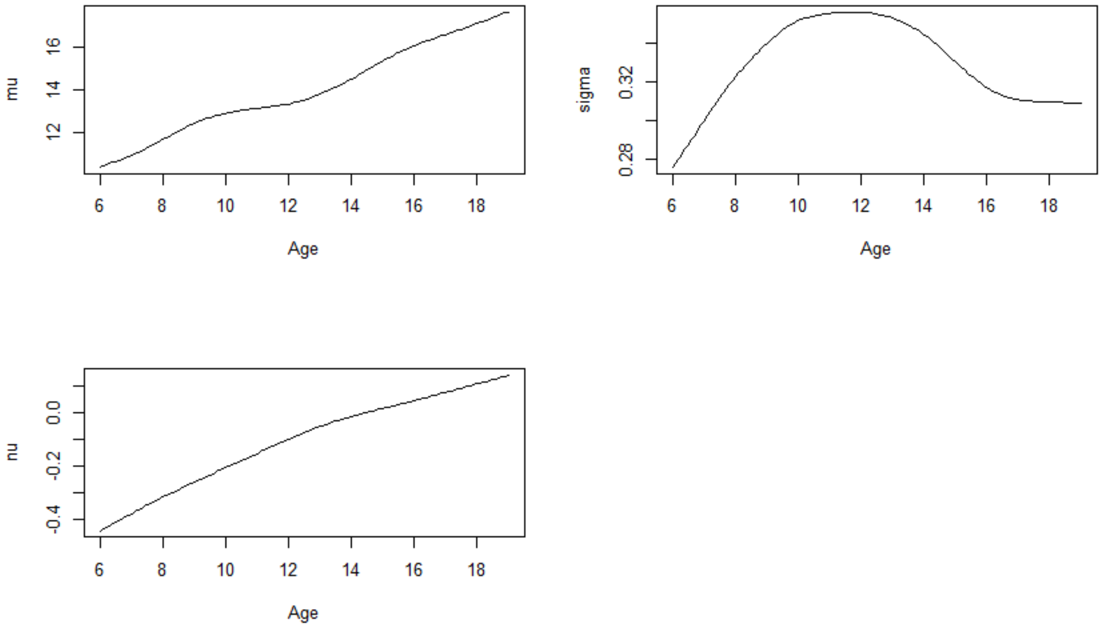


Figure 3.24: LMS(4,3,1) parameter curves for female triceps

3.2.1 BCPE Models

Previous studies have shown that the distributions of anthropometric measures can contain kurtosis (Rigby and Stasinopoulos, 2004). When using the LMS method this kurtosis goes unmodeled because there is no explicit term for it, as it is a function of μ, σ and λ. The BCPE distribution allows the models to have kurtosis where needed, without requiring the distribution to be kurtotic at all age ranges. This added flexibility helps prevent the incorrect modeling of skewness to attempt to compensate for unmodeled kurtosis. The BCEP models were fitted to each gender and measurement combination using GAIC to select the optimal model. This process functions the same as with the LMS method except that the search for optimal smoothing constants now occurs over a four dimensional parameter space rather than three. This increases the computation time significantly, which is one of the drawbacks of choosing this method. The models selected are shown in Table 3.2.

Measure	Gender	μ edf	σ edf	ν edf	τ edf	GAIC $k = 3$
BMI	Male	3	4	4	5	15621
BMI	Female	5	8	3	4	14971
TRIC	Male	4	4	4	5	10708
TRIC	Female	4	4	1	3	10769

Table 3.2: BCPE models selected though GAIC

The addition of the τ parameter did not dramatically change the amount of smoothing applied to the mean and variance parameters for any of the models.

Figures 3.25 and 3.26 show the centiles produced and the parameter curves for male BMI. Comparing them to the LMS models (Figure 3.18) shows they are nearly identical, with differences only visible in the highest centile. This is reflected in the μ and σ parameters which also have a near identical shape. The skewness parameter for the BCPE model shows a linear increase to age 12 instead of the decrease to age 8 for the LMS method. The kurtosis parameter starts near 1 indicating a heavy tailed distribution at young ages. After age 8, the kurtosis varies around two, which corresponds to a normal distribution.

Female BMI curves, using the BCPE model are also relatively similar to those from the LMS method as seen in Figures 3.27 and 3.28 (compared to Figures 3.19 and 3.20). The waves that were present in the LMS growth curves are still visible in the BCPE curves but are slightly less pronounced. The three parameter curves from the the LMS model also have an almost identical shape to the curves from the BCPE model. An exception to this is between 16 and 18 years of age where the BCPE variance curve decreases less than the LMS variance curve. In this range there is a dramatic rise in τ. The net effect of these differences is that the upper centiles decrease less than for the LMS models.

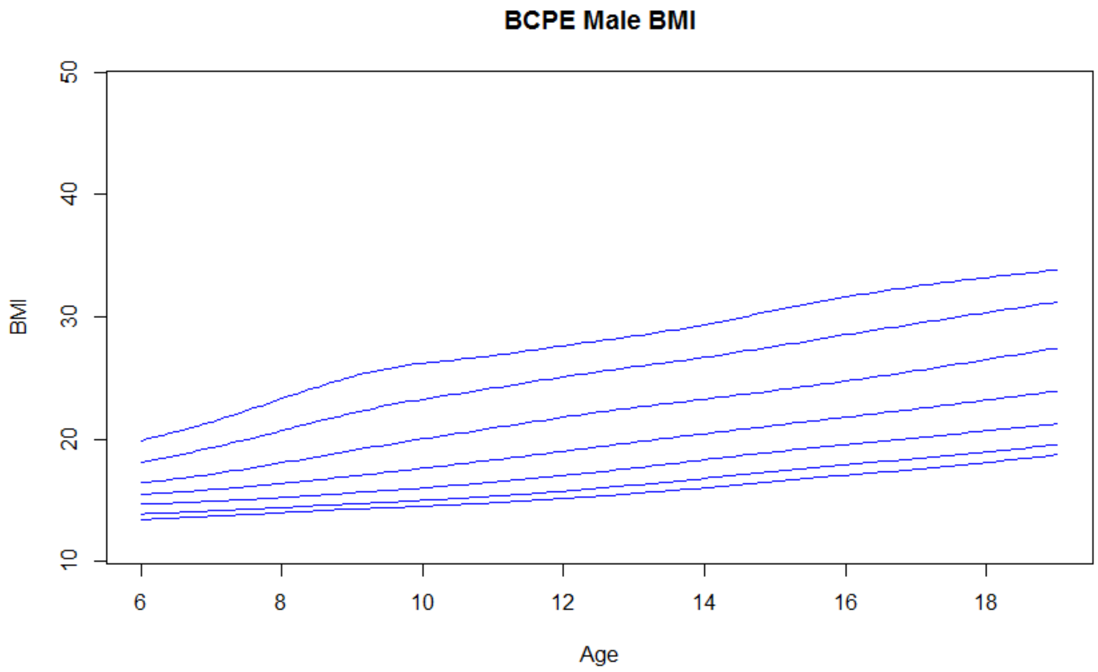


Figure 3.25: BCPE(3,4,4,4) growth curves for male BMI

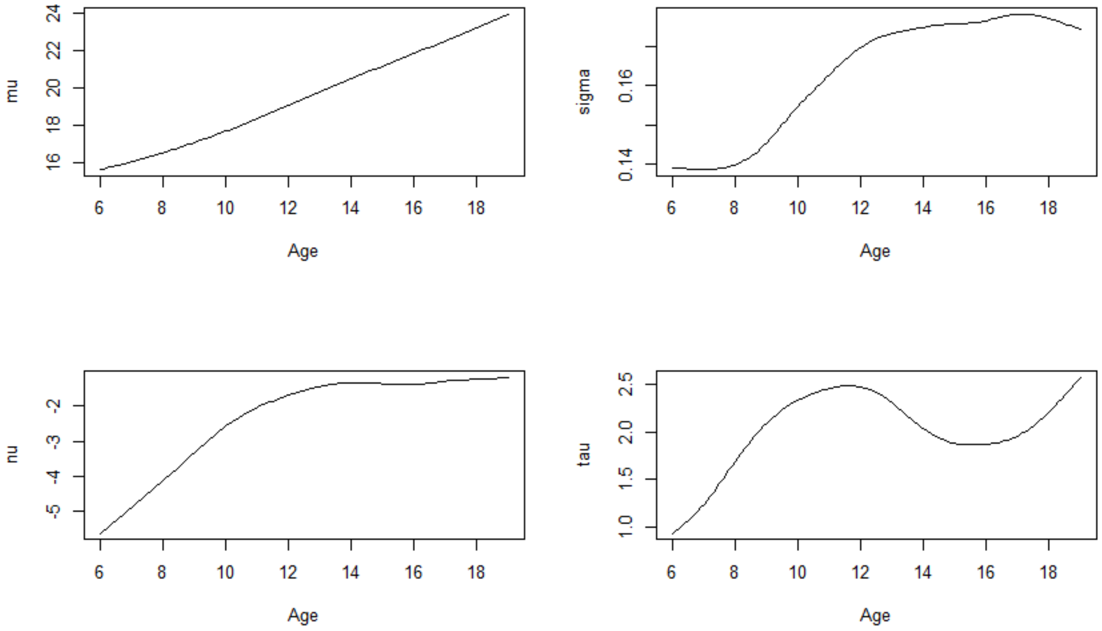


Figure 3.26: (3,4,4,4) parameter curves for male BMI

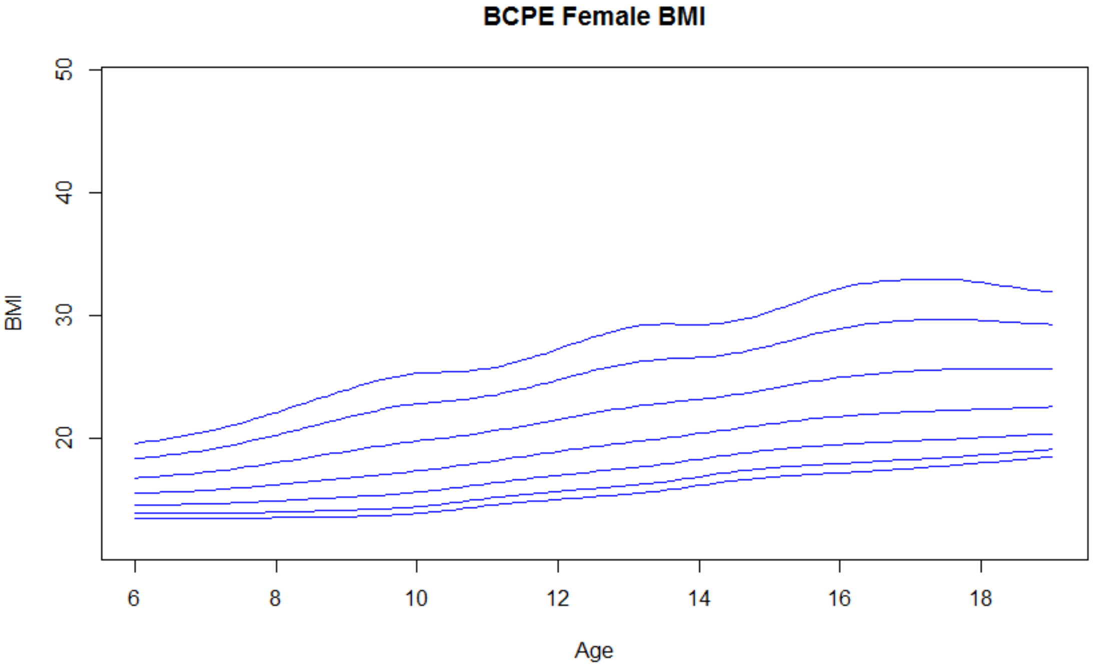


Figure 3.27: BCPE(5,8,3,4) growth curves for female BMI

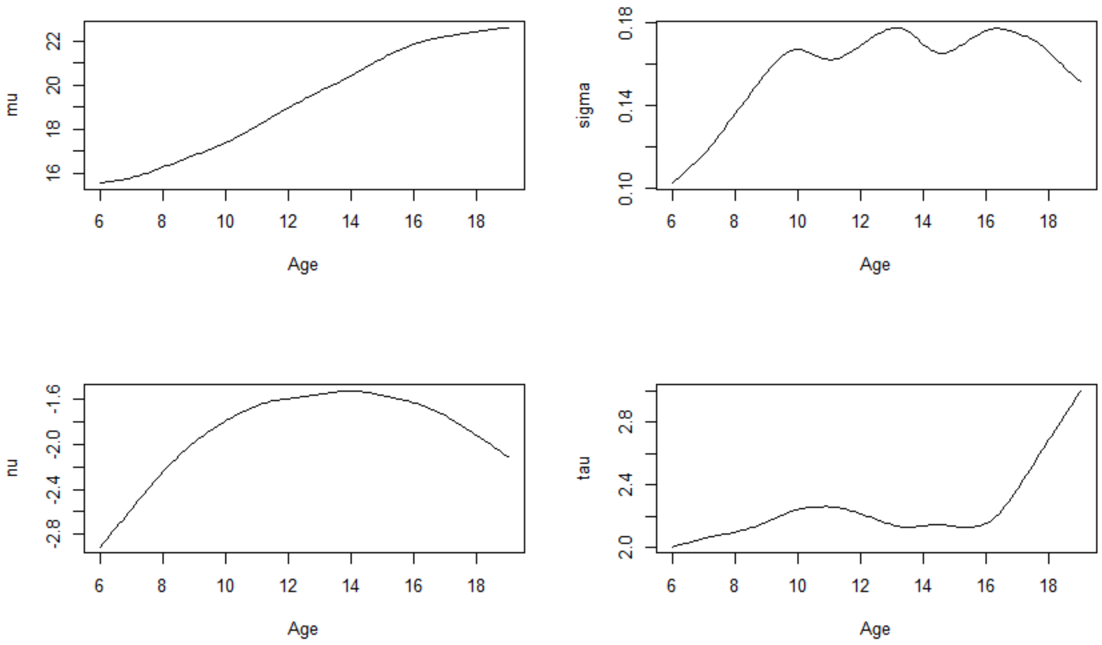


Figure 3.28: (5,8,3,4) parameter curves for female BMI

Automatic model selection for triceps BCPE models did not choose models that were very different from the LMS models for males or females. For males, one degree of freedom was added to the ν parameter and the τ parameter was choosen to have an edf of 5. The τ curve increases until peaking at approximately age 11 then drops again (Figure 3.30), coinciding with the peak seen in the centile curves themselves (Figure 3.29). The model chosen for female triceps skinfold is also very similar to the LMS model. The very low edf on the ν parameter remains at 1, meaning it retains its near linear shape (Figure 3.31). The addition of the τ parameter causes the actual shape of the curves to change very little (Figure 3.32).

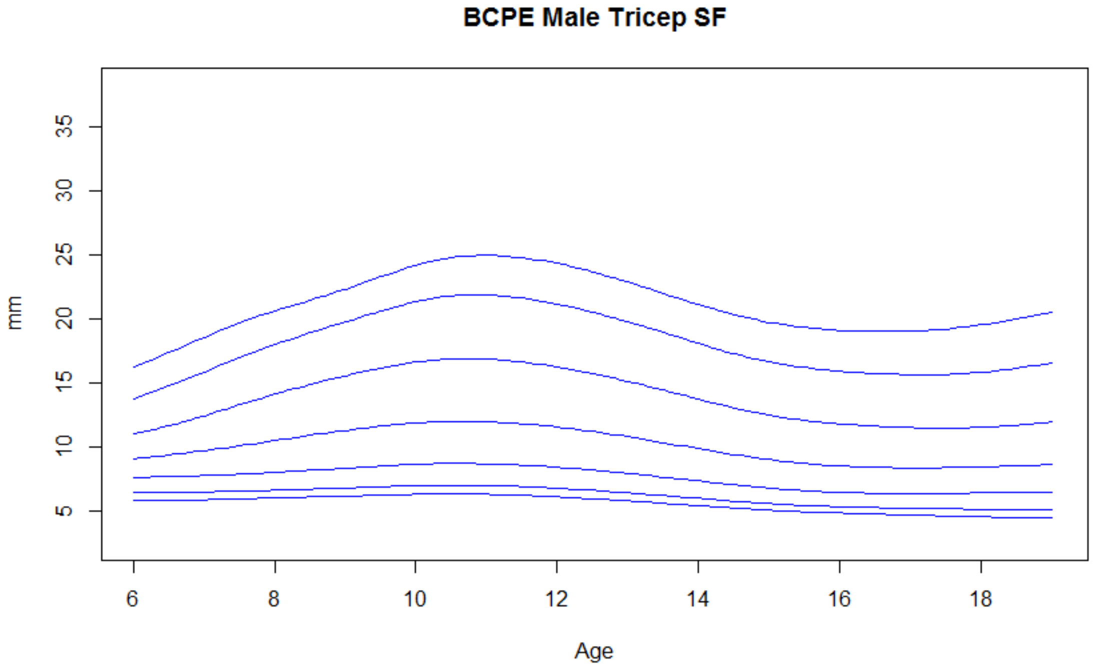


Figure 3.29: BCPE(4,4,4,5) growth curves for male triceps

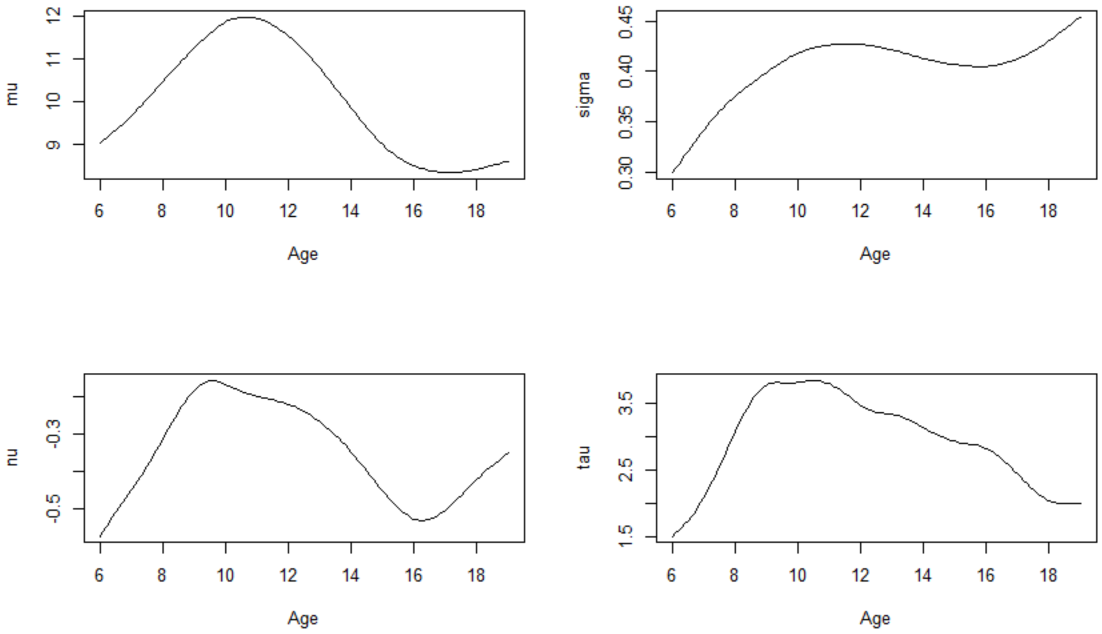


Figure 3.30: BCPE(4,4,4,5) parameter curves for male triceps

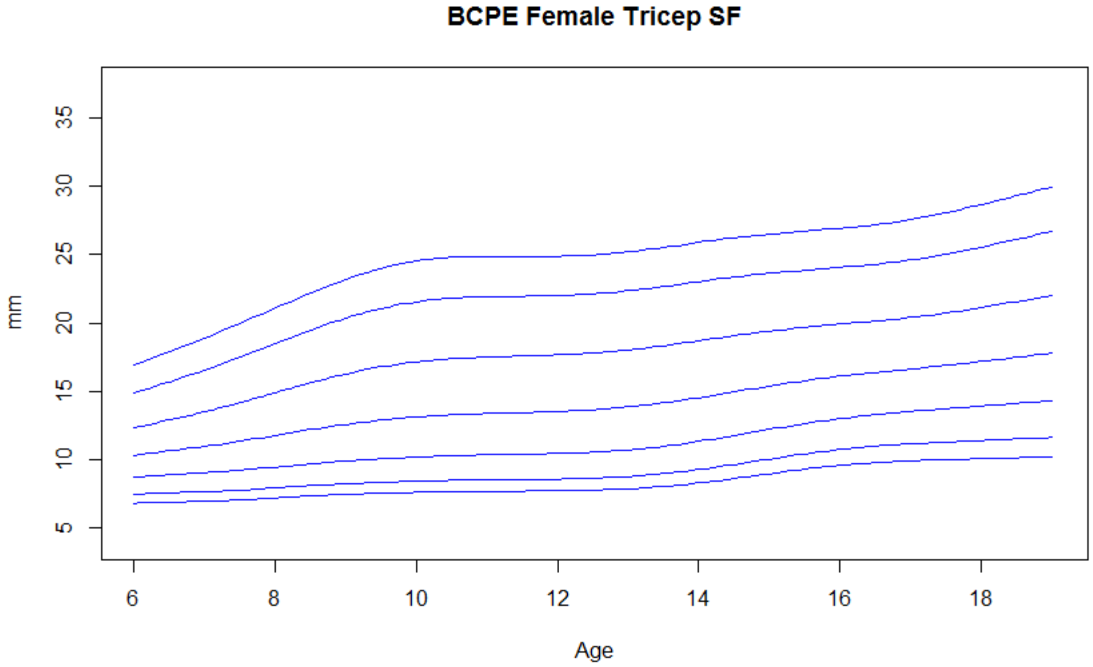


Figure 3.31: BCPE(4,4,1,3) growth curves for female triceps

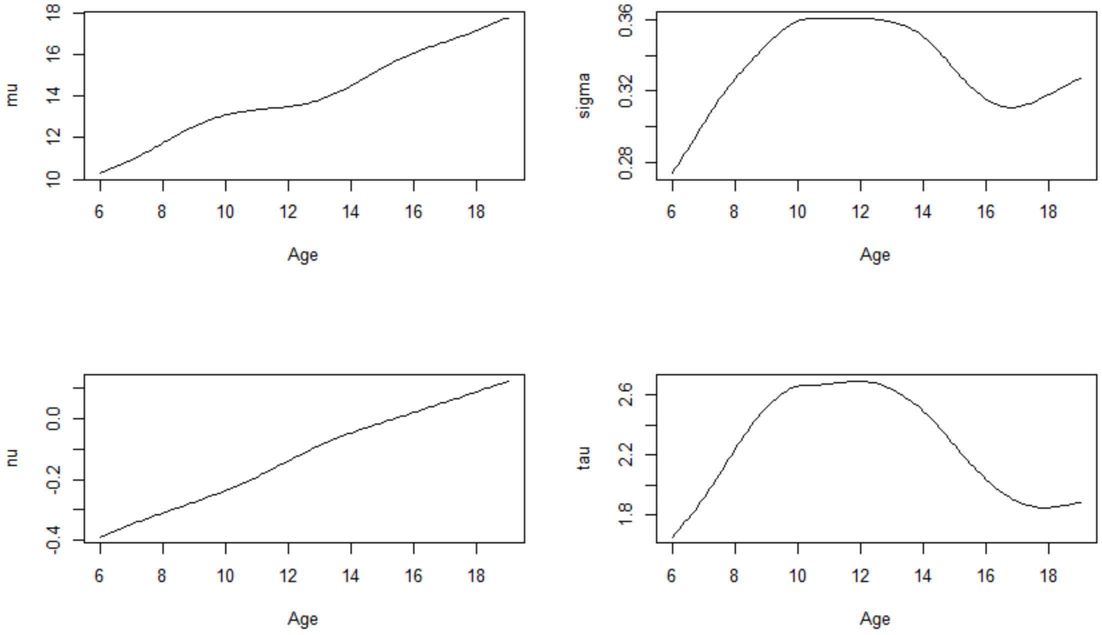


Figure 3.32: BCPE(4,4,1,3) parameter curves for female triceps

Chapter 4

Diagnostics

4.1 Worm Plots

Van Buuren and Fredriks (2001) showed how a detrended QQ-plot called a worm plot could be used to assess model fit when constructing growth curves. The paper explains how the method can be used for any model where a likelihood can be explicitly calculated, such as the LMS or other GAMLSS methods. The method was later extended to quantile regression by (Buuren, 2007). The worm plot is a useful diagnostic for fitting growth charts because it not only identifies if there is a lack of fit to the data but also where the problem occurs and what might be done to correct it.

In a normal QQ-plot the empirical quantile of each data point is plotted against its theoretical quantile. For models that fit data extremely well the empirical quantile and the theoretical quantiles should be very close and the data points will all lie on a straight line. Deviation from this straight line can be a sign of a poor fit; however, it can be difficult to tell what amount of deviation is cause for concern or what exactly causes the deviations. In a worm plot, the vertical axis is replaced with the difference of the empirical quantile and the theoretical quantile. These points form the namesake “worm” and a flat worm indicates the data follows the assumed distribution. Figure 4.1 (Van Buuren and Fredriks, 2001) is an example of a normal QQ plot and a detrended QQ plot or worm plot for male BMI over all ages. It is much easier to see the departure from the line in the detrended plot.

Creating a worm plot for the entire data set shows how well the overall model fits but does not show how well the model fits conditional on age. Splitting the data by age reveals how well the model fits for different age ranges. The empirical quantiles are recalculated for the data within each age range and the worms are plotted. If the worm for an age range dose not lie flat on the zero line there is some aspect of the model that needs to be improved in that age range. To help visualize where we

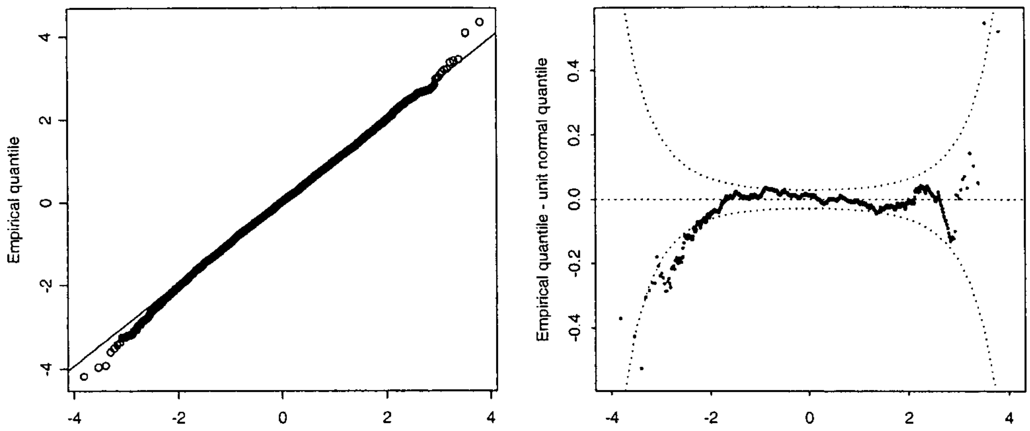


Figure 4.1: QQ-plot and worm plot (Van Buuren and Fredriks, 2001)

expect the the worms to lie the 95% confidence intervals of the theoretical quantiles are also plotted. Figure 4.2 (Van Buuren and Fredriks, 2001) shows the worm plots for the same data split up among 16 age ranges with an equal number of observations. The lowest age range is the the bottom left corner, increasing to the right then up through the rows with the highest age range in the top right panel.

The worms in these plots show that there are many more problems with the model fit than the global worm plot would indicate. Several of the age ranges have worms that do not lie on the zero line or which cross the confidence interval boundaries. The process of improving the model fit is to change the model tuning parameters one at a time and to recreate the worm plots until a relatively good fit is achieved in each age range. (Van Buuren and Fredriks, 2001) suggest the interpretation of the shapes of the worms in Table 4.1.

Moment	Worm Shape	Diagnosis
Mean	passes above the origin	fitted mean is too small
	passes below the origin	fitted mean is too large
Variance	has a positive slope	fitted variance is too small
	has a negative slope	fitted variance is too large
Skewness	has a U-shape	fitted distribution is too skewed to the left
	has an inverted U-shape	fitted distribution is too skewed to the right
Kurtosis	has an S-shape on the left bent down	tails of the fitted distribution are too light
	has an S-shape on the left bent up	tails of the fitted distribution are too heavy

Table 4.1: Interpretation of various patterns in the worm plot

For methods such as the LMS and BCPE, the tuning parameters available are the effective degrees of smoothness on each spline corresponding to the parameters

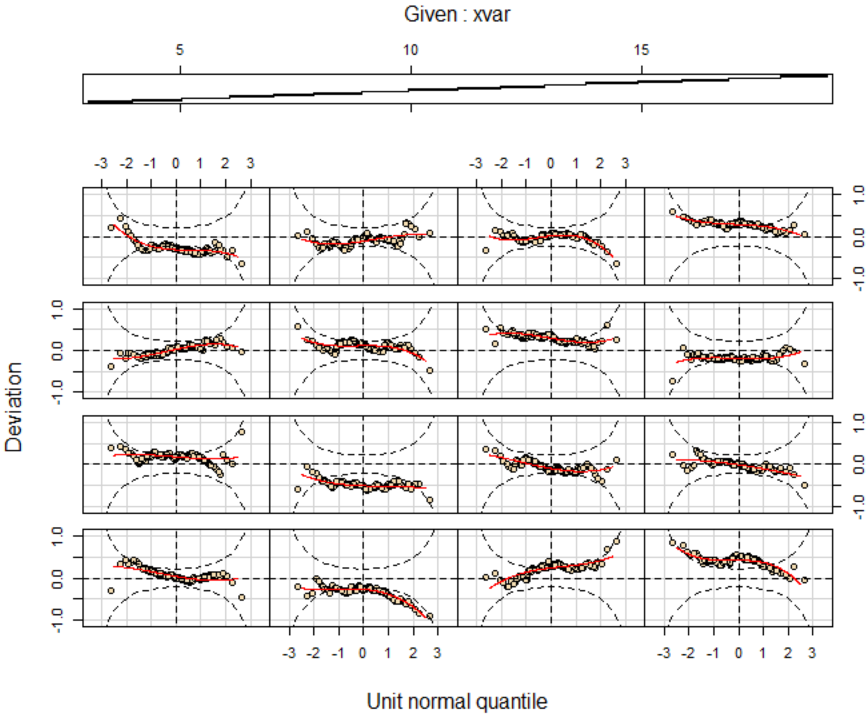


Figure 4.2: Age stratified worm plots (Van Buuren and Fredriks, 2001)

of the fitted distribution. Since the amount of smoothing applied is the same across all age ranges we have to try and fit them all simultaneously. We first begin with a model determined through some optimization program or fitting process. The order of tuning used by (Van Buuren and Fredriks, 2001) is to increase the effective degrees of freedom on the mean until the worms all intersect the zero line, then to increase the degrees of freedom on the variance until each worm's slope is flat. Finally, increase the effective degrees of freedom on any remaining tuning parameters such as kurtosis and skewness until the worms are relatively straight.

Worm Plots with Quantile Regression

Quantile regression has no underlying assumption about the distribution of the data so a method is needed to extract the theoretical quantiles to compare against the empirical quantiles. This is done by approximating a theoretical distribution by fitting many quantile regression models over a fine spacing of percentiles; for example 100 centiles each using the same set of basis splines for $\tau = \{0.01, 0.02, \ldots, 0.99\}$. Each data point falls somewhere between two of the models and linear interpolation can be used to approximate its quantile to compare against its empirical quantile (Buuren, 2007).

Improving the fit of the quantile regression model based on the information in the worm plot is done by adding knots to places where the model exhibits poor fit. Packing knots more densely allows the model to bend more in that interval and to provide a better fit. Extra knots can be useful for the ages around puberty, for example, when children's bodies change much faster than later in life. Worm plots can also be used to help protect against over-fitting by checking to see if any points lie outside the 95% confidence intervals. We would expect to see 5% of the points outside these intervals so if we see none it can be a sign that the model is over fitted and that knots can be removed from that age range to provide a smoother fit that is more reflective of the underlying biological process.

4.1.1 Using Worm Plots to Improve Model Fits

Starting with male BMI, in the LMS worm plots (Figure 4.3) some of the panels show a distinct S shape, which indicates the presence of kurtosis. This is noticeably improved in the BCPE model (Figure 4.4) as seen in panels 1, 4 and 5. Adding the fourth parameter has helped model kurtosis that is present in the data. For this reason we use the BCPE model. However this model is not without its flaws: Panels 3, 6, and 9 show positive or negative slopes and the worm of panel 9 lies below the center line. Increasing the effective degrees of freedom reduces some of these problems but does not fix them entirely. Particularly noticeable is panel 3 where a large section of the worm passes above the confidence interval. Changing the effective degrees of freedom for any of the parameters proved ineffective in correcting this problem unless they were allowed to vary so liberally that much larger problems were caused in other regions of the model. We conclude that the data deviates from the assumed model in this age range (approximately 9 to 11.5 years old). The final model chosen for male BMI is BCPE(4,6,4,4) (Figure 4.5).

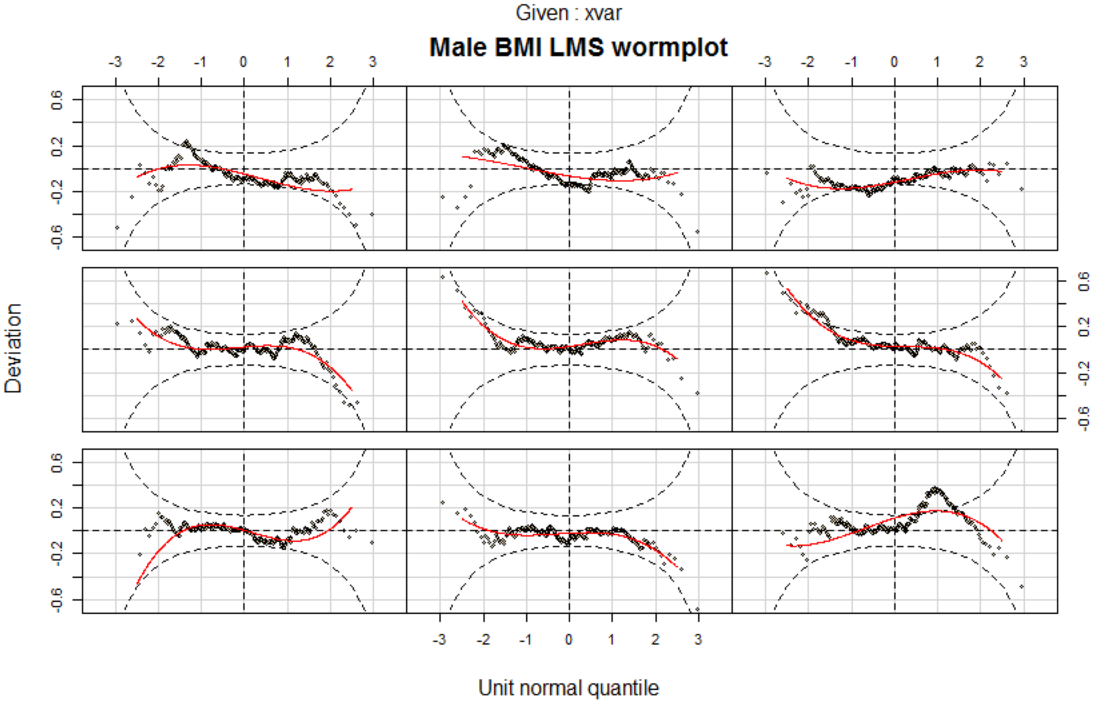


Figure 4.3: Worm plots for Male BMI LMS model

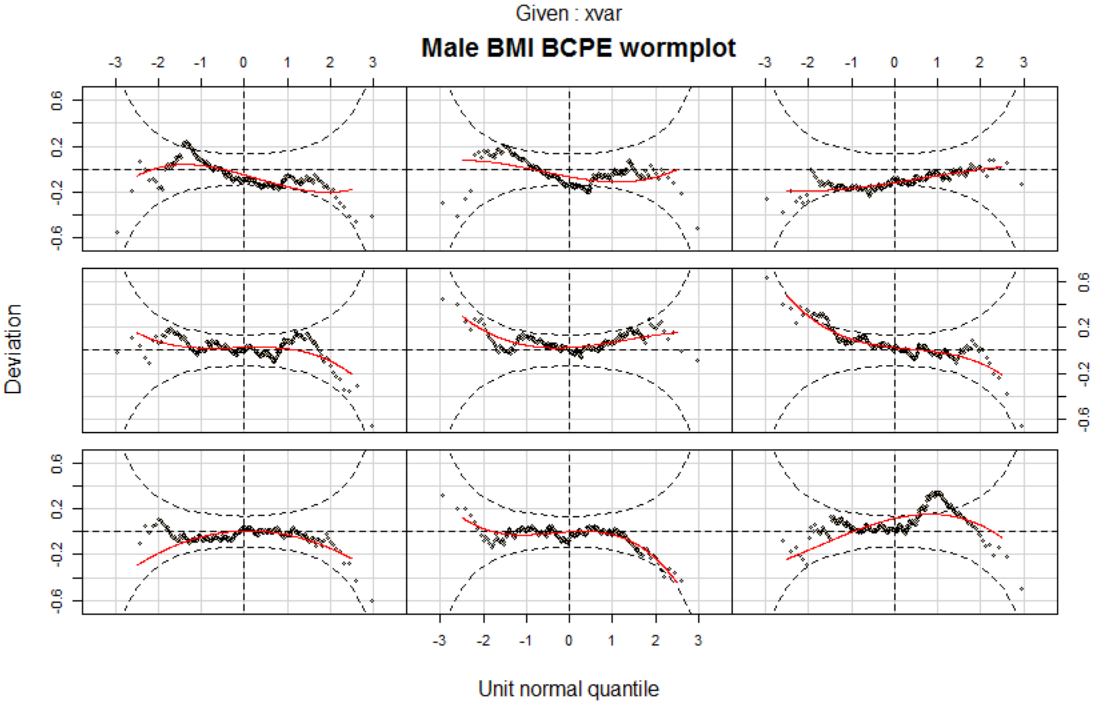


Figure 4.4: Worm plots for Male BMI BCPE model

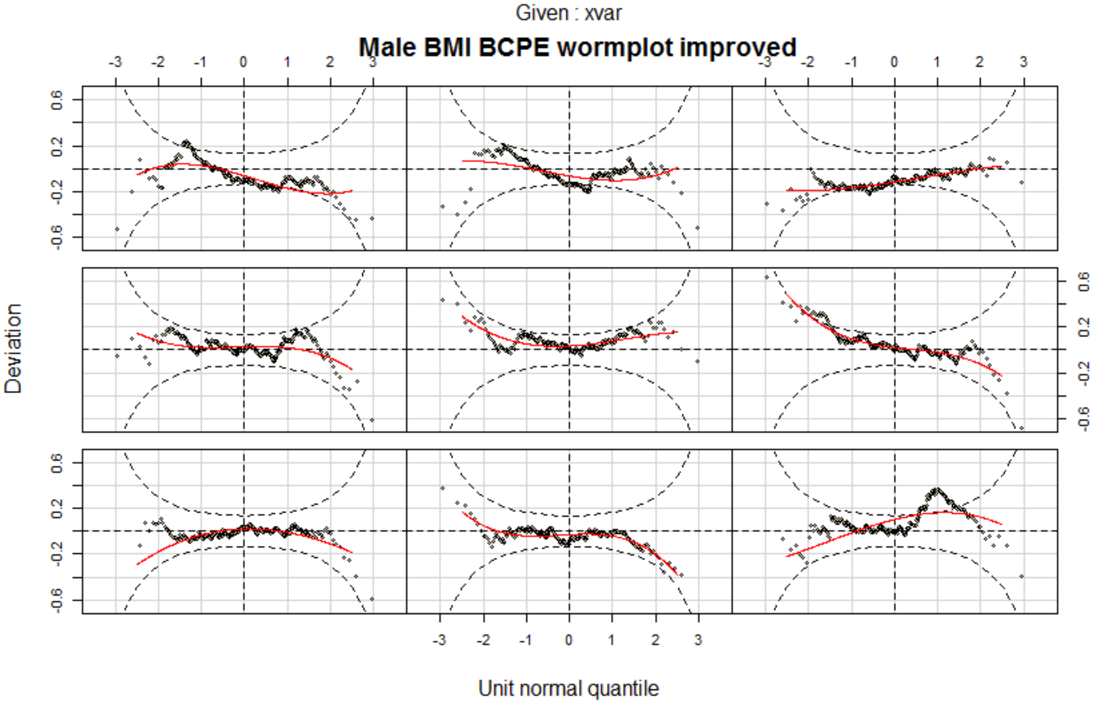


Figure 4.5: Worm plots for Male BMI improved BCPE model

Similarly to the male BMI data, the female BMI data shows evidence of kurtosis that goes unmodeled in the LMS model (Figure 4.6), indicated by the S pattern in panels 3, 5, 6, and 9 that is improved when using the BCPE method (Figure 4.7). While the BCPE model improves the problems caused by kurtosis, several of the worms still have a large positive or negative slope indicating incorrectly fitted variance in those regions. The automatic model selection method used in the previous section chose a rather large effective degrees of freedom for the σ parameter at 8. By reducing it to 4 we see in the worm plots that these problems are much abated, though not eliminated, and that all of the worms tend to lie flatter on the line. The improved model for Female BMI is BCPE(5,4,3,4) (Figure 4.8).

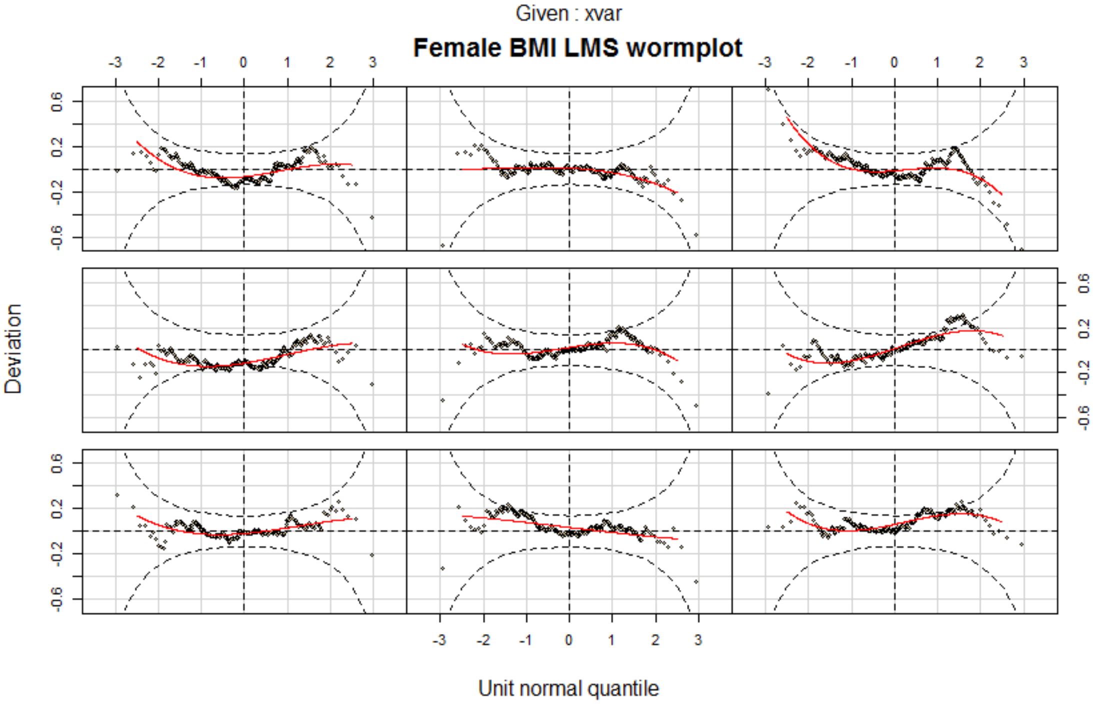


Figure 4.6: Worm plots for Female BMI LMS model

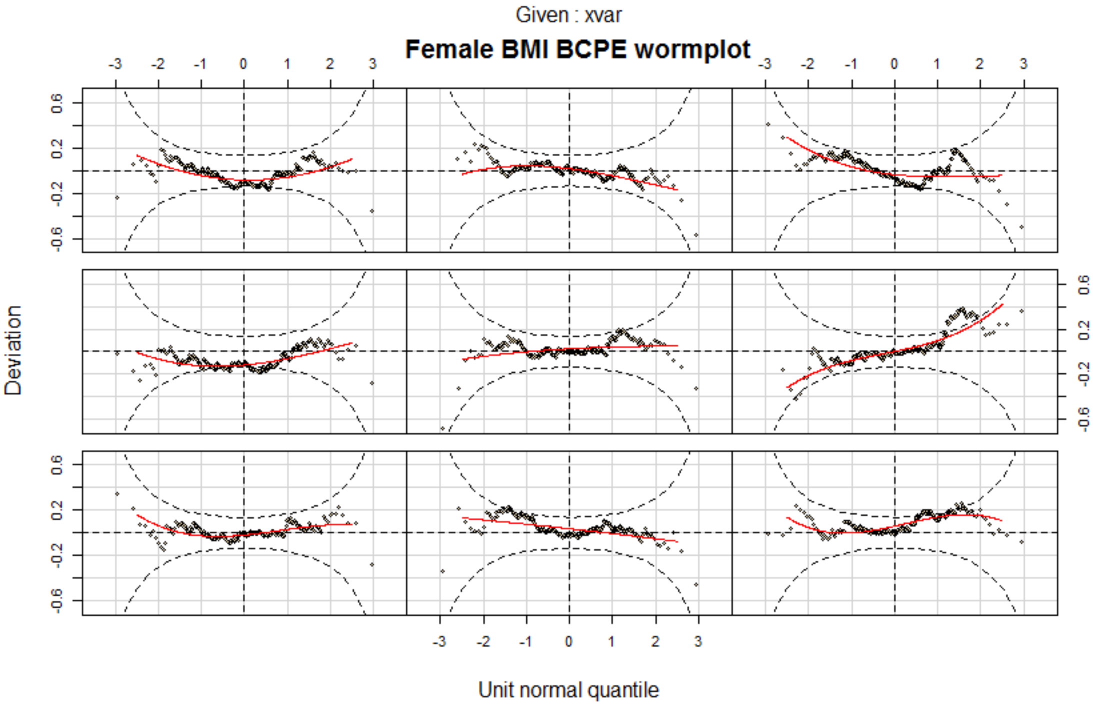


Figure 4.7: Worm plots for Female BMI BCPE model

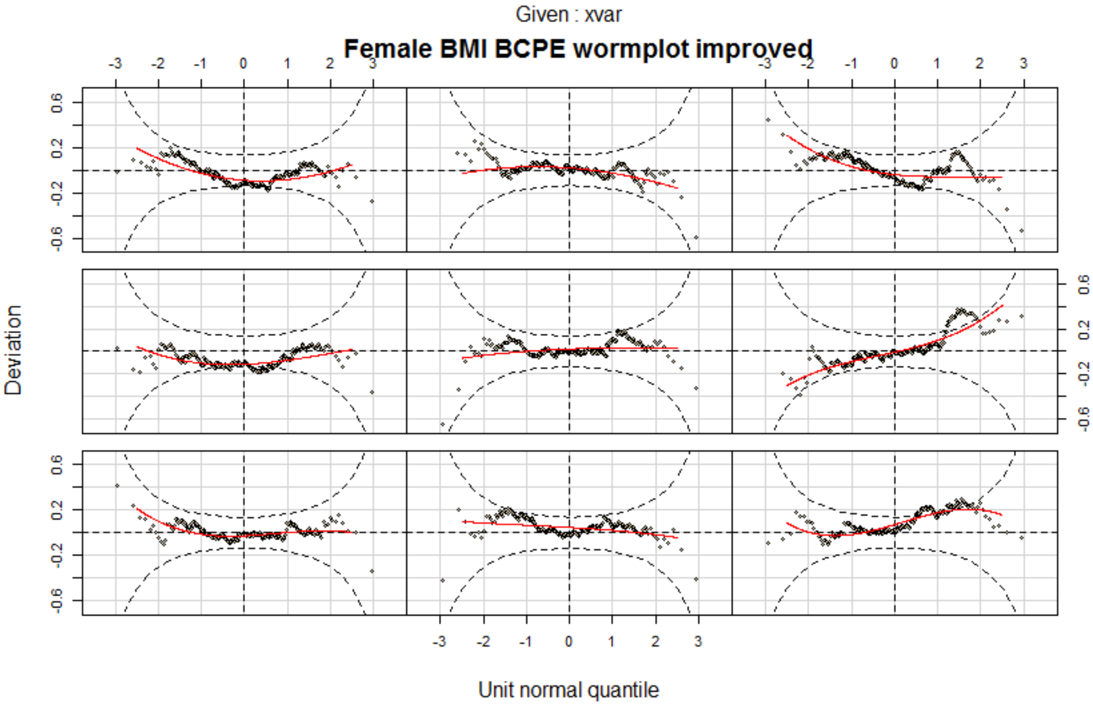


Figure 4.8: Worm plots for Female BMI improved BCPE model

The S patterns are very visible in the male triceps LMS model worm plots (Figure 4.9) and they are vastly improved by the BCPE model (Figure 4.10). This model fits reasonably well with the exception of the panels 1 and 2 where the worm falls below the origin, indicating an incorrectly fitted mean in this region. Increasing the edf of the μ parameter did not correct this problem so we will use the model originally supplied by the automatic model selection technique BCPE(4,4,4,5).

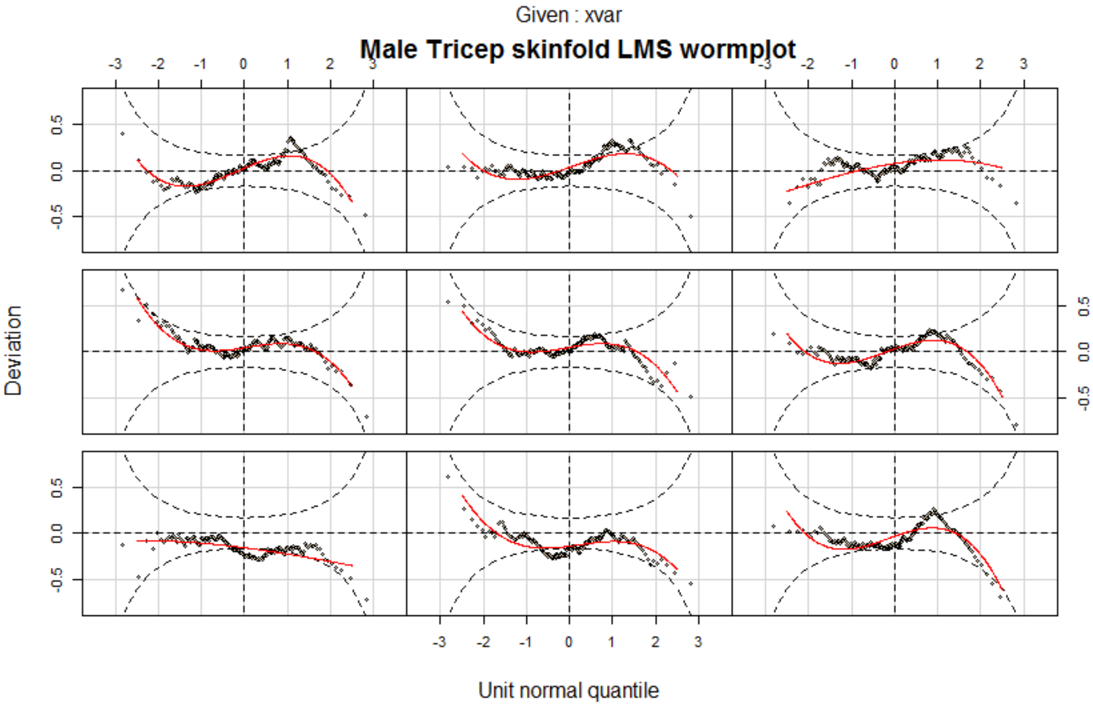


Figure 4.9: Worm plots for Male triceps LMS model

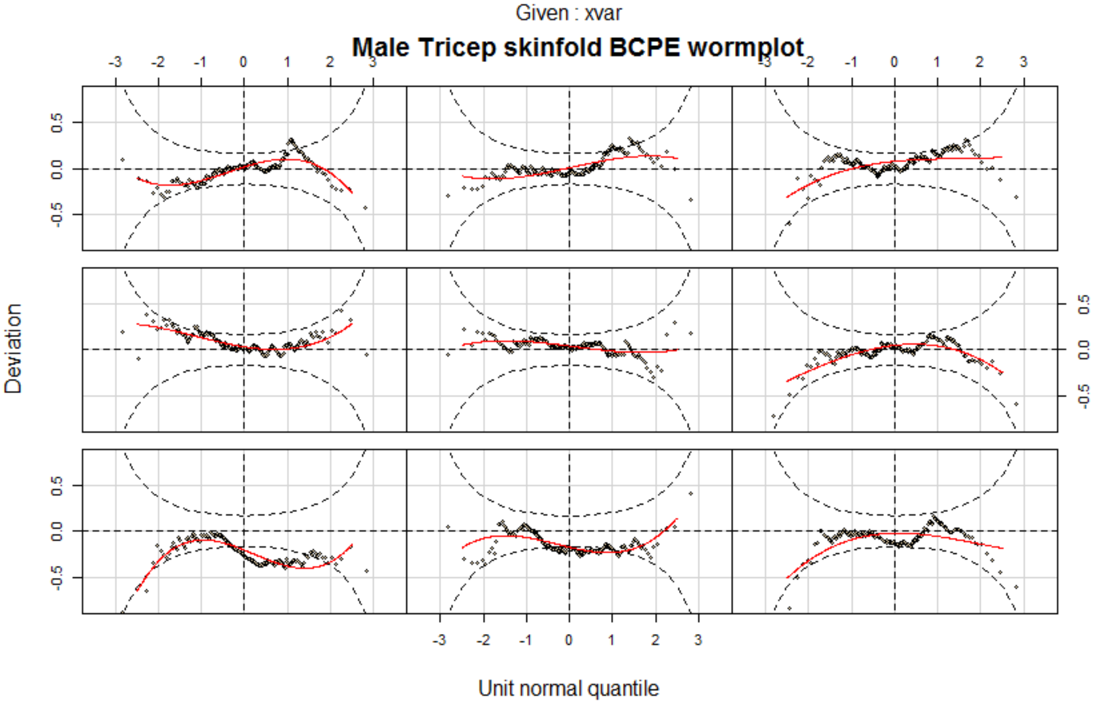


Figure 4.10: Worm plots for Male triceps BCPE model

In contrast to the other gender and variable combination the BCPE model (Figure 4.12) offers only mild improvement over the LMS model (Figure 4.11) for female triceps skinfold. Some of the S shape that is present in panels 1 and 2 of the LMS model worm plots is reduced in the BCPE model (Figure 4.12). However, this small improvement is offset by worsening of the U shape in panel 4. For the ease of comparison to the other models we choose to use the BCPE model. The automatic model selection technique chose to only have an edf of 1 for the ν parameter. Increasing the edf to 4 improves the fit but doesnt remove the problem entirely. The final model chosen for female triceps skinfolds is BCPE(4,4,4,4) (Figure 4.13).

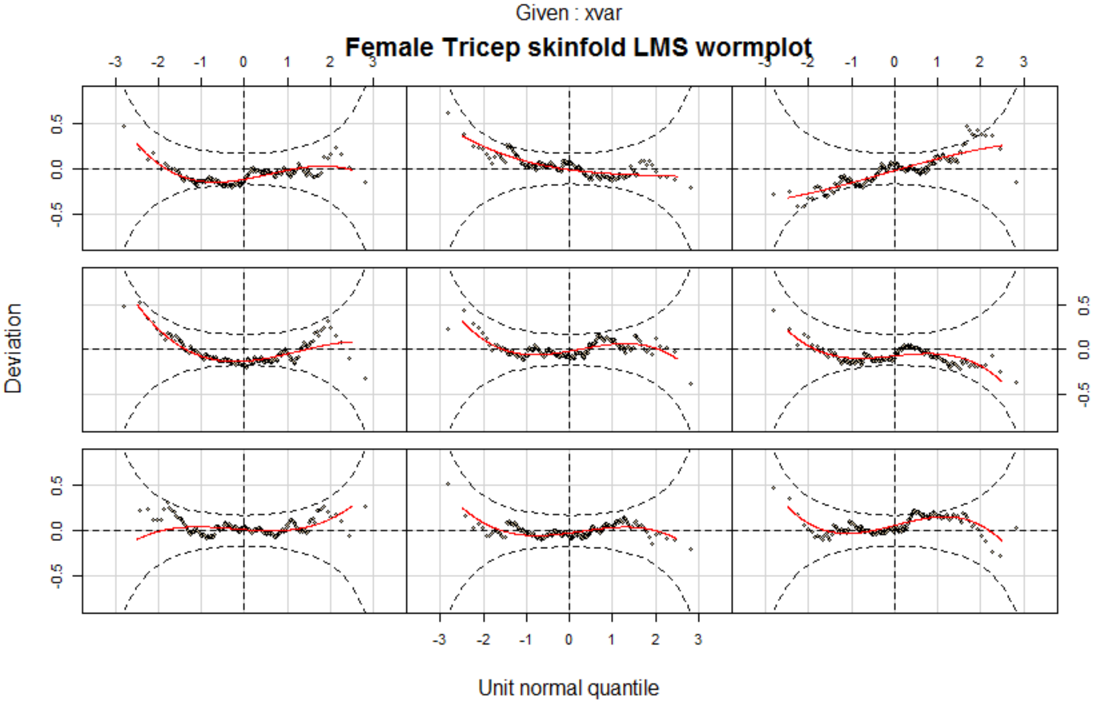


Figure 4.11: Worm plots for Female triceps LMS model

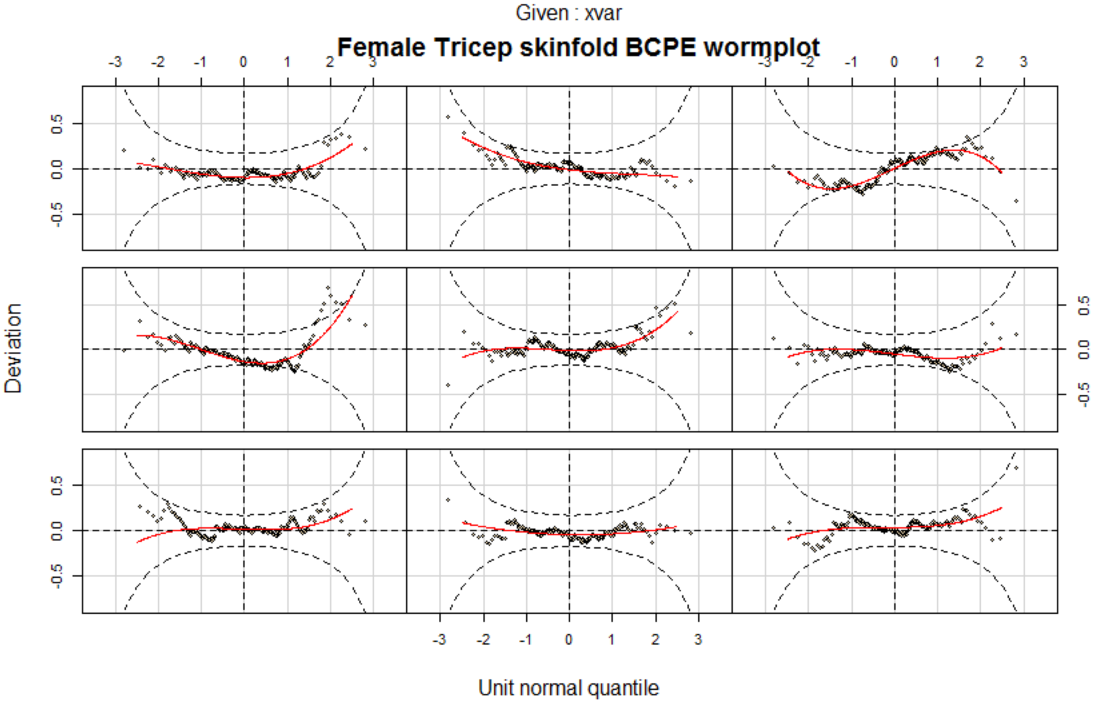


Figure 4.12: Worm plots for Female triceps BCPE model

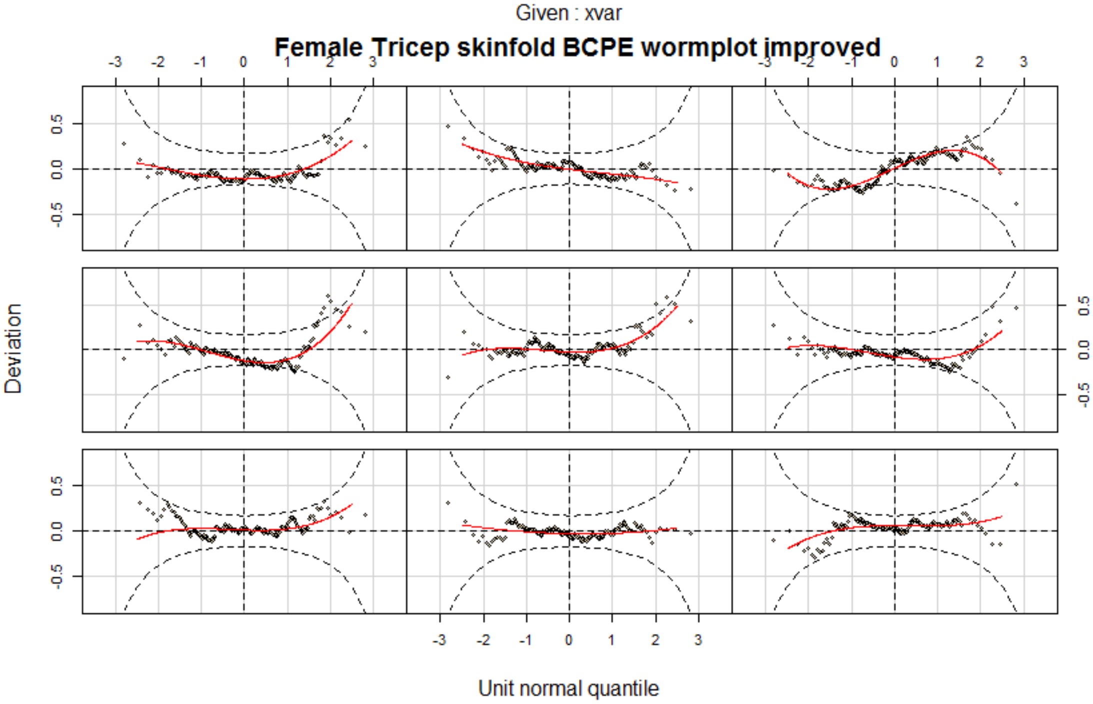


Figure 4.13: Worm plots for Female triceps improved BCPE model

The worm plots for male and female BMI and triceps skinfold show that the quantile regression models fit the data extremely well. Figures 4.14, 4.15, 4.16, and 4.17 show that all the worms lie flat and pass through the origin. Almost no points fall outside the confidence regions of any of the panels. This indicates the possibility of overfitting, as we would expect some points to fall outside these regions. For this reason, we remove extraneous knots from the knot vector used to construct the models, while requiring that the model remains a good fit. This will make the model more parsimonious and reduce undulations in the percentile curves.

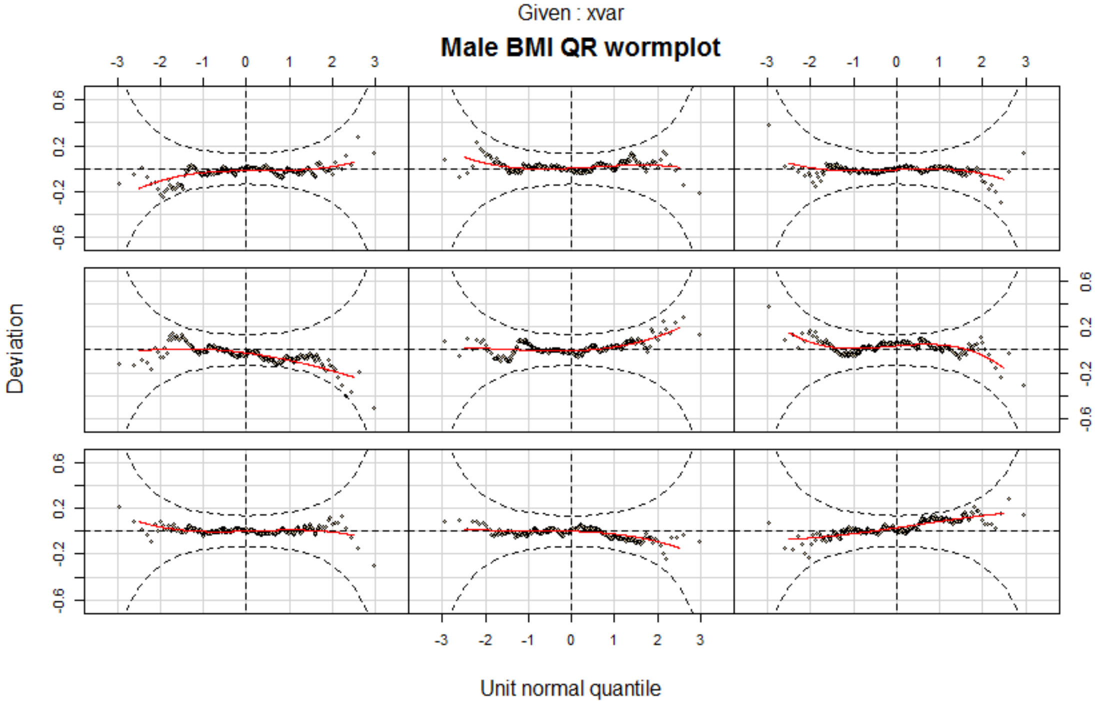


Figure 4.14: Male BMI QR worm plots

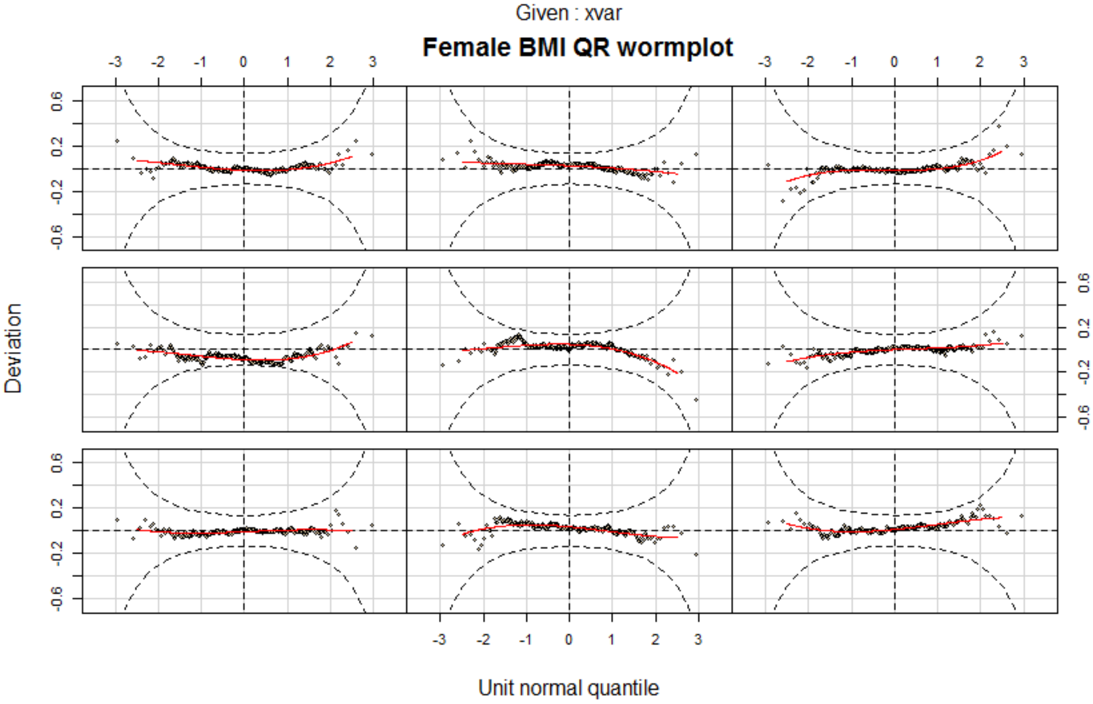


Figure 4.15: Female BMI QR worm plots

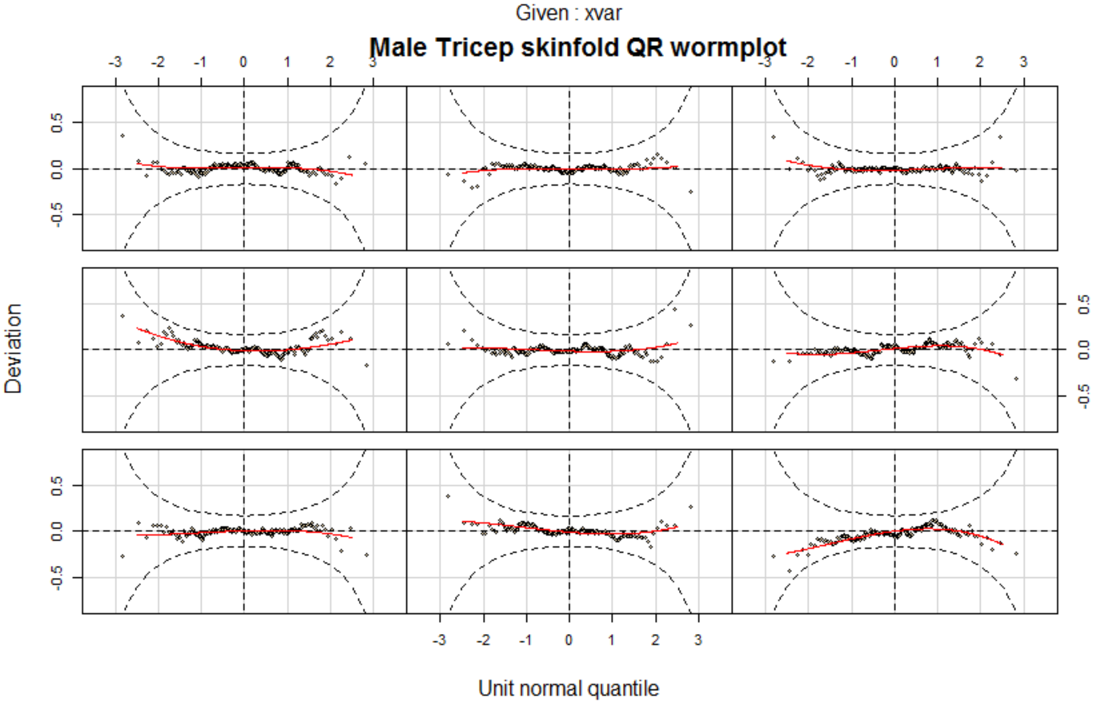


Figure 4.16: Male triceps QR worm plots

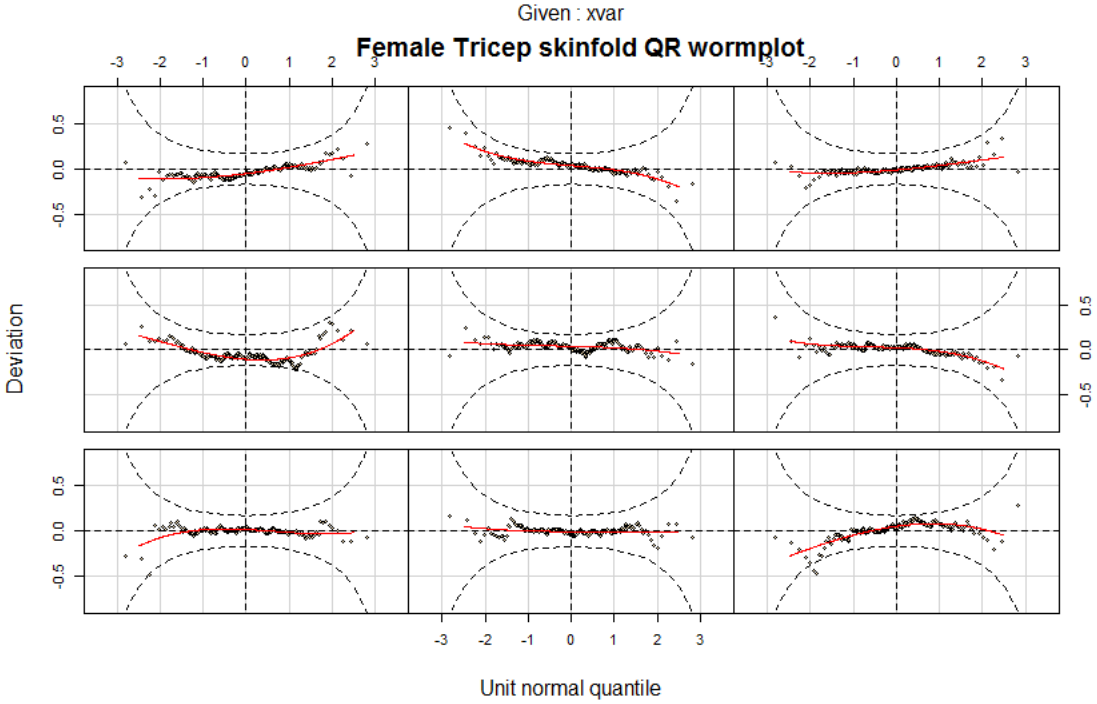


Figure 4.17: Female triceps QR worm plots

The male BMI model constructed using the quantile regression automatic model selection chooses knots at 10, 11, and 17. Dropping the knot at 11 does not cause any major problems to appear in the worm plots but gives the residuals more of the appearance that we would expect of a model that is not overfitted (Figure 4.18). Based on this, the interior knots used for the final quantile regression model for male BMI were reduced to (10, 17). A large number of knots were chosen for female BMI and the resulting plots exhibit some fluctuations. Removing the knots at 11 and 17 years does not dramatically degrade the fit of the model, but attempting to remove any more knots does, causing a serious model violation in the region surrounding that knot. This indicates that female BMI is more variable than male BMI and requires a more rapidly changing model to accurately fit the data. The final model chosen has interior knots at (10,13,14,16,18), (Figure 4.19).

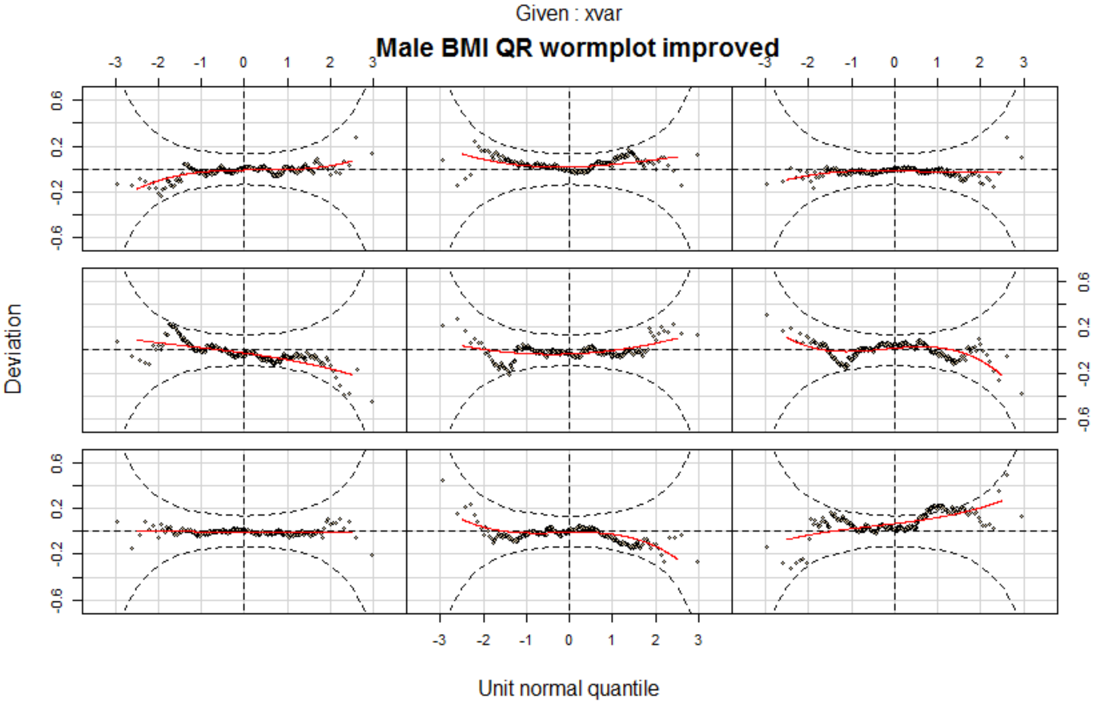


Figure 4.18: Male BMI QR improved worm plots

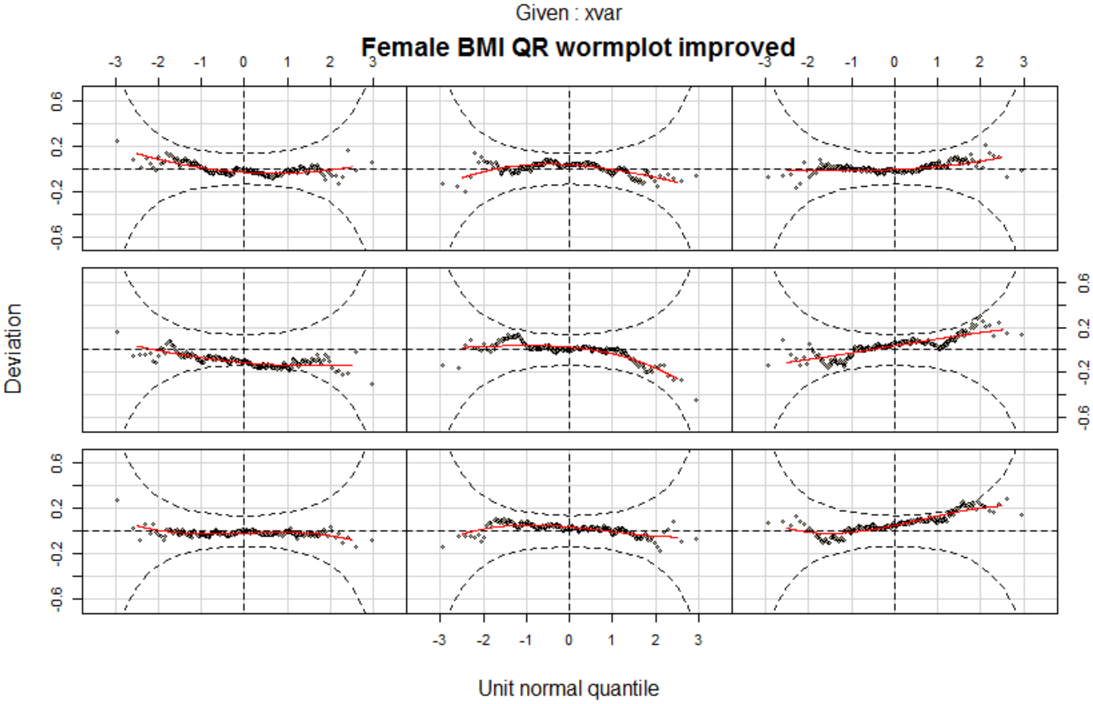


Figure 4.19: Female BMI QR improved worm plots

Male triceps skinfold worm plots for the quantile regression model behave similarly to those for male BMI. Extraneous knots at age 8 and 17 can be removed which don't cause any patterns to appear in the worm plots, but do increase the spread of the data points to what a model that is not overfitted should look like (Figure 4.20). Female triceps skinfolds is the only quantile regression model where the worm plot displays a noticeable problem with model fit. Panels 3 and 4 show an upwards facing U shape, which can be improved by adding a knot at age 10 (Figure 4.21). The final quantile regression model for female triceps skinfolds has knot vector (9, 10, 14, 15).

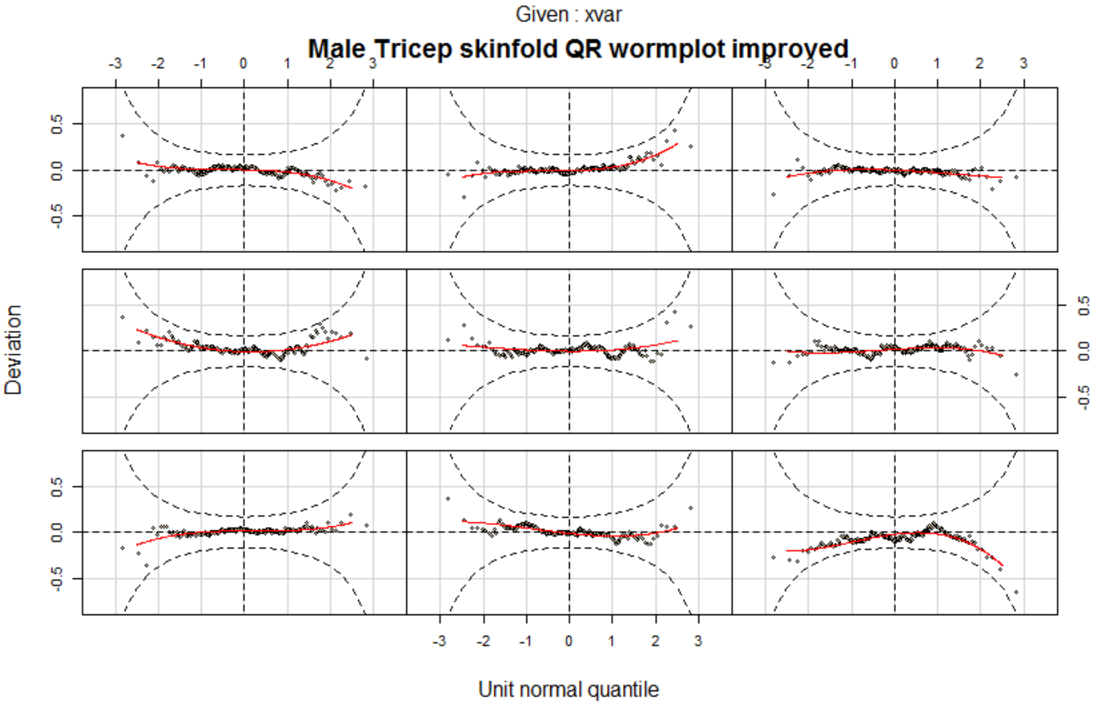


Figure 4.20: Male triceps QR improved worm plots

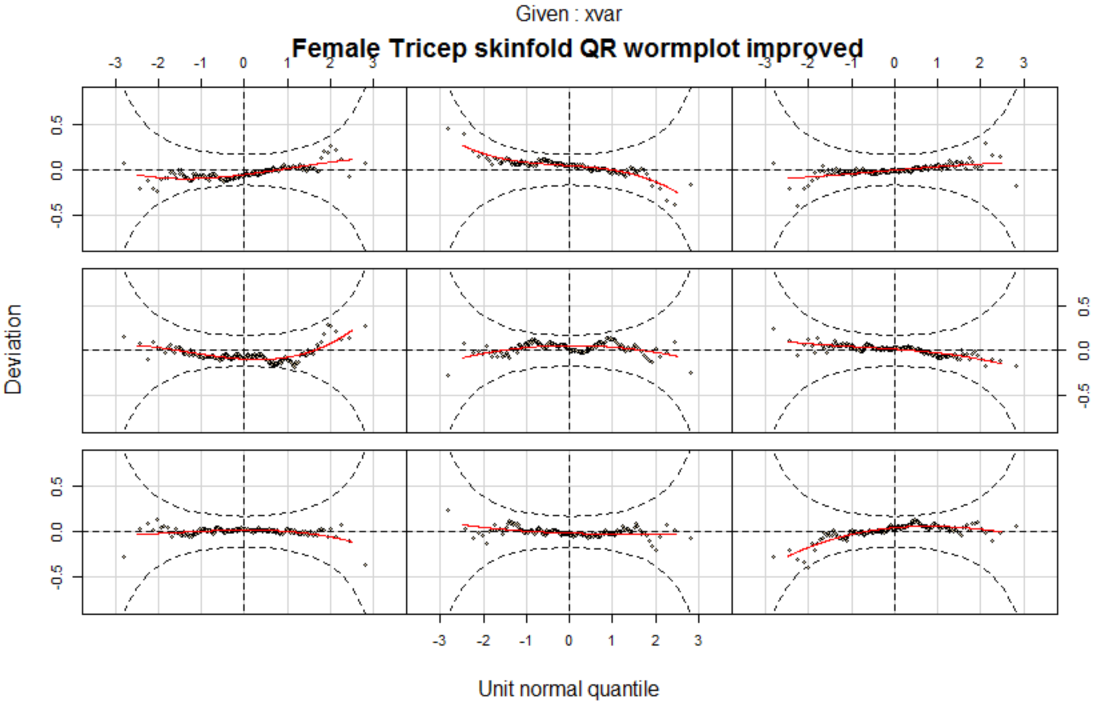


Figure 4.21: Female triceps QR improved worm plots

4.2 Measuring Smoothness

A key aspect of the growth charts produced using the methods discussed in this thesis is the degree to which the centiles are smoothed. Rapid changes in the centiles will make them less useful for clinicians as a small change in age could cause a large change in the centile that a child falls into. Additionally, large or rapid fluctuations may be seen as biologically implausible reducing the chart's usefulness. Deciding what exactly constitutes an appropriate amount of smoothing is rather subjective.

In order to remove fluctuations that may be a result of the fitting process or the randomness of the sample data and not a reflection of the underlying biological processes, researchers may add additional smoothing to the models selected by automatic model selection techniques (World Health Organization, 2008). In the GAMLSS methods this is done by lowering the edf's of the parameters, and in quantile regression by removing interior knots from the basis splines. The amount of additional smoothing deemed appropriate will vary from researcher to researcher, adding possible bias to the model.

In an effort to quantify smoothness of a particular graph, a measure of smoothness is introduced that allows the comparison of models fitted with different fitting methods, and even models fitted to different data sets. Having a objective measure of smoothness allows the researcher to select the model with the best statistical properties from among models that produce graphs with a similar or perhaps a minimum acceptable level of smoothness.

The measure used is the integral of the squared second derivative of the centile C_τ

$$w(C_\tau) = \int^x (C_\tau'')^2 dx \; .$$

Centile curves with more fluctuations will have a larger second derivative, and the integral of the square converts this to a single positive constant that can be used for comparison. The derivatives and integral were preformed numerically. Anthropometric measures tend to have a biologically defined lower limit so lower centiles tend to be relatively smooth compared to higher centiles.

This measure could be used by researchers to compare the smoothness of the graphs produced from their own study to that of previous published works, possibly performed on different populations. However, the range of ages and possible values of the outcome might be different between studies which leads to different values of the smoothness measure even for graphs with a similar level of apparent smoothness. A scaled version of the smoothness measure is more appropriate for comparisons of data created from different data sets

$$\tilde{w}(C_\tau) = \frac{(x_s)^3}{(y_s)^2} w(C_\tau)$$

where x_s and y_s are the range of observed x (age) and y (outcome) respectively. Multiplying by this constant is equivalent to rescaling the fitted curves to the unit square and recalculating the smoothness measure.

To show that this multiplicative relationship holds regardless of how the centiles are constructed, consider the scaled smoothness measure $\tilde{w}$ on a curve $\tilde{f}(\tilde{x})$, where $\tilde{x}$ is the data rescaled from $x \in (a = x_{min}, b = x_{max})$ to $\tilde{x} \in (0, 1)$ on the x-axis and $\tilde{f}(\tilde{x})$ is the curve fitted to that data. The relationship between the original and rescaled curves and data is $f(x) = \tilde{f}(\tilde{x})$ where

$$\tilde{x} = \frac{x - a}{b - a} \quad and \quad x = a + (b - a)\tilde{x}.$$

The second derivative of the curve fitted to the rescaled data is then

$$\begin{aligned} \tilde{f}''(\tilde{x}) = \frac{d^2 \tilde{f}}{d\tilde{x}^2} &= \frac{d}{d\tilde{x}} \left(\frac{dx}{d\tilde{x}} \frac{df}{dx} \right) \\ &= \frac{d}{d\tilde{x}} \left(f'(x)(b - a) \right) \\ &= (b - a) \left(\frac{df'(x)}{dx} \frac{dx}{d\tilde{x}} \right) \\ &= (b - a)^2 f''(x). \end{aligned}$$

Thus the scaled smoothness measure is

$$\begin{aligned}\tilde{w} &= \int_0^1 \left(\tilde{f}''(\tilde{x})\right)^2 d\tilde{x} &= \int\limits_a^b (b-a)^4 \left(f''(x)\right)^2 \frac{dx}{(b-a)} \\ & &= (b-a)^3 \int\limits_a^b \left(f''(x)\right)^2 dx \\ & &= (b-a)^3 w.\end{aligned}$$

Similarly, if we scale the data in the y direction from the range $y \in (0, y_{max})$ to $\tilde{y} \in (0,1)$ with the relations $\tilde{y} = y/y_{max}$ and $\tilde{f} = f/y_{max}$ the second derivative of the curve fitted to the rescaled data will be

$$\frac{d^2\tilde{f}}{dx^2} = \frac{1}{y_{max}} \frac{d^2 f}{dx^2}.$$

The smoothness measure of the curve fitted to the data that has been rescaled in the y direction is

$$\begin{aligned}\tilde{w} &= \int\limits_a^b \left(\tilde{f}''(x)\right)^2 dx \\ &= \frac{1}{y_{max}^2} \int\limits_a^b \left(f''(x)\right)^2 dx \\ &= \frac{w}{y_{max}^2}\end{aligned}$$

4.2.1 Centiles for the Final Models and their Smoothness Measures

Figures 4.22-4.29 show the fitted centiles for each final model. The BCPE models feature smooth gradually changing lines while the quantile regression plots feature much more rapid changes in the centile curves. The exception here is male BMI fitted using quantile regression, where the centile curves look very similar to the smooth curves fitted using BCPE and the smoothness measure reflects this with the results being of approximately the same magnitude.

Table 4.2 contains the smoothness measure for each of the final models, computed for each centile. An immediate observation is that the measure is typically and approximately an order of magnitude larger for quantile regression models than for BCPE models. This reflects what might be intuitively expected because the GAMLSS models are fitted with an explicit smoothness penalty, while the quantile regression models are not. Quantile regresion uses a penalty that is based on the number of parameters available to fit the curves, essentially a penalty on how much the curves might potentially fluctuate, not on how the fitted curves actually fluctuate, as is the case in the parametric models. The result is that quantile regression seems to produce models that are insufficiently smoothed.

In all the anthropometric measure, sex, and model combinations, the 90th and 95th centiles have the highest smoothness measures. The highest centiles in each model tend to be more variable so this is expected. An interesting difference between BCPE models and quantile regression models is that for the BCPE models the smoothness measure for the 95th centile is higher than the 90th for each fitted model. This is not the case for the quantile regression models. For male BMI and female triceps skinfold using quantile regression models, the smoothness measure is slightly lower for the 95th centile. This is possible because the quantile regression curves are fitted more independently of each other. If a bend is fitted in a given centile a similar bend need not necessarily be fitted in an adjacent centile. BCPE and LMS models, however, fit all the centile curves simultaneously because the value of all centiles for a given age are based on the same conditional distribution, which is non-zero for all positive values of the outcome. If the distribution is made more skewed or more variable to accommodate the data at a given age, this skewness or variation will be amplified in the extreme centiles.

Sex	Variable	Model type	5	10	25	50	75	90	95
Male	BMI	BCPE	0.189	0.135	0.075	0.033	0.157	0.858	2.679
Female	BMI	BCPE	0.193	0.209	0.245	0.257	0.326	0.806	1.887
Male	Triceps	BCPE	0.171	0.204	0.420	1.125	3.146	5.447	7.007
Female	Triceps	BCPE	0.655	0.547	0.367	0.295	0.839	2.332	4.227
Male	BMI	QR	3.803	0.631	0.970	0.713	1.616	7.543	7.335
Female	BMI	QR	3.417	7.132	1.914	15.357	38.274	306.657	629.732
Male	Triceps	QR	63.832	27.174	11.331	35.581	124.262	533.675	680.424
Female	Triceps	QR	3.011	3.558	5.836	5.196	7.423	57.512	50.748

Table 4.2: Smoothness measure for each centile

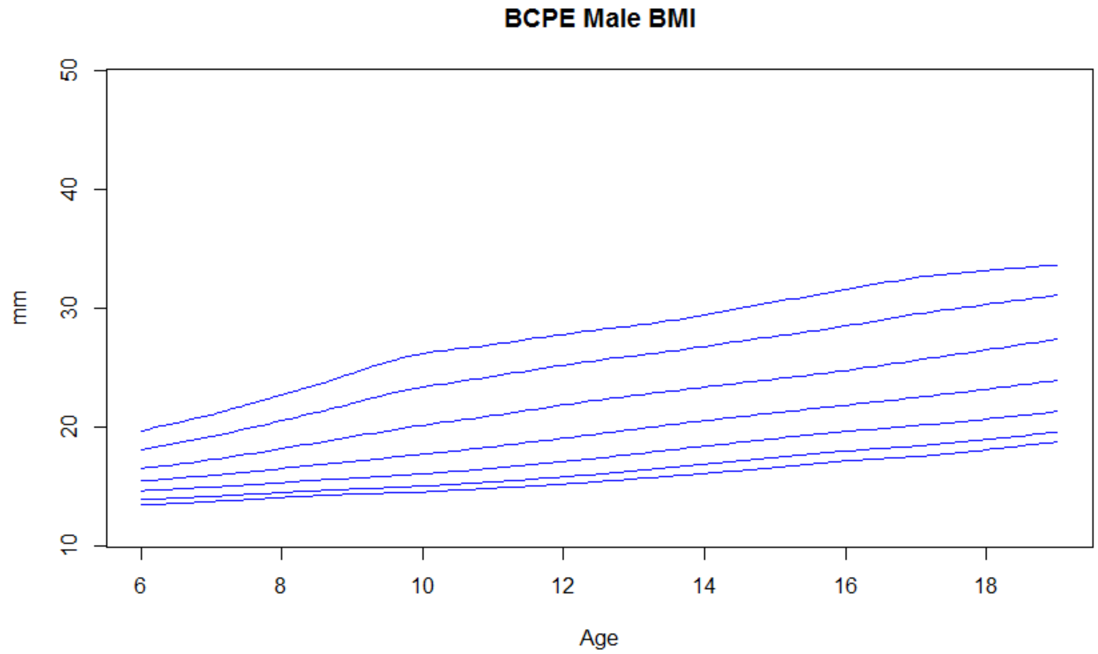


Figure 4.22: Male BMI BCPE final model

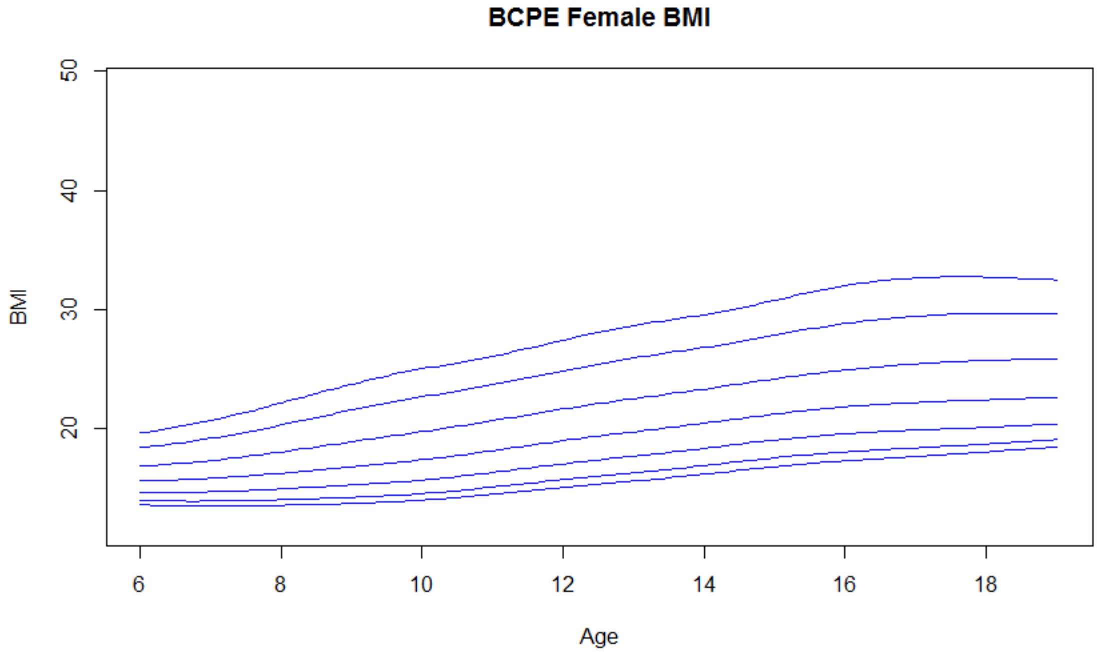


Figure 4.23: Female BMI BCPE final model

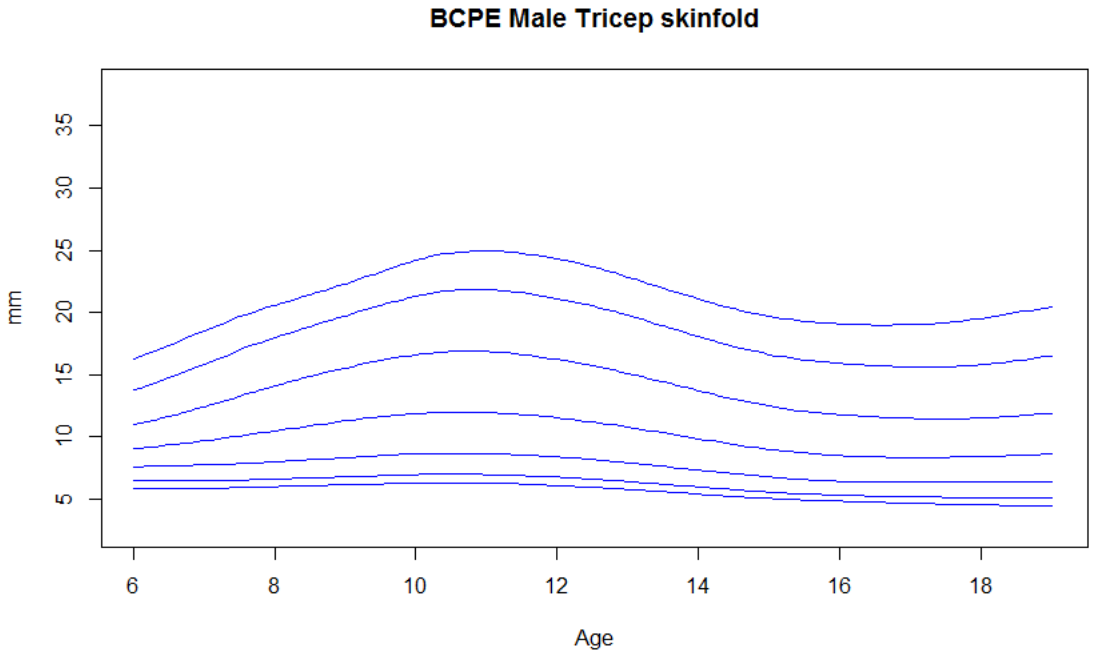


Figure 4.24: Male triceps BCPE final model

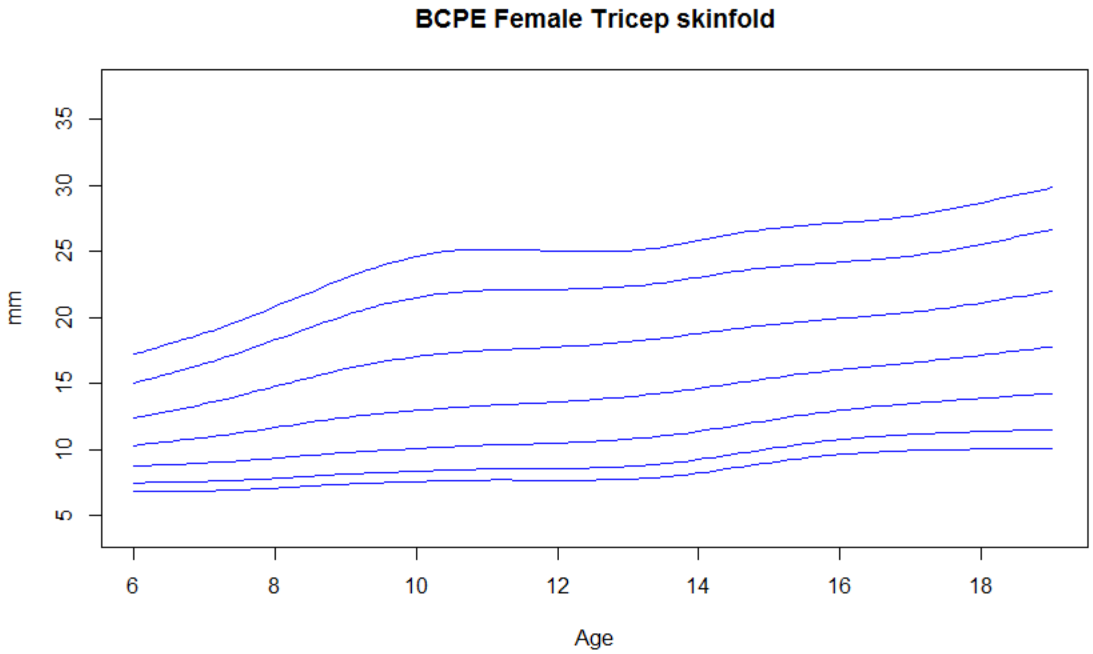


Figure 4.25: Female triceps BCPE final model

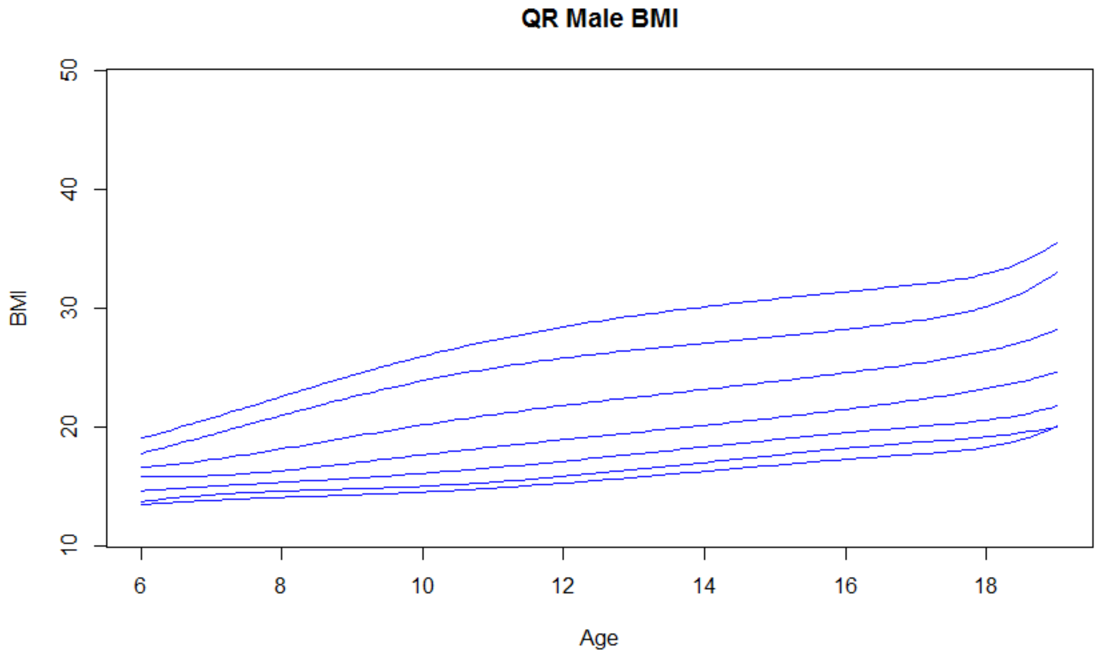


Figure 4.26: Male BMI QR final model

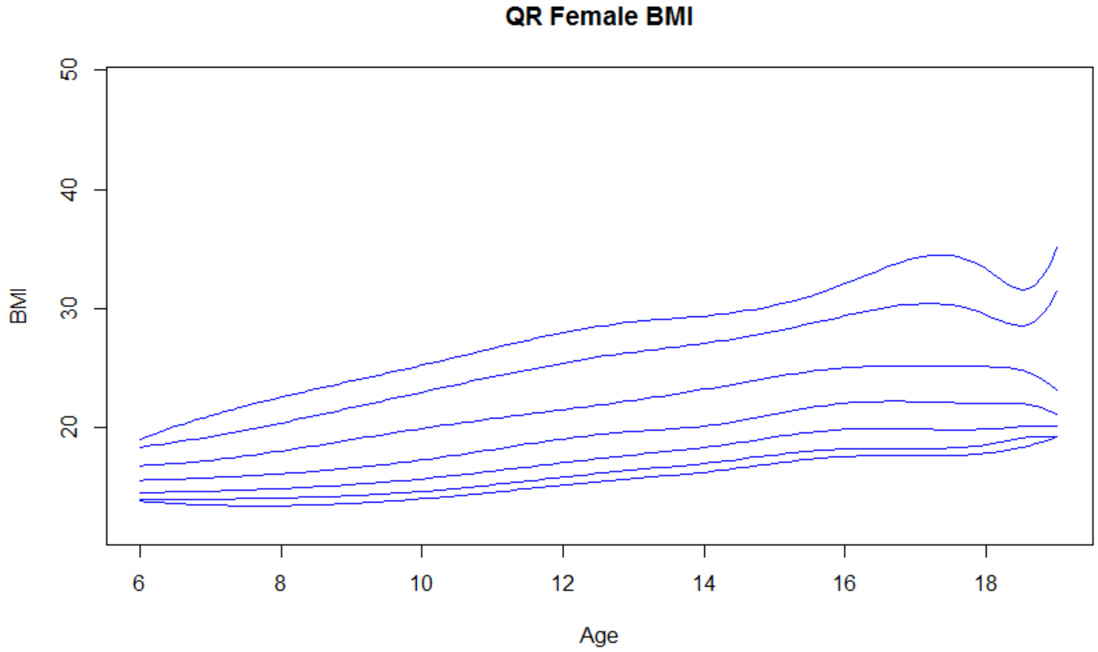


Figure 4.27: Female BMI QR final model

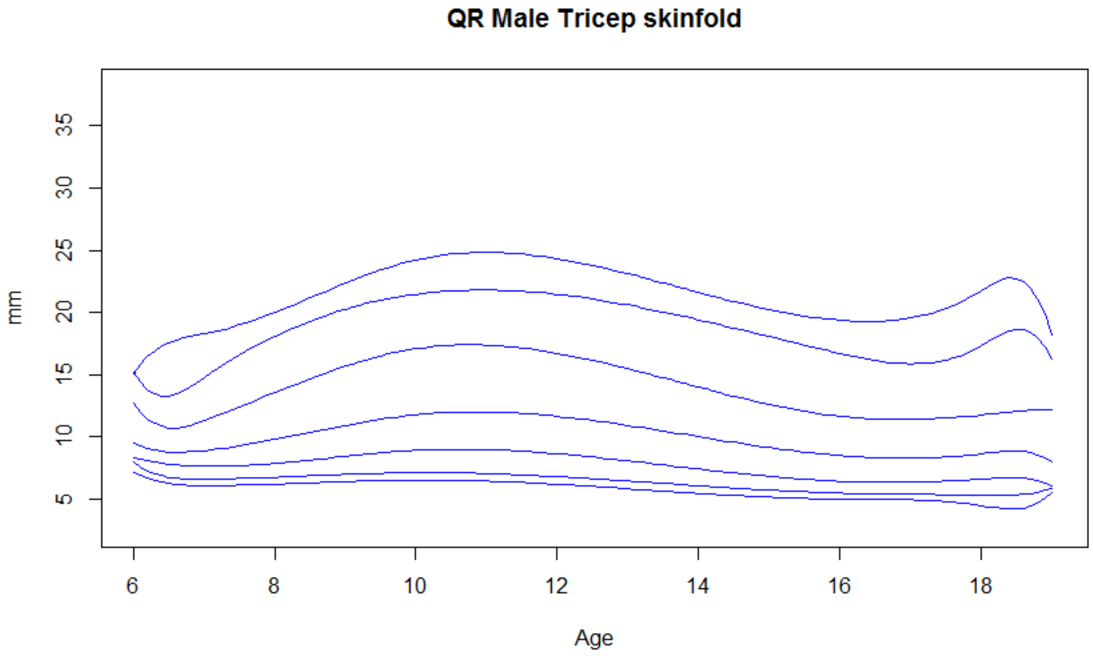


Figure 4.28: Male triceps QR final model

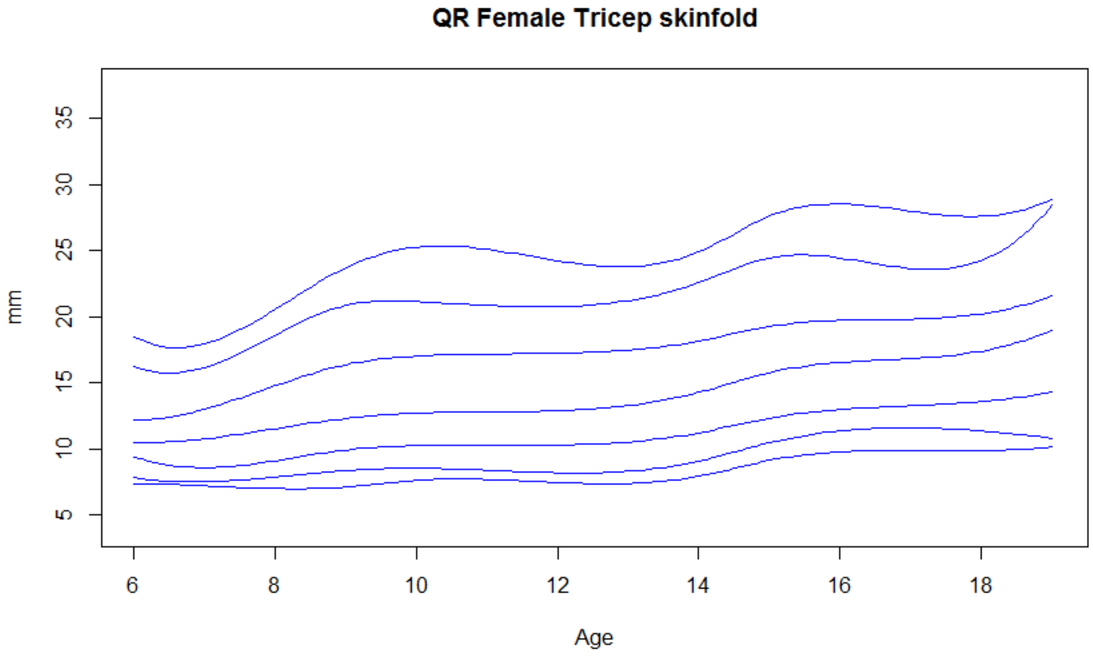


Figure 4.29: Female triceps QR final model

Chapter 5

Discussion

Quantile regression, LMS, and BCPE models were fitted for BMI and triceps skinfold for both males and females. Quantile regression utilized an exhaustive search procedure to identify optimal knot placement by searching all possible models with integer knot placements. LMS and BCPE models were fitted using a GAIC optimization procedure. These models were improved using the diagnostic worm plots and had their overall smoothness assessed.

Both quantile regression and the GAMLSS methods use these automatic model selection techniques to remove arbitrary choice from the process of fitting centile curves but they are only marginally successful. The models selected by the automatic model selection routines are still dependent on a choice of penalty parameter; a larger penalty will tend to produce a smoother graph. This may be preferable to individually choosing the edf's for each parameter in BCPE models, that dictate how much each moment will be allowed to vary, or the knot number and placement of knots in quantile regression that dictates where the curves will be allowed to bend the most, but it is still a conscious choice by the researcher.

The objective of removing choice is further undermined by the diagnostic portion of the model creation process. While the worm plots are useful for adjusting the models to obtain a better fit, they are assessed visually. Deciding what constitutes a pattern in the shape of the worms and what is just noise is left to the user, as is deciding if the adjustments made adequately solved the problem. While guidelines exist for how these procedures should be implemented, a more algorithmic method might remove some of the possible bias introduced by the user preference.

A final and perhaps overlooked choice made by the user is the assumption about the underlying distribution (this only applies for GAMLSS models). LMS and BCPE methods assume that the data are normally or BCPE distributed after an appropriate transformation. These distributions are not the only ones that can be used; the

GAMLSS package for **R** (Stasinopoulos et al., 2015) contains dozens of options that can be implemented and the choice of which can have a fundamental impact on the shape of the curves.

Nonparametric methods like quantile regression can model any distributional shape. Wei et al. (2004) showed how quantile regression could be used to construct curves for a population that was bimodal, something that could not be accurately represented with the unimodal distributions of LMS and BCPE. The population in question in Wei et al. (2004) was actually the result of two populations being combined, and the diagnostic plots created for this thesis did not reveal any evidence of a lack of fit but this is a possibility in other datasets and should always be considered.

Quantile regression may seem attractive because of the lack of assumed parametric distribution but it has some drawbacks compared to GAMLSS methods that are difficult to justify. The construction of GAMLSS methods inherently provides information about how the moments of the distribution change over time by examining the parameter curves. Understanding how the variance or skewness of a populations distribution change over time may be of interest to researchers. However, this information is obtained by visually inspecting the graphs and examining the spacing of the lines and it is therefore potentially very unreliable. If GAMLSS parameter curves are published along with the centile curves, other researchers can construct exact centile curves for any τ they desire. We chose to construct our centile curves for values of τ corresponding to round numbered percentiles (5%, 10%, 25%, 50%, 75%, 90%, 95%), but some organizations (World Health Organization, 2008) publish centile growth curves for children based on evenly spaced values of z_α. Comparing results from a GAMLSS model requires only construction of the centiles for the required value of τ whereas quantile regression model would require interpolating the desired centiles from the adjacent published centiles. Additionally, growth curves constructed using quantile regression provide no information about the shape of the distribution (or growth curves) for centiles higher than the highest published centile curve. For example, in our case using the quantile regression growth curves we constructed, one could say nothing about the 99th centile, because the 95th is the highest one we constructed.

Relatedly, GAMLSS growth charts are also easier for other researchers to use with their own data. For given observations, a z-score can be calculated using the parameters provided by other researchers. Quantile regression models require the user to approximate a centile by interpolating between the two closest centiles. While this may be acceptable in a setting where, it might not be very relevant to distinguish where exactly a point falls between two centiles, but if the variable in question was part of a larger model this procedure may not be acceptable.

5.1 Future Work

Future work in this area could focus on combining the strengths of the two methods studied in this thesis and related methods to possibly improve some shortcomings. Quantile regression's main attractive feature of no underlying distributional assumption could be used to make the distributional choice of the GAMLSS methods more informed. Quantile regression models could be fitted and cross sectional profiles obtained to illustrate the shape of the conditional profile at any given age. This would allow researchers to confirm if the intended distributional choice to be used with a GAMLSS model is appropriate. For example, if these profiles revealed that the population was bimodal, the LMS and BCPE methods outlined in this thesis would not be a good choice because they are unimodal distributions.

The smoothness measure developed in Chapter 4 can be used by other researchers to compare curves to determine if they are smoothing the curves they construct to a similar degree. The World Health Orginization (World Health Organization, 2008) and the Center for Disease Control (Grummer-Strawn M et al., 2010) both have their own published set of growth curves and it would be interesting to study how the results compared. Both the WHO and the CDC use the LMS method but the process they used for selecting the smoothing parameters are different.

A method combining a explicit smoothness penalty within the quantile regression procedure is describe in Koenker et al. (1994). The method uses the objective function

$$\sum_{i=1}^{n} \rho_\tau \left(y_i - g(t_i) \right) + \lambda_k \int |g^{''}(x)|^p dx.$$

When $p = 1$, minimizing this objective function remains a linear programming problem which can be solved very quickly. However, the optimal function $g(t)$ is a linear spline (Koenker et al., 1994). This leads to centile “curves” that are rather jagged and unsuitable for the purposes discussed in this thesis. Given the access to computing power that is now available, a similar method that incorporates a quadratic or higher order penalty similar to that used in the GAMLSS method could perhaps produce more useful results and is worth future investigation.

Appendices

Appendix A

Splines

A.1 Splines

The GAMLSS method outlined in Chapter 3 utilizes cubic smoothing splines that differ from the B-splines used in Chapter 2. The derivation of these splines is based on the work by Pollock (1999). Interpolating splines are presented first, then cubic smoothing splines are described.

A.1.1 Interpolating Spline

The explanation of how interpolating splines may be explicitly derived as presented in (Pollock, 1999) is now shown below. Given an ordered set of coordinates (x_0, y_0) $\ldots (x_n, y_n)$ we wish to create a piecewise function for interpolating the points between them. Rather than the jagged form of straight lines linking each point, a cubic polynomial with the requirements that at each point the first and second derivatives be continuous provides a smooth curve with no breaks or jumps. Let f_i denote the function over interval i, $[x_{i-1}, x_i]$, $i = 1, \ldots, n$, then these requirements are

$$f_{i-1}(x_i) = f_i(x_i) = y_i \tag{A.1}$$

$$f_{i-1}^{'}(x_i) = f_i^{'}(x_i) \tag{A.2}$$

and

$$f_{i-1}^{''}(x_i) = f_i^{''}(x_i). \tag{A.3}$$

We also need to define the behavior of the curve at the endpoints. The often-used natural cubic spline is linear at the endpoints, leading to the constraints

$$f_0^{''}(x_0) = 2b_0 = 0$$

and

$$f_{n-1}^{''}(x_n) = 2b_n = 0.$$

In the interval $[x_i, x_{i+1}]$ each piecewise polynomial can be expressed as

$$f_i(x_i) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i, \tag{A.4}$$

so its first and second derivatives are

$$f_i^{'}(x_i) = 3a_i(x - x_i)^2 + 2b_i(x - x_i) + c_i \tag{A.5}$$

and

$$f_i^{''}(x_i) = 6a_i(x - x_i) + 2b_i. \tag{A.6}$$

Combining these expressions with the constraints above allows the parameters of one polynomial segment to be defined in terms of the next. Using (6.1) and (6.4) we get

$$a_{i-1}h_i^3 + b_{i-1}h_i^2 + c_{i-1}h_i + d_{i-1} = d_i = y_i$$

where $h_i = x_i - x_{i-1}$. Similarly, using 6.2 and 6.5 gives

$$3a_{i-1}h_i^2 + 2b_{i-1}h_i + c_{i-1} = c_i$$

and 6.3 and 6.6,

$$6a_{i-1}h_i + 2b_{i-1} = b_i.$$

Equation 6.6 also implies

$$a_ih_i^3 + b_ih_i^2 + c_ih_i + d_i = y_{i+1}$$

which can be solved for c_i using that $d_i = y_i$, to give

$$c_i = \frac{y_{i+1} - y_i}{h_i} - a_i h_i^2 - b_i h_i.$$

The constant a_i, can be found by solving $6a_i h_i + 2b_i = b_{i+1}$, to get

$$a_i = \frac{b_{i+1} - b_i}{3h_i}.$$

This allows us to express c_i in terms of b_i's

$$c_i = \frac{y_{i+1} - y_i}{h_i} - \frac{1}{3}\left(b_{i+1} + 2b_i\right) h_i.$$

Now we have expressed all of our parameters in terms of our data points y_i and the parameters b_i and b_{i+1}.

Using these equations to rewrite the first order continuity requirements $f^{'}_{i-1}(x_i) = f^{'}_i(x_i)$ gives

$$b_{i-1}h_{i-1} + 2b_i(h_{i-1} + h_i) + b_{i+1}h_i = \frac{3}{h_i}(y_{i+1} + y_i) - \frac{3}{h_{i+1}}(y_i - y_{i-1}).$$

Letting i vary from 1 to $n-1$ yields a system of equations that can be reduced to a bi-diagonal system

$$\begin{bmatrix} p_1' & h_1 & 0 & ... & 0 & 0 \\ 0 & p_2' & h_2 & ... & 0 & 0 \\ 0 & 0 & p_3' & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & p_{n-2}' & h_{n-2} \\ 0 & 0 & 0 & ... & 0 & p_{n-1}' \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ ... \\ b_{n-2} \\ b_{n-1} \end{bmatrix} = \begin{bmatrix} q_1' \\ q_2' \\ q_3' \\ ... \\ q_{n-2}' \\ q_{n-1}' \end{bmatrix}$$

where

$p_i = 2(h_{i-1} + h_i) = 2(x_{i+1} - x_{i-1})$ and $q_i = \frac{3}{h_i}(y_{i+1} - y_i) - \frac{3}{h_{i-1}}(y_i - y_{i-1})$.

This system can be solved by back substitution for the values $b_1, \ldots, b_{n-1}$. Using these values and previous equations, the values $a_0, \ldots, a_{n-1}$ can be obtained as well as the value c_0. The values $c_1, \ldots, c_{n-1}$ are generated recursively using

$$c_i = (b_i + b_{i-1})h_{i-1} + c_{i-1}.$$

We have now defined a_i, b_i, c_i, d_i for all i and fully defined our interpolation spline.

A.1.2 Cubic Smoothing Splines

Pollock (1999) extends the methods used to derive interpolating splines to cubic smoothing splines that do not intersect every data point. If we imagine that the data follows a underlying function $f(x)$ with some random variability we can express each data point as the relationship $y_i = f(x_i) + \varepsilon_i$ for $i = 1, \ldots, n$ where ε_i is a independent random variable with variance $\mathrm{V}(\varepsilon_i) = \sigma_i^2$. Fitting a cubic spline strikes a balance between an exact fit to the data and a smooth function that filters out noise to reveal the underlying function. The function $f(x)$ is approximated by the spline function $S(x)$, which minimizes the function

$$L = \lambda \sum_{i=0}^{n} \left(\frac{y_i - S_i}{\sigma_i} \right)^2 + (1 - \lambda) \int\limits_{x_0}^{x_n} \left(S''\left(x_i\right)^2 \right) dx.$$

Here we can see that the two extreme cases of λ =1 and $\lambda = 0$ are illustrative of the purpose of each component of L. When λ =1, L is minimized by an piecewise polynomial that perfectly intersects each data point (an interpolating spline). When $\lambda = 0$

$$L = \int\limits_{x_0}^{x_n} \left(S''\left(x_i\right)^2 \right) dx.$$

If $S(x_i)$ is linear $S''(x_i) = 0$ and $L = 0$. Thus in this extreme with maximum smoothing L is minimized when S becomes a straight line. We choose λ such that we balance smoothness with fealty to the data.

The smoothing term is piecewise and can be rewritten as

$$\sum_{i=0}^{n-1} \int\limits_{x_i}^{x_{i+1}} \left(S_i''(x)^2 \right) dx,$$

where S_i is the component of S between x_i and x_{i+1}. Since each S_i is composed of a cubic function, its second derivative S_i'' is a linear function that takes on the value $2b_i$ at x_i and the value $2b_{i+1}$ at x_{i+1}. This allows the integral of each segment to be written as

$$\int\limits_{x_i}^{x_{i+1}} \left(S_i''(x)\right)^2 = 4\int\limits_{0}^{h_i} \left(b_i\left(1-\frac{x}{h_i}\right)+b_{i+1}\left(\frac{x}{h_i}\right)\right)^2 dx = \frac{4h_i}{3}\left(b_i^2+b_ib_{i+1}+b_{i+1}^2\right),$$

and L to be written as

$$L = \lambda\sum_{i=0}^{n}\left(\frac{y_i-d_i}{\sigma_i}\right)^2 + (1-\lambda)\sum_{i=0}^{n-1}\frac{4h_i}{3}\left(b_i^2+b_ib_{i+1}+b_{i+1}^2\right)$$

where $d_i = S_i(x_i)$. The major difference between a smoothing spline and an interpolating spline is that the ordinates d_i are not given by y_i. Similarly to the interpolating spline, the function $S_i(x_i)$ and its second derivative can be defined in terms of the coefficients of the polynomial a_i, b_i, c_i and d_i,

$S_i(x_i) = d_i$

$S_i(x_{i+1}) = d_{i+1}$

$S_i^{''}(x_i) = 2b_i$

and

$S^{''}(x_{i+1}) = 2b_{i+1}$.

Using the second and fourth condition we can solve for a_i and c_i in terms of b_i, b_{i+1} and d_i, d_{i+1} to get

$$a_i = \frac{b_{i+1}-b_i}{3h_i}$$

and

$$c_i = \frac{d_{i+1}-d_i}{h_i} - \frac{1}{3}\left(b_{i+1}-2b_i\right).$$

Using the condition that the first derivatives are continuous $S_{i-1}'(x_i) = S_i'(x_i)$ gives

$$3a_{i-1}h_{i-1}^2 + 2b_{i-1}h_{i-1} + c_{i-1} = c_i.$$

By replacing the c's and a's with the expressions above we have

$$b_{i-1}h_{i-1} + 2b_{i-1}(h_{i-1}+h_i) + b_{i+1}h_i = \frac{3}{h_i}(d_{i+1}-d_i) - \frac{3}{h_{i-1}}(d_i-d_{i-1}).$$

Summarizing this equation for all i in $1,\ldots,n-1$ and the end cases where $b_0 = b_n = 0$ in matrix form yields

$$\begin{bmatrix} p_1' & h_1 & 0 & ... & 0 & 0 \\ h_1 & p_2' & h_2 & ... & 0 & 0 \\ 0 & h_2 & p_3' & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & p_{n-2}' & h_{n-2} \\ 0 & 0 & 0 & ... & h_{n-2} & p_{n-1}' \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ ... \\ b_{n-2} \\ b_{n-1} \end{bmatrix} = \begin{bmatrix} r_0 & f_1 & r_1 & 0 & ... & 0 & 0 \\ 0 & r_1 & f_2 & r_2 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & r_{n-2} & 0 \\ 0 & 0 & 0 & 0 & ... & f_{n-1} & r_{n-1} \end{bmatrix} \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \\ ... \\ d_{n-1} \\ d_{n-2} \end{bmatrix}$$

where

$$p_i = 2\left(h_{i-1} + h_i\right), \; r_i = \frac{3}{h_i}, \; and \; f_i = -(r_{i-1} + r_i)$$

which can be expressed equivalently as

$$Rb = Q'd.$$

The function L can be expressed in matrix form

$$L = \lambda(y-d)'\Sigma^{-1}(y-d) + \frac{2}{3}(1-\lambda)b'Rb,$$

which, using the the relation $b = R^{-1}Q'd$, allows us to reexpress the function L in terms of only d, the ordinates at the knots

$$L = \lambda(y-d)'\Sigma^{-1}(y-d) + \frac{2}{3}(1-\lambda)d'QR^{-1}Q'd.$$

The optimal value of the of the knot ordinates occurs when L is minimized. To find these values we differentiate with respect to d, giving

$$-2\lambda(y-d)'\Sigma^{-1} + \frac{4}{3}(1-\lambda)d'QR^{-1}Q' = 0$$

which implies that

$$\lambda\Sigma^{-1}(y-d) = \frac{2}{3}(1-\lambda)QR^{-1}Q'd$$

and

$$\lambda\Sigma^{-1}(y-d) = \frac{2}{3}(1-\lambda)Qb.$$

Premultiplying by $\lambda^{-1}Q'\Sigma$ and rearranging gives

$$(\mu Q'\Sigma Q + R)b = Q'y$$

where $\mu = 2(1-\lambda)/3\lambda$. This expression can be solved for b using the fact that $\mu Q'\Sigma Q + R$ is symmetric with 5 diagonal bands. Once the values for b are obtained we can substitute it back into our previous equations to find the ordinates of the splines using

$$d = y - \mu\Sigma b$$

and the remaining coefficients can be obtained from our previous equations and we have fully defined our smoothing spline for all knot intervals.

Bibliography

Buuren, S. V. Worm plot to diagnose fit in quantile regression. *Statistical Modelling*, 7(4):363–376, 2007.

Canadian Society for Exercise Physiology. Canadian Society for Exercise Physiology Canadian Physical Activity, Fitness & lifestyle approach. Technical report, Canadian Society for Exercise Physiology, Ottawa, 2003.

Cole, T. J. and Green, P. J. Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, 11(10):1305–19, 1992.

Daniels, S., Morrison, J., Sprecher, D., Khoury, P., and Kimball, T. Association of body fat distribution and cardiovascular risk factors in children and adolescents. *Circulation*, 99(4):541–545, 1999.

Grummer-Strawn M, L., Reinold, C., and Krebs, N. F. Use of World Health Organization and CDC growth charts for children aged 0-59 months in the United States. Technical Report 9, 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20829749`.

Koenker, R. and Bassett, G. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. ISSN 1098-6596. doi: 10.1017/CBO9781107415324.004.

Koenker, R. and Hallock, K. F. Quantile regression, an introduction. *Journal of Economic Perspectives*, 15(4):43–56, 2001.

Koenker, R., Ng, P., and Portnoy, S. Quantile smoothing splines. *Biometrika*, 81(4): 673–680, 1994.

Koenker, R., Portnoy, S., and Zeileis, A. quantreg: quantile regression, R package version 5.21, 2016.

Pollock, D. S. G. Smoothing with Cubic Splines. In *Handbook of Time Series Analysis, Signal Processing and Dynamics*, pages 293–322. 1999. ISBN 9780125609906. doi: 10.1016/B978-012560990-6/50013-0.

R Core Team. R: A language and environment for statistical computing, 2016. URL `http://www.r-project.org/`.

Reilly, J. J., Methven, E., McDowell, Z. C., Hacking, B., Alexander, D., Stewart, L., and Kelnar, C. J. H. Health consequences of obesity. *Archives of Disease in Childhood*, 88:748–752, 2003. ISSN 1468-2044. doi: 10.1136/adc.88.9.748.

Rigby, R. A. and Stasinopoulos, D. M. Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54(3):507–554, 2005.

Rigby, R. A. and Stasinopoulos, D. M. Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. *Statistics in Medicine*, 23(19):3053–3076, 2004.

Roberts, F. and Barrodale, I. An improved algorithm for discrete L1 linear approximation. *SIAM J*, 10(5):839–848, 1973.

Stasinopoulos, M., Rigby, B., Voudouris, V., Akantziliotou, C., Marco, E., and Kiose, D. gamlss: Generalised Additive Models for Location Scale and Shape, R package version 4.3-8, 2015.

Statistics Canada. Canadian Health Measures Survey (CHMS): instructions for combining cycle 1 and cycle 2 data. Technical report, 2013.

Van Buuren, S. and Fredriks, M. Worm plot: A simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20(8):1259–1277, 2001. ISSN 02776715. doi: 10.1002/sim.746.

Wei, Y., Pere, A., Koenker, R., and He, X. Quantile regression methods for reference growth charts. *Statistics in Medicine*, 25(8):1369–1382, 2004. ISSN 0277-6715. doi: 10.1002/sim.2271.

World Health Organization. The new WHO child growth standards. Technical report, WHO, Department of Nutrition for Health and Development, 2008. URL `http://hpps.kbsplit.hr/hpps-2008/pdf/dok03.pdf`.