# Stepwise Modeling of Child Growth with Response Modeling Methodology (RMM)


## Haim Shore

*Department of Industrial Engineering and Management,*

*Ben-Gurion University of the Negev.*

POB 653, Beer-Sheva 84105, Israel (shor@bgu.ac.il)





**ABSTRACT**

Response Modeling Methodology (RMM) is a new method to model monotone convex relationships. Unique to RMM is its continuous-monotone-convexity property, which allows the data determine, via the estimated parameters, the final form of the model. This lends RMM modeling versatility that qualifies it to serve as general platform for nonlinear modelling. In this article, RMM is applied to model child growth curves. RMM quantile function is first fitted to existent reference centiles (as given at CDC and WHO web sites). Good accuracy is obtained. RMM is then used to model Dutch male head circumference data, based on a sample of 7040 observations. The poor fit obtained for a single model, combined with medical evidence, indicate existence of two separate growth processes. Therefore stepwise modeling is used with a cubic spline linking the two models. Multiplicative residuals from the estimated RMM median models are normal and independent of the growth process. This analysis is compared to former GAMLSS analyses that used semi-parametric models (expressed in terms of age) for the response first four moments. Given that comparison, a discussion of the risk of modeling noise, when using a-parametric or semi-parametric models, is given.



**Keywords:** *Child growth reference centiles; GAMLSS; quantile regression; RMM*


*Supporting information for this article is available in Supplementary Material, at the journal web-site.*

# 1. Introduction

Children growth charts consist of a series of percentile curves that illustrate the distribution of selected body measurements in children, like height and weight, as function of covariates, like age. Both the American Department of Health and Human Services (Centers for Disease Control and Prevention, CDC) and the World Health Organization (WHO) periodically publish detailed data tables for various children characteristics and affiliated growth charts. There is rich literature about methods to construct child growth charts. A comprehensive overview of these has appeared in [4]. Table 1 therein summarizes 30 existing methods that "could potentially be used for the construction of the WHO attained growth curves." A recent newcomer to this arsenal of methodologies is generalized additive models for location, scale and shape (GAMLSS). It has been introduced in [1], [12] and [14] as a way of overcoming some of the limitations associated with the popular generalized linear models (GLM) and generalized additive models (GAM) ([11] and [10], respectively). In 2006, GAMLSS was adopted by WHO [29] as a general methodology to construct world standards for child growth curves. More recently, WHO [30] has used, in the framework of GAMLSS, the Box-Cox-power-exponential distribution (BCPE, [13]), with cubic spline smoothing, to construct growth standards for growth velocity based on weight, length and head circumference.

Child growth curves, expressed in terms of age or age-dependent covariate, are often monotone concave as the growth rate tends to slow down with advancing age. However, the diversity of shapes one can find amongst children growth percentile curves makes it difficult to adhere to parametric modeling, resulting in the use of a-parametric or semi-parametric modeling (as the numerical example below, relating to

GAMLSS implementation, testifies). This leads to loss of the unique advantages traditionally linked to parametric modeling, like statistical power. Furthermore, the sharp distinction between modeling of signal (which is desirable) and modeling of noise (extremely undesirable) is often blurred when purely parametric modeling is abandoned. On the other hand, a-parametric or semi-parametric modes of modeling, like those used in GAMLSS, offer modeling versatility that parametric modeling, with its a-priori specified model, often cannot provide.

Recently, a new platform for parametric modeling of monotone convex relationships has been introduced, denoted Response Modeling Methodology (RMM [17]; "Convex" will henceforth relate, unless otherwise specified, to both convex and concave). RMM may be viewed a hybrid of the parametric and the non-parametric approaches. On the one hand, it is a-parametric in the sense that the data decide the shape of the final model. On the other hand, it is parametric in the sense that a parametric model is used to model the nonlinear relationship. This strange combination becomes accessible via RMM due to its unique "continuous monotone convexity (CMC)" property. To understand the nature of this property and its implications to modeling child growth curves, consider the inverse Box-Cox (BC) transformation (a special case of RMM and cornerstone to modeling growth curves by the most widely implemented LMS method; [6], [7]):

$$y = [1 + \lambda(\eta + \sigma z)]^{(1/\lambda)} \quad , \qquad (1)$$

where y is a specified percentile of the response (the modeled variable), z is the corresponding standard normal percentile, $\eta$ is a linear combination of covariates (often denoted the linear predictor) and $\{\sigma, \lambda\}$ are parameters. It is easy to realize that this transformation comprises three distinct models (linear: $\lambda$=1, power: $\lambda \neq 0,1$, and exponential: $\lambda \to 0$), all having monotone convexity that increases in intensity as we

move from linear to power to exponential relationships. The inverse BC function transforms these models into mere points on the continuous spectrum spanned by lambda ($\lambda$), the BC parameter. Thus, data solely determine the final form of the model. RMM expands considerably the span of models represented by the inverse BC transformation. This is achieved by delivering representation to a hierarchy of models that may be arranged, according to their degree of convexity, on the "Ladder of monotone convex functions" (refer for details to Appendix A and [21]). As with the inverse BC transformation, the transition from one model to the next is linked to changes in the shape parameters of the model. Since these parameters are estimated from data, the latter determine the final shape of the model. This lends RMM versatility in describing monotone convex relationships that is not shared by other current models or modeling methodologies (like generalized linear models, GLM). Thus, major merits of a-parametric modeling (like modeling versatility) and of parametric modeling (like statistical power) are preserved in the RMM approach.

The purpose of this article is twofold. First, to demonstrate RMM capability of modeling child growth curves. This will be done using two sets of reference centiles that appear at the web-sites of CDC (American Centers for Disease Control and Prevention) and WHO (World Health Organization ). No sample data are used in this part of the paper. The second objective is to demonstrate RMM stepwise modeling of real data where apparently different growth processes require different models. This is done by analyzing a data set formerly analyzed with GAMLSS (refer to Stasinopoulos and Rigby [27], henceforth SR2007, and references therein). A comparison of the results obtained via the two methodologies may highlight the differences in approach between them. In particular, the risk of modeling noise when using a-parametric or semi-parametric methods is addressed. Accordingly, this article is divided into two major parts. In Section 3, RMM is used to model (approximate) current age-dependent reference centiles that appear at CDC and WHO web sites. Section 4 delivers a

detailed RMM analysis of the data set used by SR2007, and compares the results to those of GAMLSS. Section 4 summarizes this article with some conclusions. The next Section 2 provides background overview of reported applications of RMM modeling.

## *2. Modeling with RMM – a brief historic overview*

RMM models deliver the quantile of a random variable (not necessarily normal) in terms of the respective standard normal quantile and a linear combination of covariates (LP, the linear predictor). Since its introduction some years ago (Shore [17]), RMM has proved to be a versatile and effective modeling platform in diverse areas, such as distribution fitting ([18], [20], [24]), chemical engineering [26], quality engineering ([19], [16]), ecology [23] and fetal growth modeling and monitoring [2,25]. In all these references, RMM has been shown to approximate well existent (published) parametric monotone convex relationships. Furthermore, RMM delivers estimated parametric models that often have goodness-of-fit and stability comparable to or better than those of competitive parametric models, even when the latter have a larger number of parameters. Some examples are [3], [21], [22], [23], [26] and references therein. Two recent articles give an advanced overview of RMM [21] and its estimation procedures [22]. A brief tutorial on RMM is given in Appendix A.

## *3. Modeling growth reference centiles with RMM*

In this section we *fit* RMM models to two sets of growth curves, comprising 1179 and 660 reference *(population)* centiles, and evaluate the accuracy obtained. The purpose of the fitting is to demonstrate RMM versatility in delivering satisfactory representation to age-dependent differently-shaped child-growth curves. It is noted that RMM is fitted to existent *population* centiles that had been derived from data using smoothing

functions (like cubic splines). Therefore, model residuals are not expected to exhibit the characteristic pattern of random scatter seen when modeling sample (raw) data.

### *3.1. Example 1 (WHO)*

The centiles set for this example is taken from the CDC site, at page: Data Table for Boys Weight-for-length and Head Circumference-for-age Charts . The set analyzed here relates to “Boys weight-for-length”, with length in the range 45-110 cm. It includes 131 percentile sets, each associated with a certain age-dependent covariate value (length) and comprising 9 weight percentiles corresponding to:

$p_j$ (%)= {2.3, 5, 10, 25, 50, 75, 90, 95, 97.7}.

Altogether there are 131x9= 1179 weight percentile values. To fit the RMM model we minimize the sum of squared deviations of the given centiles from the corresponding RMM quantile, namely:

$$OF = \sum_{j=1}^{9} \sum_{i=1}^{131} [y_{i,j} - Q(\eta_i, z_j)]^2 \rightarrow Minimum \qquad (2)$$

where $y_{i,j}$ is child-weight $p_j$ percentile, corresponding to length $x_i$, $Q(\eta_i,z_j)$ is the RMM quantile (eqs. A.3 or A.6 in Appendix A) with linear predictor (LP) $\eta_i$ (corresponding to $x_i$) and standard normal percentile $z_j$, corresponding to $p_j$ (refer for a definition of the LP to Appendix A, eq. A.5 ). To facilitate the numerical routine, parameters are identified in two stages: First the median is modeled by RMM median (A.7), using all median values in the data set (131 values). In that stage LP and the RMM parameters *{m,b}* (and possibly also a response location parameter, L) are identified, assuming, without loss of generality: *a=1*. In the second stage, parameters {*c,d*} are found that minimize (2), using RMM quantile function (A.6), the fitted median model (A.7) taken from the first stage and all percentile values in the set. In that stage, the complete RMM quantile function is estimated (relate to [22] for discussion of the merits of the two-stage fitting

procedure; see also comment below). Note, that since we model existing *population* centiles, we minimize the mean squared error (the discrete analog of the L2 norm), rather than mean absolute deviation (MAD), which should have been minimized on estimating a median model from raw data (refer for details to the numerical example in Section 4). Applying this routine with: $\eta=\beta_0+\beta_1 x$ *(a=1,L=0)*, we obtain the parameters given in Appendix B. Figure 1a displays source length-related median child weights (upper plot) and the relative errors (in %) obtained from the fitted RMM quantile model.

*Insert Figure 1a about here*

*Insert Figure 1b about here*

We realize that for all 1179 percentiles in the WHO set, the fitted RMM centiles do not deviate more than about ±5% from actual values. For very small child weight values, the *percentage* errors are somewhat larger, as could be expected. Note that a column of residuals in Figure 1a represents deviations associated with the nine centiles particular to the respective length. As formerly noted, lack of complete randomness in deviations scatter is expected since the data used to derive the reference centiles had been extensively smoothed (to obtain smooth growth curves). A different scatter plot is obtained in the next Example 2.

## *3.2 Example 2 (WHO)*

This data set is taken from WHO site, at page: Head circumference-for-age: Birth to 5 years . This set relates to “Girls head circumference-for-age”, with age, in months, in the range 0-5 years. It includes 60 percentile sets (originally 61, however the first, with covariate x=0, was removed from the data set). Each set comprises 11 percentiles corresponding to:

p (%)= {1, 3, 5, 15, 25, 50, 75, 85, 95, 97, 99}.

Altogether there are 60x11= 660 percentile values. Applying the fitting routine (as described earlier), with: $\eta=\beta_0+\beta_1 x$ *(a=1,L=0)*, we obtain the parameters given in Appendix B. Figure 1b displays source age-related median head circumference (upper plot) and the relative errors (in %) of the fitted model. For all 660 percentiles in the data set, RMM fitted centiles do not deviate more than about $\pm$0.5% from the actual values. The deviations seem more randomly scattered than in the previous example, testifying to a smaller degree of raw data smoothing. However, one should still bear in mind that the input for the analyses are reference centiles that have been subject to data smoothing during their derivation.

# *4. RMM Modeling of Dutch males head circumference*

## *4.1 The data set and its history of analysis*

The introduction herewith pursues SR2007. The Fourth Dutch Growth Study ([8], [9] [28]) is a cross-sectional study that measures growth and development of the Dutch population between ages 0 and 22 years. The study measured, among other variables, height, weight, head circumference and age for 7482 males and 7018 females. There are 7040 observations, as there were 442 missing values for head circumference. Scatter plot of the data is given in Figure 2.

*Insert Figure 2 about here*

The data were previously analyzed by van Buuren and Fredriks [28], who found strong evidence of kurtosis which they were unable to model. The data were subsequently analyzed by Rigby and Stasinopoulos [15], using a Box-Cox t (BCT) distribution to model the kurtosis, and by SR2007, using same response distribution. Head circumference (y) of the males is analyzed in these analyses with explanatory variable (a transformed age): x = age$^{\xi}$ ($\xi$ is a parameter that needs estimation). In Section 4.2 we expound RMM modeling of this data set and in Section 4.3 GAMLSS

analysis, as conducted in SR2007, is presented. Section 4.4 compares the results obtained by the two approaches.

## *4.2 RMM modeling and estimation*

### *4.2.1 Preliminary analysis (single model vs. two models?) and RMM first-step estimation (estimating the median)*

Modeling the data set requires certain pre-processing in order to obtain sample estimates of local dispersion. These estimates may be used (if required) for weighting in the first stage of estimation (estimating the median) or for second-stage estimation (estimating the final quantile function using quantile regression, as detailed in [22]). Fortunately, SR2007 have already partitioned the data into 20 age groups of about equal sizes (about 350 observations each). We assume that within each group systematic variation comprises small portion of the overall variation so that measures of local dispersion can be obtained (specifically, MAD and standard deviation). Table 1 in Supplementary Material displays various statistics by age group. Some of these are used for weighting in various stages of estimating the RMM model.

For first stage estimation, an estimate of the RMM median model (eq. A.7) was obtained. Estimates of the median parameters are given in Appendix B. Plot of the model's predicted values, together with data scatter plot, are shown in Figure 2. Observing the data scatter, one may find evidence that perhaps two separate growth processes have generated the data: Children aged 0-10 years (first ten groups) and children/adults older than that (last ten groups). This distinction is corroborated by medical evidence, according to which "the start of adolescent growth" is "at about 10 years of age" (Cameron and Demerath [5]). Furthermore, given the flexibility of the RMM median model, due to its CMC property, failure of the estimated median model to deliver adequate representation to all median values (as shown in Figure 2, upper plot), particularly for age>10 years, is evidence that probably two separate models are

needed for adequate representation of the data (an expanded discussion of this point is given in the comments below). Therefore separate RMM median models were re-estimated for the two sets of age groups. The estimated median models cross at age 9.08 years. This indicates that the correct partition is into groups1-9 (ages 0 to 9 years) and groups 10-20. RMM median models were re-estimated for this new partition and found to cross at age 8.557 years, thus validating the new partition of the data. Parameters' estimates for the two median models are given in Appendix B. Table 1 delivers averages of MAD (mean absolute deviation per age group) for the two sets of age groups: An average based on sample medians and that based on the modeled medians.

*Insert Table 1 about here*

One realizes that the estimated median models deliver adequate representation to the sample MAD values. Scatter plots of the data together with the estimated median models are displayed at the bottom two plots of Figure 2. Figure 3 plots age-group sample medians and the fitted models as function of group's average age.

*Insert Figure 3 about here*

Obviously, fit of the RMM median is improved, relative to the initial partition of the data. Henceforth we will model the data set using two different RMM models relating to data partition as delineated above. A cubic spline will smooth the transition between the two median models, as detailed in Section 4.2.3.

**Comments**

**(i)** RMM stepwise modeling, as implemented above, may raise three concerns. First, perhaps lack of good fit in modeling the complete sample derive from the inadequacy of the RMM model and not from existence of two growth processes. Secondly, what is the justification of a two-stage estimation procedure instead of a single stage

procedure. Thirdly, why not switch to semi-parametric quantile regression estimation and use MLE procedure for the complete sample?

Relating to the first concern, we first note that while change-point methods exist for linear regression and mixture linear regression procedures are widely available (for example, via R *regmix* routine), we are unaware of similar routines for non-linear modeling. This can be attributed foremost to the fact that there is no guarantee that a change point does not affect a change in the form of the model itself (as indeed happens in the current example). Since RMM has been fitted to scores of nonlinear parametric models in various disciplines (as expounded in the Introduction) and shown to result in negligible loss in accuracy, it may be asserted that lack of satisfactory fit for the upper half of the sample is evidence for two separate growth processes rather than for inadequacy of the RMM model. As noted earlier, this conclusion is medically corroborated.

Relating to the second concern, RMM modeling allows separation of the estimation of the median from estimation of the general quantile function. The two-stage estimation procedure reduces appreciably the number of parameters that need to be estimated at each stage. Consequently, the probability of obtaining local minima in the minimization routine (instead of a global one) is reduced. Furthermore, separation of estimation into two steps allows an increase in the number of covariates that may be included in the LP (estimated completely in the first stage).

Regarding the third concern, we show later in the paper that use of a semi-parametric modeling procedure for the data in Example 2 had led in the past to seemingly incorrect conclusions about the dependence of the growth centiles on moments higher than the first. In other words, we show that semi-parametric modeling can lead to modeling of noise. A discussion of this point is given in Section 4.4.

**(ii)** Observing the data for age groups≥10 years, one is tempted to use linear regression instead of RMM modeling. However, linearity is a special case of the RMM model (relate to eq. A2). Therefore, if indeed a linear model is suggested by the data this will show in the estimated RMM parameters for the upper part of the sample.

### *4.2.2 RMM estimation - second step (estimating the quantile function)*

Having estimated the median (eq. A.7), stage 2 estimates the last two parameters: *c* and *d* (refer to A.6). RMM model used at this stage is the approximate model (last part of A.6), which has explicit probability density function (pdf), needed to calculate likelihood functions. Also it has one less parameter. In this subsection we show results from estimation using nonlinear quantile regression. The procedure first find estimates for $\{c_p, d_p\}$, the parameters of the RMM quantile model (eq. A.6) for the 100p% centile. Four centiles are modeled, corresponding to CDF values of:

p = {0.10, 0.25, 0.75, 0.90}.

Estimates are found by minimizing the sample average:

$$\widehat{E}_p = \frac{1}{n}\sum_{i=1}^{n} \rho_p(y_i - \hat{y}_p), \qquad (3)$$

where $\hat{y}_p$ is the RMM model for the 100p% centile, and $\rho_p$ is the "Tilted absolute value function", or the "Check function", for that centile. This function is cornerstone for quantile regression (see details, for example, in [22]), and it is generally defined by

$$\rho(u) = u[p - I(u<0)], \qquad (4)$$

where *I* is the indicator function (*I(C)=1*, if condition *C* is true; *I(C)=0*, otherwise). Estimates of {$c_p$, $d_p$} for the four different values of p are given in Appendix B. To obtain a single RMM model, with {c,d} independent of p, we minimize (find details in [22]):

$$OF = \sum_{k=1}^{4} \sum_{i=1}^{n} (w_k)(\frac{1}{n})\rho_{p_k}(y_{i,j} - \hat{y}_k) \rightarrow Minimum, \quad (5)$$

where weights {$w_k$} are reciprocal values of the estimated $E_p$ (p= 0.10, 0.25, 0.75, 0.90; k=1,2,3,4), as obtained for the separate four models (eq. 3), $\hat{y}_k$ is RMM model for the standard normal k-th centile (A.6), and n is the sample size (equal for all estimated quantile models). Note that weighting is needed since dispersion around the separately estimated four models is not homogenous (similar weighting is implemented in linear regression if dispersion of different observations, as given by the error standard deviations, is non-homogenous; find more details in [22]). The final estimates of {c,d} are displayed in Appendix B. Figure 4 shows RMM estimated quantiles, based on a single RMM quantile model for all p values, for some of the above four p values, together with the data scatter plots. Figure 5 shows sample quantile values, for the different age-groups, together with RMM quantile values, calculated at the groups' mean average age.

*Insert Figure 4 about here*

*Insert Figure 5 about here*

### *4.2.3 Adding a cubic spline for smooth transition between the estimated median models*

Two RMM median models had been estimated in a step-wise procedure that resulted in two different models with a common intersection point (IP) at age $x_0$= 8.557. Since the two models have different slopes at their IP, a transition model is required that guarantees smooth transition from one model to the next with respect to both models'

values at transition points and their slopes. We fit a cubic equation to the interval of $\{x_0 \pm \Delta\}$, where $\Delta$ is arbitrarily set as $\Delta$=1/2:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3,\ x_0 - \Delta \le x \le x_0 + \Delta \qquad (6)$$

Parameters of the cubic equation are found that minimize the sum of relative deviations of the values and the slopes of (6) at $x_0 \pm \Delta$ from respective values of the estimated RMM models (valid at each transition point). We obtained:

$\beta_0$= 64.94; $\beta_1$= -3.223; $\beta_2$= 0.1994; $\beta_3$= 7.8E-4; $x_0$=8.557; $\Delta$= 1/2.

Note that the model slope at the lower transition point ($x_0$-$\Delta$) is 0.1409, and the slope at the upper transition point ($x_0$+$\Delta$) is 0.5794. The rapidly changing slope emphasizes that two growth processes are at hand here and justifies the adoption of two separate median models with a cubic spline that links them together. Figure 6 displays the two models and the cubic transition spline. We realize that the cubic equation allowed smooth transition at the transition points from one RMM median model to the next.

*Insert Figure 6 about here*

### *4.2.4 Evaluation of RMM Analysis - the final model*

Values of estimated {*c,d*}, as shown in Appendix B, are generally small (except for *d* for the lower model). Observing these parameters as they appear in RMM quantile model (first part of eq. A.6), we realize that the coefficients of the standard normal centile, *z*, are extremely small. These are ($\eta$ is the linear predictor):

For the lower model 1: $c_1/\eta = 0.04316/\eta$, $d_1 = 0.6742$; $(\eta>1)$;

For the upper model 2: $c_2/\eta = 0.04288/\eta$, $d_2 = 0.02609$. $(\eta>1)$.

These small values indicate that the RMM median model (c=d=0 in A.6) may provide adequate and satisfactory fit to the data. In other words, deviations around the median are not age-dependent and therefore the z quantile vanishes from the RMM estimated quantile function (which leave us with a model for the median). Since the median is a multiplicative term in the RMM model A.6, we define the RMM error (residual) for observation $y_i$, assuming that the RMM median is the adopted model, as:

$$\varepsilon_i = \frac{y_i}{M_Y(\eta_i)}, \qquad (7)$$

where $M_Y$ is the RMM median (function of $\eta_i$, the LP value associated with observation i). Table 2 shows some statistics relating to the errors. Note in particular the near zero values of the correlations (of error with age), which indicate no relationship between error and age. *In other words: An attempt to model any error moment in terms of age is equivalent to modeling of noise!*

*Insert Table 2 about here*

Figure 7 displays a scatter plot of the errors, Figure 8 is a Q-Q plot (assuming normality), and Figure 9 shows the error mean, standard deviation, skewness and kurtosis for the separate age-groups (as function of group average age). All deliver evidence that the errors have distributions that are not related to age (the covariate) and that these distributions are approximately normal. Expressing the errors on a log scale does not improve normality. Figure 8 shows some departure from normality at the tails of the plot. However, this is expected for a sample that large (7040 observations).

*Insert Figures 7-9 about here*

It is easy to calculate the log-likelihood associated with model (A.7). The density function, $f_Y(.)$, of observation $y_i$ is (with *c=d=0*, assuming normality of errors):

$$f_Y(y_i) = \frac{1}{\widehat{\sigma}_\varepsilon M_Y(\eta_i)} \phi[\frac{\varepsilon_i - \widehat{\mu}_\varepsilon}{\widehat{\sigma}_\varepsilon}], \quad (8)$$

where $\phi$ is the standard normal density function and $\widehat{\mu}_\varepsilon$ and $\widehat{\sigma}_\varepsilon$ are estimates of the errors' mean and standard deviation, respectively (as these are given in Appendix B, separately for the two models). Using (8) we obtain a total log-likelihood value of -13544 (-5914. 6 for first model and -7629 for second model). This is about 1.3% smaller than the value of -13366, obtained via a ML procedure for the non-parametric cubic-spline-based GAMLSS model (refer to SR2007 and the next Section 4.3). Note again that in estimating the RMM models normality of the models' multiplicative errors was assumed (for both models). This assumption is supported by the evidence provided earlier.

## *4.3 GAMLSS Modeling (based on SR2007)*

SR2007 assume that the response has distribution BCT, with the $\alpha$ quantile given by:

$$y_\alpha = \begin{cases} \mu(1+\sigma\nu t_{\tau,\alpha})^{1/\nu}, & \nu \neq 0 \\ \mu \exp(\sigma t_{\tau,\alpha}), & \nu = 0, \end{cases} \quad (9)$$

where $t_{\tau,\alpha}$ is the $100\alpha$ centile of $t_\tau$, a standard t distribution with degrees of freedom parameter $\tau$. The four parameters ($\mu,\sigma,\nu,\tau$) may be interpreted as relating to location (median), scale (centile-based coefficient of variation), skewness (power transformation to symmetry) and kurtosis (degrees of freedom), respectively. These parameters are estimated via R GAMLSS package, assuming the smooth non-parametric functions of $x=(age)^{\xi}$:

$$g_k(p_k) = h_k(x),\ k=1,2,3,4\ , \quad (10)$$

where $g_k(p_k)$ is the link function of the k-th parameter ($p_k$) of the above t distribution (k=1 for the median; $k$=2 for scale; $k$=3 for skewness; $k$=4 for kurtosis). For the final model, SR2007 selected identity link functions for $\mu$ and $\nu$, while log link functions were

assumed for $\sigma$ and $\tau$ (to ensure $\sigma > 0$ and $\tau > 0$). Note that no parametric functions of co-variates are used in this model. The final model fitted (with penalty value 3 in the penalized maximum likelihood procedure):

$$BCT(df_\mu, df_\sigma, df_\nu, df_\tau, \xi) = BCT(15.77, 8.05, 2, 2, 0.28), \quad (11)$$

where $df_p$ is the total (effective) degrees of freedom for the smooth non-parametric cubic spline functions for parameter $p$ (an element of $\{\mu, \sigma, \nu, \tau\}$). The associated optimal global deviance, -2($LL$*), where $LL$* is the optimized log-likelihood, is 26732.04. Various goodness-of-fit criteria, relating to the standardized quantile residuals, are provided in SR2007 to establish that residuals within the various age groups are normally distributed. The estimated parameters are given graphically as function of age, and may serve to derive conclusions about how shape characteristics of the response distribution vary with age. For example, from Figure 13 therein one deduce that the skewness parameter, $\nu$, monotonically decreases with age (plot c), while the kurtosis parameter, $\tau$, monotonically increases with age (plot d). Further details about GAMLSS analysis of this data set may be found in SR2007.

## *4.4 Comparing two approaches: Parametric (RMM) vs. semi-parametric (GAMLSS)*

RMM analysis of the data in the numerical example demonstrates some of the relative merits of RMM for modeling child growth:

(1) No a-priori specification of the response distribution needed (this is determined by the data);
(2) No a-priori specification of the response median model needed (this is determined by the data);
(3) A single LP is used.

The main results from the RMM analysis are:

(1) Two simple median models (for the (0-9) and the (10-22) age-groups), with a cubic spline to ensure smooth transition between them, have delivered adequate representation to the data and the two underlying child growth processes;
(2) Median models' multiplicative residuals are approximately normal and are not correlated with age.

The adequacy of the two RMM estimated models helped gain new insight regarding the modeled growth process. Although the process seems to differ (in terms of the median) between the two sets of age groups, the response distribution around the medians is independent of age. This is a new finding that stands in stark contrast to the GAMLSS model of SR2007, where *a-parametric* modeling of scale, skewness and kurtosis presumes that the response distribution continuously varies with age (beyond changes in location). As judged by the results obtained from RMM *parametric* modeling, application of semi-parametric GAMLSS in SR2007 apparently led to modeling of noise.

## *5. Conclusion*

Two examples of fitting RMM to approximate population centiles, as delivered by WHO, demonstrate that RMM provides a proper platform to model these centiles. (Further examples may be provided from the author on request.) A data set of 7040 observations of children growth, formerly analysed using GAMLSS framework, is re-analysed, using RMM. An RMM median model, estimated separately for data partitioned into two sets, has produced adequate fit, notwithstanding the small number of parameters (5 for each model, including 2 for the LP). The multiplicative residuals have been shown to constitute a single normal distribution, rendering highly questionable modeling (as function of age) of parameters associated with second degree moment or higher (as done in GAMLSS analysis). The only linkage with age

seems to be the change point around 9 years of age, which required fitting a different median model. Once this change has been addressed, the residuals are unrelated to age.

The unique combination of advantages of parametric and a-parametric nonlinear modelling as provided by RMM, discussed in the Introduction and demonstrated by the numerical example, seems to suggest that RMM may provide adequate platform for modelling child growth curves.

## *Appendix A. A tutorial on RMM modeling*

Exposition herewith pursues Shore [21]. RMM models a monotone convex relationship between the percentile of a response, *Y*, a linear combination of predictor variables (the linear predictor, LP, denoted $\eta$) and the respective standard normal percentile. The basic RMM model describes a modeled response, *Y* (a random variable) in terms of the LP, two possibly correlated zero-mean normal errors, $\varepsilon_1$ and $\varepsilon_2$ (with correlation $\rho$ and standard deviations $\sigma_{\varepsilon 1}$ and $\sigma_{\varepsilon 2}$, respectively), and a vector of parameters $\{\alpha, \lambda, \mu\}$:

$$W = \log(Y) = \mu + \frac{\alpha}{\lambda}[(\eta + \varepsilon_1)^{\lambda} - 1] + \varepsilon_2. \qquad \text{(A.1)}$$

Note that $\varepsilon_1$ implies that there is uncertainty (either measurement imprecision or otherwise) in the explanatory variables (included in the LP). This is in addition to uncertainty associated with the response ($\varepsilon_2$). One may realize that various common scientific and engineering models can be derived from (A.1). To demonstrate that, let us *temporarily* ignore the errors and the scale parameter, $\mu$. One obtains from (A.1) for $\lambda$=0:

$$\log(Y) = \lim_{\lambda \to 0}\{(\frac{\alpha}{\lambda})[\eta^{\lambda} - 1]\} = (\alpha)\log(\eta) = \log(\eta^{\alpha}). \qquad \text{(A.2)}$$

From (A.2), a linear relationship between the response and the LP is obtained for $\alpha$=1 and a power relationship for $\alpha\neq1$. An exponential relationship is obtained from (2) for $\lambda$=1; an exponential-power relationship for {$\lambda\neq1$, $\lambda\neq0$} and so on. In fact, all models that appear on the "Ladder of monotone convex relationships", a core concept of RMM, may be derived from (A.1). A detailed explanation of the ladder, models comprising the ladder and the ladder's relationship to (A.1) may be found in Shore ([17], [21]). Note that the power and exponential relationships, associated with the inverse Box-Cox transformation (and used in the LMS method), are special cases of RMM. The quantile function associated with (A.1) is ([21]):

$$\log(y) = \log(M_Y) + \frac{(a\eta^b)}{b}\{[1+(c/\eta)z]^b - 1\} + (d)z + \varepsilon \cong \log(M_Y) + \frac{(ac\eta^{b-1})}{(bc/\eta)}[e^{\frac{(bc)z}{\eta}} - 1] + (d)z + \varepsilon, \quad \text{(A.3)}$$

where {*y,z*} are respective percentiles of the response, *Y*, and a standard normal variate, *Z,* $\varepsilon$ is the model's zero mean normal error with constant variance, $\sigma^2$, {*a,b,c,d*} are parameters and:

$$\log(M_Y) = \mu + \left(\frac{a}{b}\right)[\eta^b - 1] = \log(m) + \left(\frac{a}{b}\right)[\eta^b - 1], \quad \text{(A.4)}$$

$M_Y$ being the median of Y (the 50% percentile of *Y*, corresponding to *z=0*), and $\mu$ (or *m*) is an additional parameter that needs estimation. Note that the LP (with *k* predictor variables) is defined by:

$$\eta = \beta_0 + \beta_1 X_1 + .. + \beta_k X_k \, . \quad \text{(A.5)}$$

This implies that in estimating the median (A.4), apart from the parameters of LP only two additional RMM parameters need estimating (assuming, without loss of generality: *a*=1). If the response data contain values that change sign, or if the lowest

response value is far from zero (for example, when data are left truncated), a response location parameter, *L*, may be added so that (A.3) and (A.4) become, respectively:

$$\log(y-L)=\log(M_Y-L)+\frac{(a\eta^b)}{b}\{[1+(c/\eta)z]^b-1\}+(d)z+\varepsilon\cong \log(M_Y-L)+\frac{(ac\eta^{b-1})}{(bc/\eta)}[e^{\frac{(bc)z}{\eta}}-1]+(d)z+\varepsilon, \quad \text{(A.6)}$$

$$\log(M_Y-L)=\mu+\left(\frac{a}{b}\right)[(\eta)^b-1]=\log(m)+\left(\frac{a}{b}\right)[(\eta)^b-1]. \quad \text{(A.7)}$$

In Shore [22], several estimation procedures are described. All have two-stages. The first stage is common to all and that is when RMM median is estimated via minimization of the mean absolute deviation (MAD) (sum of absolute deviations from the median divided by the sample size). Minimization of a weighted MAD may also be needed. At this stage the LP is estimated plus two RMM parameters ($\mu$ and b, assuming, without loss of generality: a=1). In the second stage, using estimates from stage 1, the remaining parameters are estimated (two parameters, *c* and *d*). It is at the second stage that RMM estimation may pursue different procedures. Three different second-stage routines are described in [22]: ML estimation, two-moment matching and nonlinear quantile regression. The latter option is used in this article. This method initially estimates four different RMM quantile functions (excluding the median) and then uses weights derived from this initial estimation to concurrently estimate all four quantile functions so that a single set of estimates of RMM parameters{*c,d*} are obtained. This method is demonstrated in the numerical example of Section 4.

# *APPENDIX B. Parameters of RMM models (from fitting or estimating)*

| RMM Parameters (Fitting; Section 2) | | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | m | b | L | c | d |
| **Example 1 (Section 2.1)** | -.5883 | .01370 | 19.41 | .2884 | 0 | 5.450E-4 | 8.288E-2 |
| **Example 2 (Section 2.2)** | 38.87 | 18.18 | 5.220 | -0.4198 | 0 | 1.229 | 2.905E-2 |

| RMM Parameters (Estimating; Section 3) | | | | | | | |
|---|---|---|---|---|---|---|---|
| **RMM Model** | $\beta_0$ | $\beta_1$ | m | b | L | c | d |
| **Median (Initial model)** | -0.5282 | 12.48 | 1.906 | -0.3775 | -36.93 | | |
| **Median (groups 1-9)** | 1.731 | 5.138 | 12.85 | -0.9149 | 15.30 | | |
| **Median (groups 10-20)** | 18.78 | 1.182 | 7.808 | -0.3804 | 0.3165 | | |
| **10% percentile (1-9, 10-20)** | | | | | | (.05993, .7769) | (.04676, .02563) |
| **25% percentile (1-9,10-20)** | | | | | | (0.01931, .3282) | (.04546, .02881) |
| **75% percentile (1-9,10-20)** | | | | | | (.02645, .3576) | (.04219, .02878) |
| **90% percentile (1-9,10-20)** | | | | | | (.02988, .4383) | (.03986, .02708) |
| **Final Model (1-9,10-20)** | | | | | | (.04316, .6742) | (.04288, .02609) |

## *References*


[1] Akantziliotou C, Rigby RA, Stasinopoulos DM. The R implementation of generalized additive models for location, scale and shape. In M Stasinopoulos, G Touloumi (eds.), *Statistical Modeling in Society: Proceedings of the 17th International Workshop on Statistical Modeling*. 75-83. Chania, Greece. **2002**

[2] Benson-Karhi D, Shore H, Malamud M. Modeling fetal growth biometry with Response Modeling Methodology (RMM). Submitted **2013.**

[3] Benson-Karhi D, Shore H, Shacham M. Modeling temperature-dependent properties of water via response modeling methodology (RMM) and comparison with acceptable models. *Industrial & Engineering Chemistry Research* **2007**. 46(10): 3446-3463.

[4] Borghi E, de Onis M, Garza C, Van den Broeck J, Frongillo EA, Grummer-Strawn L, Van Buuren S, Pan H, Molinari L, Martorell R. Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. *Statistics in Medicine* **2006**. 25: 247–265.

[5] Cameron N, Demerath EW. Critical periods in human growth and their relationship to diseases of aging. *Yearbook of Physical Anthropology* **2002**. 45:159–184.

[6] Cole TJ. Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society A* **1988**. 151:385–418.

[7] Cole TJ. The LMS method for constructing normalized growth standards. *Eur J Clin Nutr.* **1990**. 44:45–60.

[8] Fredriks AM, van Buuren S, Burgmeijer R, Meulmeester J, Beuker R, Brugman E, Roede M, Verloove-Vanhorick S, Wit JM. Continuing positive secular change in the Netherlands, 1955-1997*. Pediatric Research* **2000**. 47:316-323.

[9] Fredriks AM, van Buuren S, Wit J, Verloove-Vanhorick SP. Body index measurements in 1996-7 compared with 1980. *Archives of Childhood Diseases* **2000**. 82:107-112.

[10] Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman and Hall, London. **1990**.

[11] Nelder JA, Wedderburn RWM. Generalized linear models. *Journal of the Royal Statistical Society A* **1972**. 135:370-384.

[12] Rigby RA, Stasinopoulos DM. The GAMLSS project: a flexible approach to statistical modeling. In B Klein, L Korsholm (eds.): *New Trends in Statistical Modeling: Proceedings of the 16th International Workshop on Statistical Modeling* , 249-256. Odense, Denmark. **2001**.

[13] Rigby RA, Stasinopoulos DM. Smooth centile curves for skew and kurtotic data modeled using the Box-Cox power exponential distribution. *Statistics in Medicine* **2004**. 23:3053-3076.

[14] Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *Applied Statistics* **2005**. 54:507-554.

[15] Rigby RA, Stasinopoulos DM. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modeling* **2006***,* 6: 209-229.

[16] Shauly M, Parmet Y. Comparison of Pearson distribution system and Response Modeling Methodology (RMM) as models for process capability analysis of skewed data*. Quality and Reliability Engineering International* **2011**. 27:68-687.

[17] Shore H. *Response Modeling Methodology (RMM) ─ Empirical Modeling for Engineering and Science.* World Scientific Publishing Co. Ltd, Singapore, 435. **2005**.

[18] Shore H. Comparison of Generalized Lambda Distribution (GLD) and Response Modeling Methodology (RMM) as general platforms for distribution fitting. *Communications in Statistics (Theory & Methods)* **2007**. 36(15): 2805-2819.

[19] Shore H. Comparison of linear predictors obtained by data transformation, generalized linear models (GLM) and Response Modeling Methodology (RMM). *Quality and Reliability Engineering International* **2008**. 24(4): 389-399.

[20] Shore H. Distribution fitting with the quantile function of Response Modeling Methodology. Chapter 13 in Karian, ZA, Dudewicz, EJ (Eds.): *Handbook of Fitting Statistical Distributions with R.* Taylor and Francis Group, LLC., 537-556. **2010**.

[21] Shore H. Response Modelling Methodology (RMM) ─ an advanced review. *WIREs Computational Statistic* **2011**. 3(4):357-372.

[22] Shore H. Estimating Response Modeling Methodology models ─ a focus article. *WIREs Computational Statistic* **2012**. 4(3):323-333.

[23] Shore H. Modeling and monitoring ecological systems - a statistical process control approach. *Quality and Reliability Engineering International* **2013**. On line: July, 2013.

[24] Shore, H. A general model of random variation. *Communications in Statistics (Theory & Methods)* **2013**. Forthcoming.

[25] Shore, H., Benson-Karhi, D., Malamud, M., Bashiri, A. Customized fetal growth modeling and monitoring - a statistical process control approach. *Quality Engineering* **2013**. Forthcoming.

[26] Shore H, Benson-Karhi D. Modeling temperature-dependent properties of oxygen, argon and nitrogen via Response Modeling Methodology (RMM) and comparison with acceptable models. *Industrial & Engineering Chemistry Research* **2010**. 49(19): 9469-9485.

[27] Stasinopoulos DM , Rigby RA. Generalized Additive Models for Location Scale and Shape (GAMLSS). R. *Journal of Statistical Software* **2007**. 23(7).

[28] van Buuren S, Fredriks M. Worm plot: a simple diagnostic device for modeling growth reference curves. *Statistics in Medicine* **2001**. 20: 1259-1277.

[29] WHO Multicentre Growth Reference Study Group. WHO Child Growth Standards: Methods and Development. World Health Organization, Geneva, Switzerland. **2006.**

[30] WHO Department of Nutrition for Health and Development. *Child Growth Standards Growth Velocity Based on Weight, Length and Head Circumference: Methods And Development*. World Health Organization, Geneva, Switzerland. 2009.

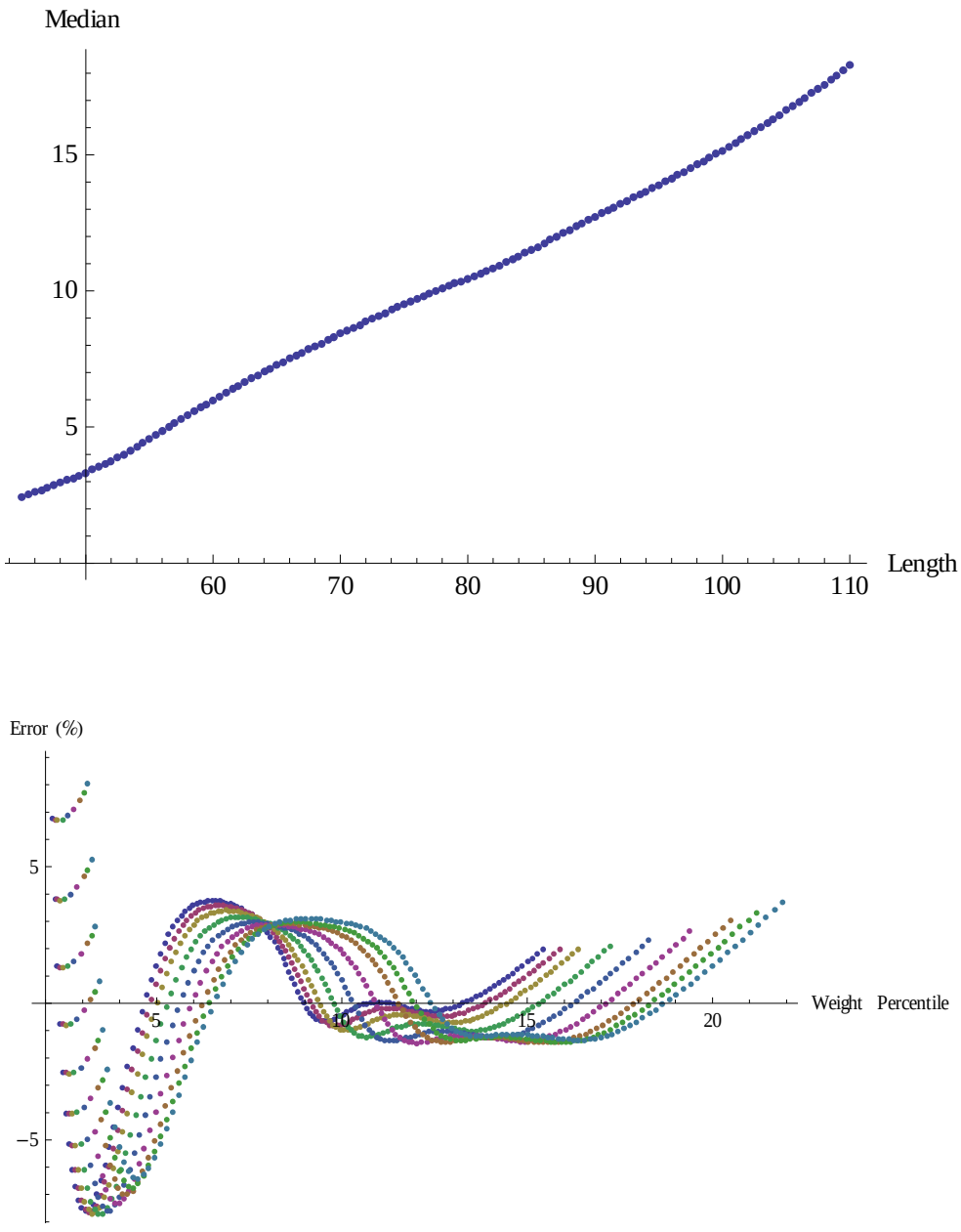


**Figure 1a.** Plots of length-related median weight (upper figure) and relative error (%) from fitting RMM quantile model to Sample 1.

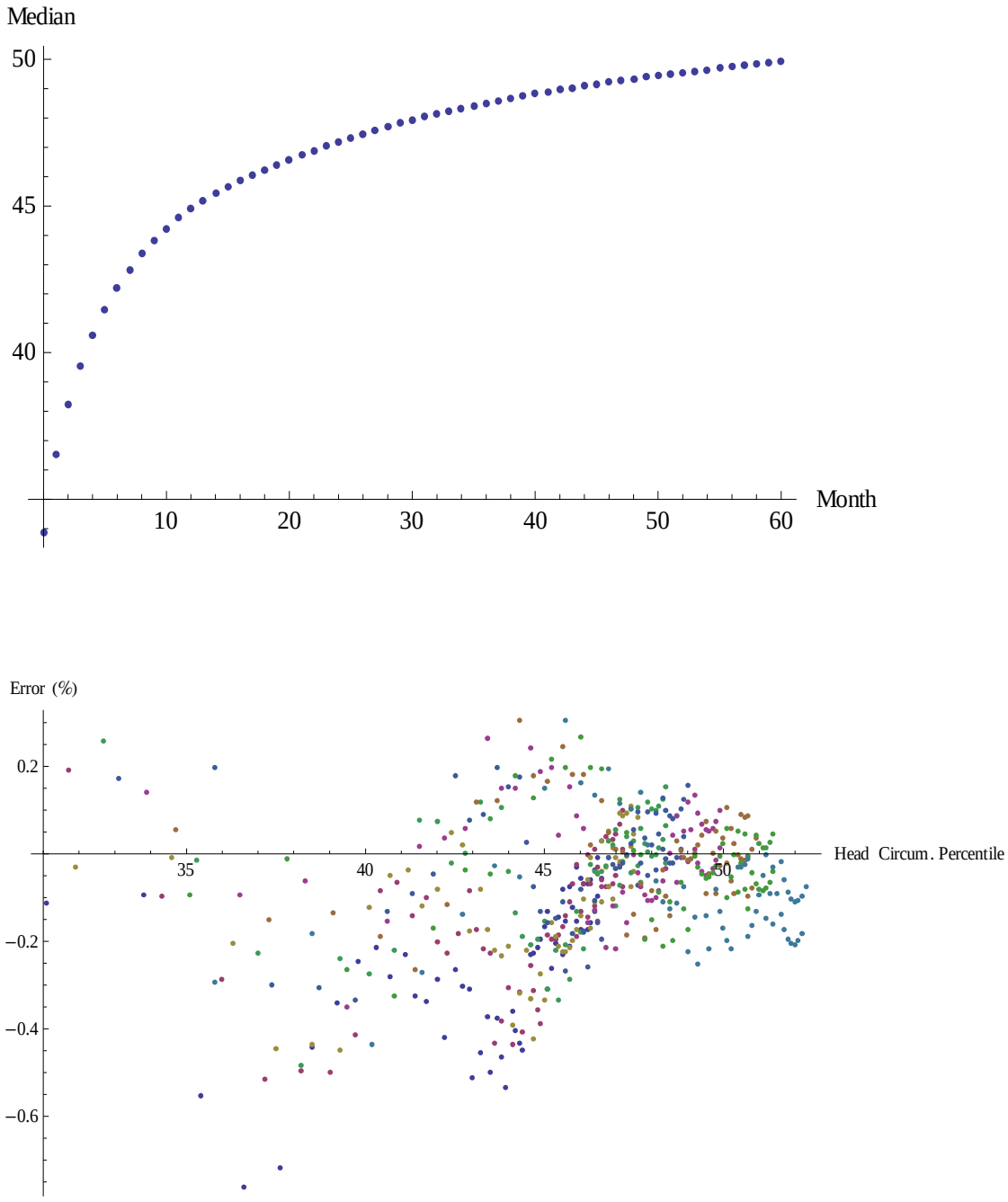


**Figure 1b.** Plots of age-dependent median head circumference (upper figure) and relative error (%) from fitting RMM quantile model to Sample 2.

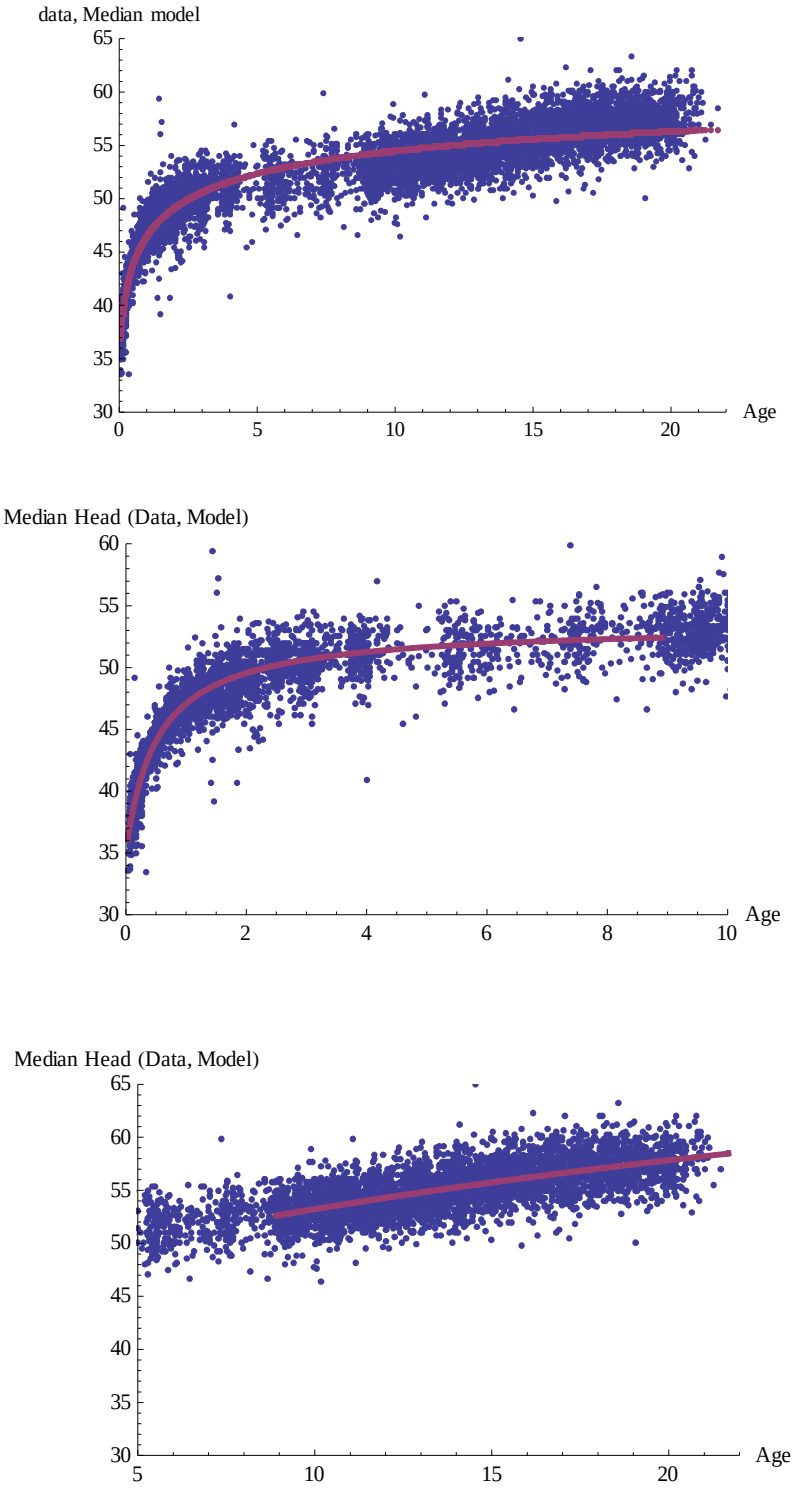


**Figure 2.** Original "Head" data (n=7040) with RMM-estimated median: Single model (upper plot); Model for age group 0-9 years; Model for age group 9-22 (lowest plot).

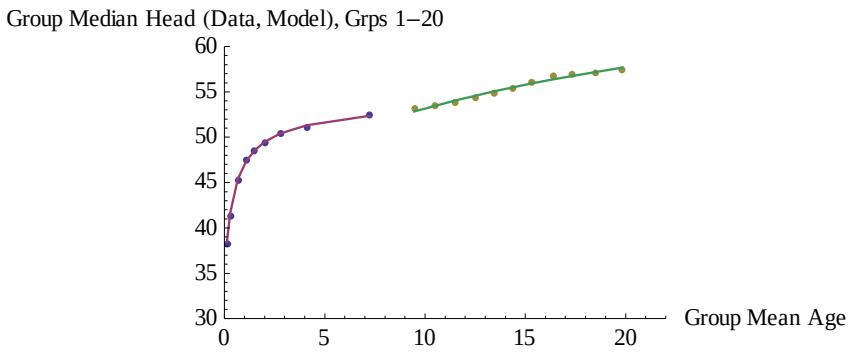


**Figure 3.** Age-group sample medians (points) and estimated RMM median (curve), calculated for group's mean age.

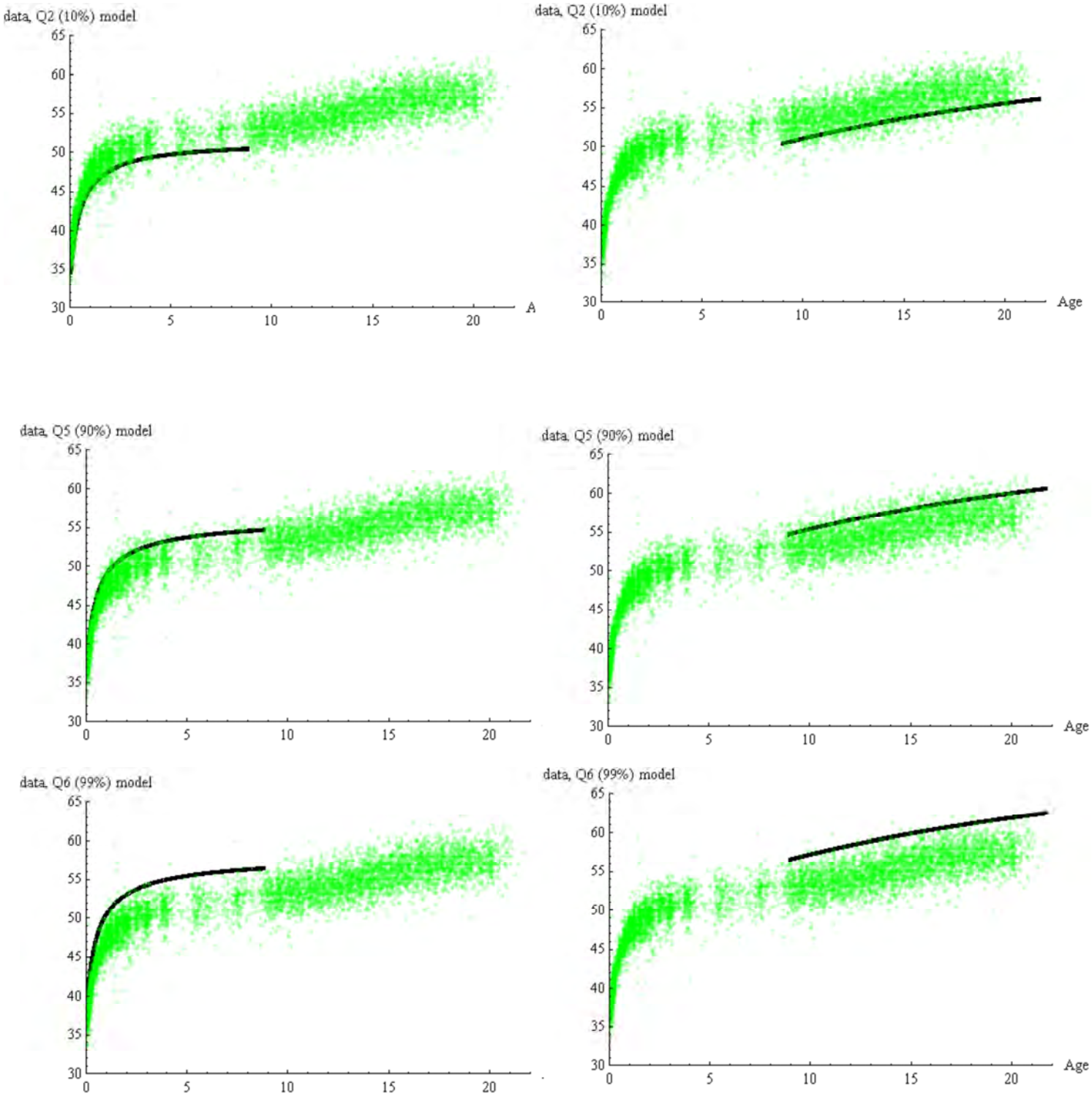


**Figure 4.** RMM estimated quantile functions (curves) with data scatter plots for some quantile models.

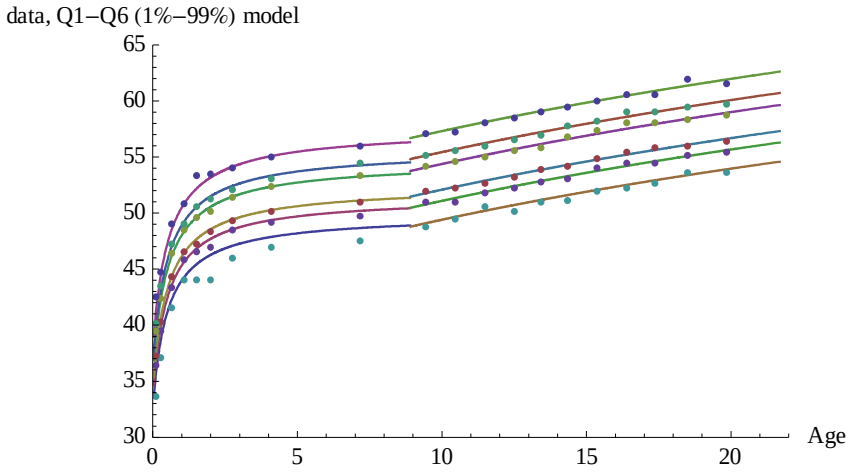


**Figure 5.** Age-group quantiles (points) and RMM quantiles (curves), for 1%, 10%, 25%, 75%, 90% and 99% percentiles (denoted Q1-Q6). Separate models for age (0-9) and (9-22) years.

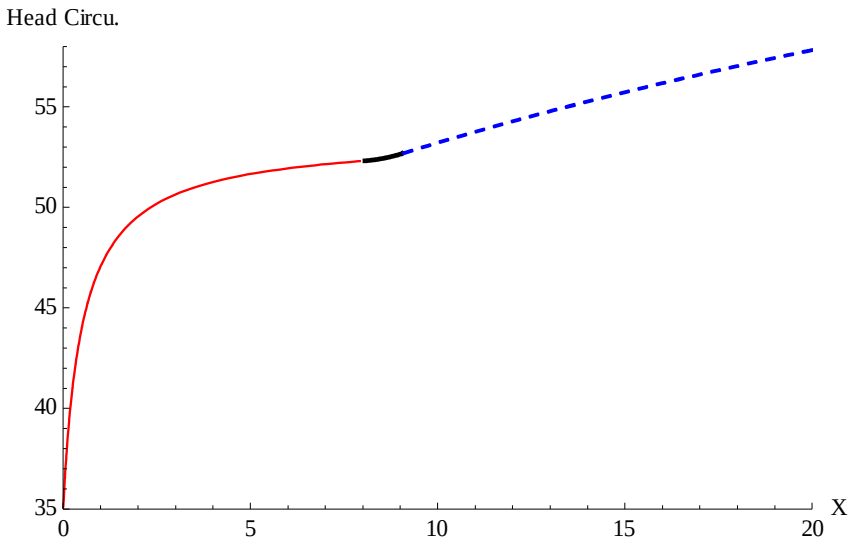


**Figure 6.** The two RMM models with the transition cubic spline that preserves both model's value and slope at the transition points (8.557±0.5).

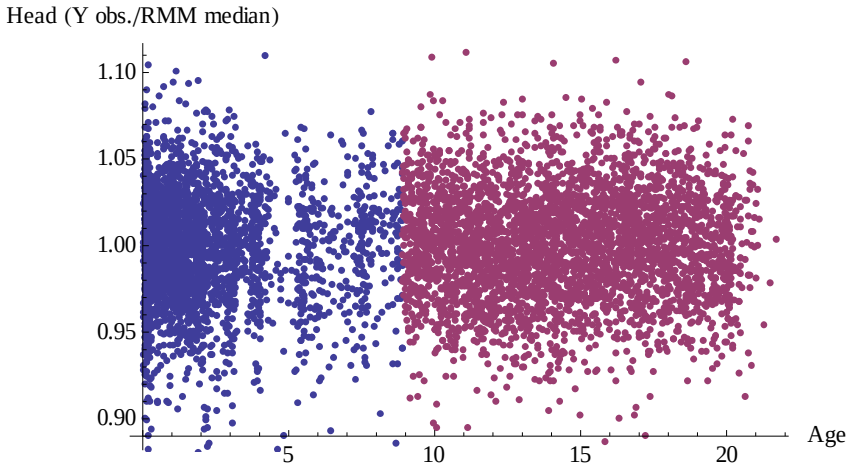


**Figure 7.** Scatter plot of errors (multiplicative) from fitting RMM median model.

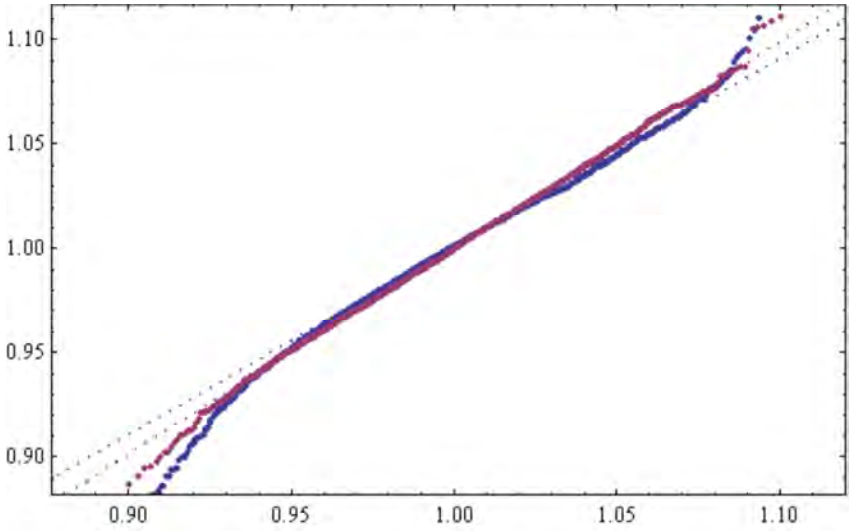


**Figure 8.** Error Q-Q plot (assuming normality). Predicted normal values are on the horizontal axis. No outliers deleted from data.

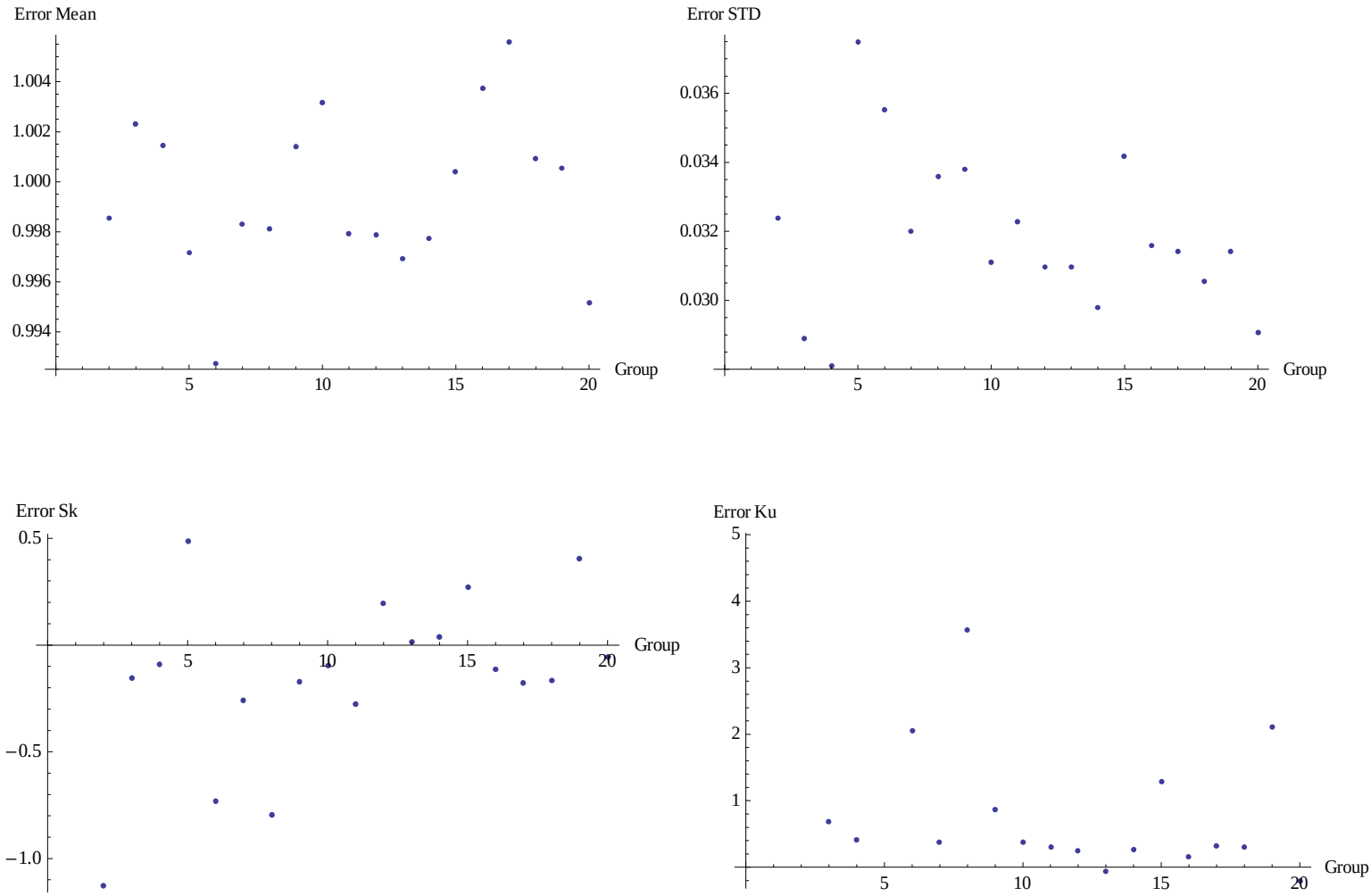


**Figure 9.** Error scatter plots, by group (numbered 1-20), for mean, STD, skewness and kurtosis. For the normal distribution: Sk=Ku=0.