Appl. Statist. (1999)
48, *Part* 4, *pp.* 533–551

Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children

Patrick J. Heagerty

University of Washington, Seattle, USA

and Margaret S. Pepe

Fred Hutchinson Cancer Research Center, Seattle, USA

[Received April 1998. Revised February 1999]

Summary. The appropriate interpretation of measurements often requires standardization for concomitant factors. For example, standardization of weight for both height and age is important in obesity research and in failure-to-thrive research in children. Regression quantiles from a reference population afford one intuitive and popular approach to standardization. Current methods for the estimation of regression quantiles can be classified as nonparametric with respect to distributional assumptions or as fully parametric. We propose a semiparametric method where we model the mean and variance as flexible regression spline functions and allow the unspecified distribution to vary smoothly as a function of covariates. Similarly to Cole and Green, our approach provides separate estimates and summaries for location, scale and distribution. However, similarly to Koenker and Bassett, we do not assume any parametric form for the distribution. Estimation for either cross-sectional or longitudinal samples is obtained by using estimating equations for the location and scale functions and through local kernel smoothing of the empirical distribution function for standardized residuals. Using this technique with data on weight, height and age for females under 3 years of age, we find that there is a close relationship between quantiles of weight for height and age and quantiles of body mass index (BMI = weight/height^2) for age in this cohort.

Keywords: Empirical distribution; Estimating equations; Kernel smoothing; Reference quantiles; Standardizing weight

1. Introduction

Diagnostic classification often relies on the comparison of an observed measurement with 'normal' or reference ranges. A simple example is the classification of an adult as obese if their body mass index BMI, defined as weight/height^2, exceeds the 85th percentile of 20–29-year-old adults based on a national survey (National Center for Health Statistics, 1987). Here the evaluation of an individual's weight depends on the ability to standardize for concomitant variables, in particular gender and height. The standardization depends on the availability of age- and gender-specific quantiles of weight for an appropriate reference population. Standardization of measures using reference population quantiles is intuitive and commonly used in anthropometric research (Cole, 1988). Consider that a child's height varies naturally with both age and gender. The National Center for Health Statistics (NCHS) produces age- and gender-specific growth charts (Hamill *et al.*, 1977) which display percentiles of height.

Address for correspondence: Patrick J. Heagerty, Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, USA.
E-mail: heagerty@biostat.washington.edu

 0035–9254/99/48533

These percentiles are widely used clinically to quantify a child's size relative to a reference population. Longitudinal monitoring of height and weight percentiles are used clinically to diagnose growth abnormalities.

Standardization of measures by using reference population quantiles is also important in data analysis. Measures at the same quantile but for different covariate values are often considered comparable. Rather than using a raw measurement Y in data analysis, one generates a standardized measure Y^*, which is the covariate-specific percentile corresponding to Y in a reference population. Equivalently, any appropriate transformation of the percentile, such as the z-score (normal deviate corresponding to the percentile), can be used as a standardized measure and used in data analysis. Thus, for example, a simple comparison of height for two groups of children can be made when each group contains children of various ages and genders by conversion of their raw heights into age- and gender-specific percentiles or into z-scores.

This paper describes a new semiparametric method for constructing covariate-specific quantiles (i.e. regression quantiles) in a reference population. The method allows percentiles to be estimated as a smooth function of covariates without imposing parametric distributional assumptions. We evaluate the methodology by an application to the standardization of weight for both height and age for female infants under 3 years of age. Of particular interest for our application are regression quantile estimation methods that can be used with multiple covariates.

2. Regression quantiles

2.1. Definition

The probability distribution of a random quantity Y in a population is described by its quantiles. For $\alpha \in [0, 1]$ the αth quantile of Y is defined as the value Y^α such that

$$P(Y \leqslant Y^\alpha) = \alpha.$$

For example, if Y has a standard normal distribution, the 0.95 quantile, or 95th percentile, is 1.64. If covariates denoted by X are relevant, then covariate-specific quantiles may be of interest. The αth regression quantile is defined as the function $Y^\alpha(X)$ such that for all x

$$P\{Y \leqslant Y^\alpha(x) | X = x\} = \alpha.$$

A familiar use of regression quantiles is in the clinical assessment of childhood height and weight.

2.2. Bin and smooth estimation

Perhaps the most widely implemented method for estimating regression quantiles is to divide the domain of X into narrow intervals, to calculate the empirical quantiles within each interval and to connect the empirical quantiles across the values of X in some smooth way. We call this the 'bin and smooth quantiles' approach. The NCHS-published percentile curves for height were calculated in this fashion (Hamill *et al.*, 1977). For each gender, subjects from national health surveys were grouped into narrow intervals of age, and empirical percentiles of height were calculated for each age and sex group. Regression splines were fitted to the empirical quantiles to yield smooth percentile curves.

A related approach involves modelling the conditional distribution of Y given X parametrically with a parameter vector θ. The parameter is estimated separately within

each interval of X and smoothed across X to derive $\theta(X)$. This approach can be called the 'bin and smooth parameters' approach. For example, the model may be that Y is normally distributed with mean and variance that are smooth functions of X. Here $\theta(X) = \{\mu(X), \sigma(X)\}$ and the resulting regression quantiles are then calculated as $Y^{\alpha}(X) = \mu(X) + \sigma(X)z^{\alpha}$ where z^{α} is the quantile for the standard normal distribution. The method of Cole (1990) is a prime example of the bin and smooth parameters approach. In Cole's model there are three parameters characterizing the location, scale and skewness of the distribution.

The bin and smooth approaches are intuitive and quite straightforward to implement. They also appear quite flexible. The main drawback of bin and smooth methods is that they require substantial sample sizes within each covariate interval. Since covariate intervals must be narrow for the method to work, there will usually be a large number of such intervals and hence the overall sample size required can be very large. When only a single continuous covariate is involved, data sets such as those derived from the National Health and Nutrition Examination Survey (NHANES) can yield regression quantiles by using the bin and smooth approach. However, when more than one continuous covariate is involved, the covariate intervals become regions in multidimensional space and data can become sparse in some intervals even with very large data sets. As an example, bin and smooth approaches would probably not be feasible for calculating quantiles of weight for height and age, even with sample sizes of the order of those obtained by combining different cycles of the NHANES.

2.3. Parametric and nonparametric estimation

Some more sophisticated approaches to regression quantile estimation have appeared in the statistical literature. These methods have primarily either assumed that a parametric model for the complete conditional distribution $f(Y_i|X_i)$ could be estimated (Cole and Green, 1992) or have made no distributional assumptions and estimated each regression quantile $Y^{\alpha}(X)$ separately for all desired values of α (Koenker and Bassett, 1978; Efron, 1991).

Cole (1990) introduced a parametric method that has been termed the '*LMS*' method since the approach is based on three functions of the covariates: $M(X)$ the median response, $S(X)$ the approximate coefficient of variation and $L(X)$ the Box–Cox power transformation required to achieve normality. This method assumes that the transformation of Y defined by

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

yields a random variable Z with a standard normal distribution. The key advantages of this method are that summaries are obtained through the functions L, M and S and that all quantiles are obtained from these functions with

$$Y^{\alpha}(X) = M(X)\{1 + z^{\alpha}\,L(X)\,S(X)\}^{1/L(X)},$$

where z^{α} is the αth percentile of the standard normal distribution. Model building focuses on the specification of the form of the L-, M- and S-functions. Cole and Green (1992) discussed flexible smoothing spline approaches. Possible disadvantages of the *LMS* method include the strong assumption that a transformation to normality can be achieved for each value of X and the sensitivity of the choice of transformation $L(X)$ to outliers in the data (Carroll, 1982).

A distributionally nonparametric approach introduced by Koenker and Bassett (1978) is based on M-estimation similarly to least absolute deviation methods. An estimate for a regression quantile is the function $\hat{Y}^{\alpha}(X)$ that minimizes

$$\sum_i \rho_\alpha\{Y_i - Y^\alpha(X_i)\}$$

where $\rho_\alpha\{x\} = \alpha x^+ + (1-\alpha)x^-$, $x^+ = \max(0, x)$ and $x^- = \max(0, -x)$. This approach yields consistent estimates of the regression quantiles under general conditions, without requiring that the form of the distribution of Y_i be specified. However, a major drawback is that a separate specification and estimation are required for each α of interest. Without special restriction, the estimates $\hat{Y}^\alpha(X_i)$ need not be monotone in α, although the true quantiles $Y^\alpha(X_i)$ are defined to be so. Finally, this method requires specialized software based on linear programming algorithms (though they are publicly available).

In an effort to modify the Koenker and Bassett algorithm to ensure that quantiles would not cross, He (1997) employed a location–scale model for Y:

$$Y = \mu(X) + \sigma(X)\epsilon$$

where the location and scale functions were defined as the median μ and median absolute deviation σ respectively. The estimation of these functions uses linear programming techniques. He's formulation does not allow the distribution of ϵ to depend on X. Our approach generalizes this model by allowing the distribution of ϵ to vary as a function of the covariates.

2.4. Our semiparametric estimator

Our approach to regression quantiles is quite simple and relies on the generic representation

$$Y_i = \mu(X_i) + \sigma(X_i)\,\epsilon(X_i). \tag{1}$$

Here $\mu(X_i)$ is a location function, $\sigma(X_i)$ is a scale function and $\epsilon(X_i)$ is a random variable that determines the distribution of Y_i. We further assume that $\epsilon(X_i)$ has zero mean and unit variance. This assumption identifies $\mu(X_i)$ as $E(Y_i|X_i)$ and $\sigma(X_i) = \sqrt{\text{var}(Y_i|X_i)}$. The quantiles of Y_i are then completely determined by $\mu(X)$, $\sigma(X)$ and the base-line distribution function

$$F_0(z, X) = P\{\epsilon(X) \leqslant z|X\}.$$

Since we can write $\epsilon(X_i) = \{Y_i - \mu(X_i)\}/\sigma(X_i)$, we can interpret F_0 as the distribution for standardized values of Y_i.

In Section 3 we provide details on methods for estimating the location $\mu(X)$, the scale $\sigma(X)$ and the distribution $F_0(z, X)$ as smooth functions of the covariates X. These estimates in turn provide an estimate for any quantile, $Y^\alpha(X)$.

Our methods combine the strengths of both the parametric *LMS* method (Cole, 1988; Cole and Green, 1992) and the distributionally nonparametric methods of Koenker and Bassett (1978). Similarly to Cole (1988), we provide three functions that summarize the complete distribution, namely $\mu(X)$, $\sigma(X)$ and $F_0(z, X)$, and these determine all the regression quantiles. However, similarly to Koenker and Bassett (1978), we do not require the specification of a parametric base-line distribution.

A full description of the semiparametric approach is presented in Section 3. In Section 4 the method is applied to the problem of standardizing BMI for age where it is shown to have certain advantages over other approaches. Section 5 deals with a more complex example in which two continuous standardizing covariates are used. Quantile estimation for multidimensional covariates is rarely (if ever) considered in the applied statistical literature and we address some of the technical challenges that arise. Some interesting substantive conclusions are obtained concerning the relationship between weight and height in young females.

3. Semiparametric estimation of regression quantiles

To estimate the regression quantiles, i.e. the quantiles of Y conditional on X, we model the conditional distribution as being derived from a location–scale family with location and scale being smooth functions of X. Our representation is given as

$$Y_i = \mu(X_i) + \sigma(X_i)\,\epsilon(X_i)$$

where $\mu(X)$ and $\sigma(X)$ are the location and scale functions and

$$F_0(z, X) = P\{\epsilon(X) \leqslant z|X\}$$

is the base-line distribution function.

3.1. Estimation of $\mu(X)$ and $\sigma(X)$

We model $\mu(X)$ and $\sigma(X)$ parametrically by using regression splines. Specifically,

$$\mu(X) = \sum_{k=1}^{p} \beta_k\, R_k(X),$$

$$\log\{\sigma(X)\} = \sum_{k=1}^{q} \gamma_k\, S_k(X)$$

where $\mathcal{R}(X) = \{R_1(X), R_2(X), \ldots, R_p(X)\}$ and $\mathcal{S}(X) = \{S_1(X), S_2(X), \ldots, S_q(X)\}$ are regression spline basis functions. Many options are available to construct basis sets and the interested reader may refer to Dierckx (1995) for a detailed description of spline methodology. In the application that follows in Section 4, X is unidimensional and we focus attention on natural splines where the basis elements are determined by the selection of a finite number of points, or 'knots'. Any linear combination of natural spline basis elements yields a piecewise cubic polynomial with the additional constraint that the function is linear beyond the extreme covariate values, $\min(X_i)$ and $\max(X_i)$. We use the function `ns( )` provided by the statistical package S-PLUS (MathSoft, 1996) to obtain the bases $\mathcal{R}(X)$ and $\mathcal{S}(X)$ for chosen knots $\{r_1, r_2, \ldots, r_{p-2}\}$ and $\{s_1, s_2, \ldots, s_{q-2}\}$ respectively.

When X is multidimensional, a set of basis functions can be calculated by first calculating basis functions for each univariate component of X and then constructing a multivariate basis by including the univariate basis elements and the possible pairwise interactions (products) of the separate basis sets. Although the use of nonparametric or flexible parametric methods in higher dimensions can be difficult, we show in an application in Section 4 how this can be achieved.

It is important to recognize that specifying $\mu(X)$ and $\sigma(X)$ as regression splines with fixed knots is to specify them as parametric functions. This permits the use of standard quasi-likelihood methods for estimation (Heyde, 1997). Similarly to Davidian and Carroll (1987) and Smyth (1989) we adopt the estimating equations

$$\sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \beta} \mu(X_i) \right\}^{\mathrm{T}} \frac{Y_i - \mu(X_i)}{\mathrm{var}(Y_i)} = 0,$$

$$\sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \gamma} \sigma^2(X_i) \right\}^{\mathrm{T}} \frac{\{Y_i - \mu(X_i)\}^2 - \sigma^2(X_i)}{\mathrm{var}[\{Y_i - \mu(X_i)\}^2]} = 0,$$

the solution of which defines the estimators $\hat{\beta}$ and $\hat{\gamma}$. Use of the identity link for the mean function and a log-link for the variance function results in

$$\sum_{i=1}^{n} \mathcal{R}(X_i)^{\mathrm{T}}\{Y_i - \mu(X_i)\}/\sigma^2(X_i) = 0,$$

$$\sum_{i=1}^{n} \mathcal{S}(X_i)^{\mathrm{T}}[\{Y_i - \mu(X_i)\}^2 - \sigma^2(X_i)]/\sigma^2(X_i) = 0$$

where we adopt the Gaussian higher moment relationship $\mathrm{var}[\{Y_i - \mu(X_i)\}^2] = 2\,\mathrm{var}(Y_i)^2$. These estimating equations are score equations under the assumption of normality, yet they provide valid estimation more generally as a straightforward application of quasi-likelihood methods. Standard errors for the estimates $\hat{\beta}$ and $\hat{\gamma}$, the solution to these estimating equations, can be obtained for either longitudinal samples or cross-sectional samples by using an empirical variance estimator (Huber (1981), pages 127–135, and Liang and Zeger (1986)).

In practice, the choices of $\mathcal{R}(X)$ and $\mathcal{S}(X)$ are important. We have chosen to use natural splines and have varied the number and placement of the knots as a sensitivity analysis. Although the focus of our method is on the regression quantiles, testing parameters is possible using standard quasi-likelihood methods since $\mu(X)$ and $\sigma(X)$ are entirely parametric (Heyde, 1997). In the following section we describe methods for the estimation of the base-line distribution F_0 which is useful in revealing misspecifications of either $\mu(X)$ or $\sigma(X)$.

3.2. Estimation of F_0

Under our model the αth regression quantile is

$$Y^{\alpha}(X) = \mu(X) + \sigma(X)\, Z^{\alpha}(X)$$

where $Z^{\alpha}(X)$ is the αth quantile of $\epsilon(X)$. This quantile is determined by F_0:

$$\alpha = F_0\{Z^{\alpha}(X),\, X\}.$$

We leave the base-line distribution function F_0 unspecified. However, since we model $\mu(X)$ and $\sigma(X)$ as smooth functions of the covariates, it is desirable also to assume that $F_0(z,\, X)$ varies smoothly in X. We base the estimation of F_0 on the standardized residuals, applying kernel weighting to allow the estimate $\hat{F}_0(z,\, X)$ possibly to vary over the domain of X.

If the distribution of $\epsilon(X_i) = \{Y_i - \mu(X_i)\}/\sigma(X_i)$ did not vary as a function of X_i then the empirical distribution function

$$\hat{F}_0(z,\, X) = \sum_i \frac{1}{n}\, \mathbf{1}(\hat{e}_i \leqslant z),$$

where $\hat{e}_i = \{Y_i - \hat{\mu}(X_i)\}/\hat{\sigma}(X_i)$, would be a natural estimator for F_0. This important special case assumes that F_0 does not depend on X. However, when this assumption appears violated then alternative more flexible methods should be considered.

If a large number of observations were available for each covariate value X_j, then a restricted empirical distribution function could be used:

$$\hat{F}_0(z,\, X = X_j) = \sum_{i:\ X_i = X_j} \frac{1}{n_j}\, \mathbf{1}(\hat{e}_i \leqslant z)$$

where n_j is the number of observations with $X_i = X_j$. However, the primary motivation for regression modelling is that an insufficient number of observations are available for any X_j, this being particularly true for multivariate X. An intuitive solution is not only to consider

observations where $X = X_j$ but also to consider those observations that are 'close', perhaps giving weight to observations according to their distance to the value of interest. This leads to

$$\hat{F}_0(z, X; \lambda_1) = \sum_i \frac{1}{W(X)} w_{\lambda_1}(X, X_i) \mathbf{1}(\hat{e}_i \leqslant z)$$

where $W(X) = \Sigma\, w_{\lambda_1}(X, X_i)$ for a weight function w depending on a tuning parameter λ_1. Without loss of generality, we assume a Gaussian kernel function

$$w_{\lambda_1}(X, X_i) = \exp\left\{ -\frac{1}{2}\left(\frac{X - X_i}{\lambda_1} \right)^2 \right\}.$$

The empirical distribution function is a special case of this more general class where $w_{\lambda_1}(X, X_i) = 1$ (corresponding to $\lambda_1 \to \infty$).

If we use $w_{\lambda_1}(X, X_i) = 1$, then the quantile function $\hat{Z}^{\alpha}(X)$ is monotone in α and constant as a function of X. If local weighting is used then $\hat{Z}^{\alpha}(X)$ will be a discontinuous function of X having support only on $\{\hat{e}_i\}_{i=1}^n$. We usually desire regression quantiles that are strictly continuous in X and we may therefore choose to replace the indicator function used in the empirical distribution function with a continuous distribution function:

$$\hat{F}_0(z, X; \lambda_1, \lambda_2) = \sum_i \frac{1}{W(X)} w_{\lambda_1}(X, X_i) K_{\lambda_2}(z, \hat{e}_i)$$

where $K_{\lambda_2}(z, \hat{e}_i)$ is a continuous distribution function in z that may depend on a second tuning parameter λ_2. Without loss of generality we assume $K_{\lambda_2}(z, \hat{e}_i) = \Phi\{(z - \hat{e}_i)/\lambda_2\}$. This approach amounts to using locally weighted kernel density estimation.

In our applications, we have used kernel weight functions w_{λ_1} that specify either a fixed distance λ_1 or have a variable distance $\lambda_1(X)$ based on obtaining a fixed proportion of the observations for estimation at each unique value of X. The first approach is similar to fixed window width methods whereas the latter is similar to using a fixed number of nearest neighbours (see Härdle (1991), pages 42–48). For a single covariate the metric used for local weighting is naturally $|X - X_i|$. However, for multiple covariates, there can be several choices. In the example discussed in Section 5, we used $X = (X_1, X_2)$ and explored the variation in $\hat{F}_0$ as a function of X_1 alone, as a function of X_2 alone and as a function of the combination $\hat{\mu}(X_1, X_2)$.

The choice of K_{λ_2} is much less critical than that of w_{λ_1} and only of interest if the resulting discontinuity from the use of $\mathbf{1}(\hat{e}_i \leqslant z)$ is a concern. Further research into the possible advantages and trade-offs among members of the class $\hat{F}_0(z, X; \lambda_1, \lambda_2)$ is warranted.

4. Standardizing the body mass index for age and gender

There has been increased interest recently in the potential for childhood measures of obesity to predict the risk for adult obesity. We recently reported on an observational longitudinal study of obesity from birth to adulthood in subjects born between 1965 and 1971 at Group Health Cooperative, a large health maintenance organization in Seattle, Washington, USA (Whitaker *et al.*, 1997). All lifetime height and weight measurements (excluding measurements during pregnancy or during emergency room visits) were abstracted from the outpatient medical records of study subjects. Full details about the data can be found in Whitaker *et al.* (1997).

BMI is considered to be a measure of weight standardized for height in young adults and is used to define obesity. In adults, obesity based on BMI is defined by BMI $\geqslant 27.8$ kg m^{-2} for

males and BMI $\geqslant 27.3$ kg m^{-2} for females (National Institutes of Health Consensus Development Panel, 1985). In our study, the childhood predictors of adult obesity (CPAO) study, we used BMI as a measure of weight standardized for height for children in addition to adults. However, BMI varies dramatically with age and gender in children and thus needs to be standardized for these concomitant variables for analysis. There are, unfortunately, no detailed published reference quantiles of BMI for children from 1 to 36 months of age in the USA, although reference quantiles for British children have been published (Cole *et al.*, 1995).

We next calculate quantiles of BMI for age and gender using data on 1–36-month-old female children in the CPAO study itself. Our main purpose is to illustrate our regression quantile estimation method on real data and to compare our method with some alternatives. Secondarily, it will allow us later to address briefly the question of whether or not BMI adequately standardizes weight for height in young children.

4.1. Methods and results

For brevity, we present results only for females since the analysis for males yielded qualitatively similar results. The data are from $N = 540$ subjects with an average of 6.3 measurements per child (3425 observations in total). Clustering of data at 6 weeks and at 3, 6, 9, 12, 18 and 24 months is due to the schedule of well child visits and vaccinations offered by the health care facility at the time that the study subjects were 1–36 months old during 1965–1973. The semiparametric quasi-likelihood procedure for quantile estimation was applied with Y being BMI and the covariate X being age. The mean function $\mu(X)$ was modelled as a natural cubic regression spline with knots at 4, 9 and 18 months and $\log\{\sigma(X)\}$ had knots at 6 and 18 months. Fig. 1 shows the estimated mean $\hat{\mu}(X)$ and the estimated standard deviation $\hat{\sigma}(X)$. 95% pointwise confidence bands were calculated based on

$$\mathrm{se}\{\hat{\mu}(X)\} = \mathrm{se}\{\mathcal{R}(X)\hat{\beta}\} = \mathcal{R}(X)\,\mathrm{se}(\hat{\beta})\,\mathcal{R}(X)^{\mathrm{T}}$$

and

$$\mathrm{se}[\log\{\hat{\sigma}(X)\}] = \mathrm{se}\{\mathcal{S}(X)\hat{\gamma}\} = \mathcal{S}(X)\,\mathrm{se}(\hat{\gamma})\,\mathcal{S}(X)^{\mathrm{T}}.$$

The standard errors for $(\hat{\beta}, \hat{\gamma})$ were based on a 'sandwich estimator' that accounts for the correlation of repeated measurements (Huber, 1981; Liang and Zeger, 1986).

The choice of knots is an issue that arises in modelling the parametric functions as regression splines. Data-driven automatic procedures are available in certain circumstances such as for likelihood-based estimation as described by Kooperberg *et al.* (1995). However, our data involve multiple observations per subject and the probability models are marginal (i.e. cross-sectional) so the estimation based on estimating equations is not likelihood estimation. Thus, automatic algorithms for choosing knots do not appear to be available at this time. We take a pragmatic approach. We believe that the mean and standard deviation functions are very smooth functions of age and we choose the number and location of the knots to provide adequate flexibility for our application. Sensitivity of the results to the placement of the knots can be assessed by changing them and recalculating the key model summaries. We compared parameter estimates for $\mu(X)$ and $\sigma(X)$ that both included more knots and different knot placements and found little difference in the estimates or their standard errors.

Fig. 2(a) shows two estimates of the base-line distribution function $F_0(z, X)$. The broken lines display $\hat{F}_0\{z, X; \lambda_1(X)\}$ based on Gaussian kernel weighting

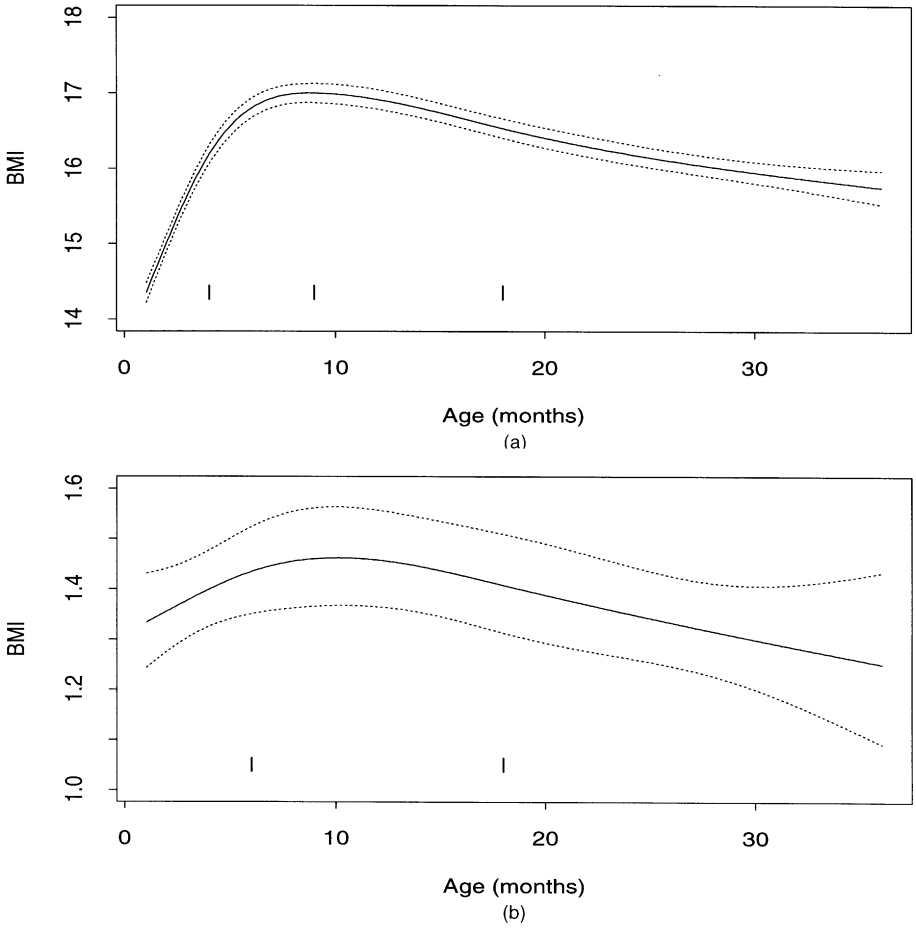

Fig. 1. (a) Mean and (b) standard deviation of BMI as a function of age for girls between 1 and 36 months in the CPAO study: the functions were estimated with natural regression splines with knots placed at locations denoted by the vertical tick marks (·····, pointwise 95% confidence intervals)

$$w_{\lambda_1}(X, X_i) = \exp\left[-\frac{1}{2}\left\{\frac{X - X_i}{\lambda_1(X)}\right\}^2\right],$$

where the tuning parameter $\lambda_1(X)$ is chosen such that 15% of the covariate values X_i fall within $\lambda_1(X)$ units of X. This allows a stable number of observations to contribute to the estimation of $\hat{F}_0$ without oversmoothing in regions that have many observations. Fig. 2(a) also plots the quantiles of the empirical distribution function estimator which assumes that the distribution of standardized values, $F_0(z, X)$, is constant over X. We see that the assumption that F_0 does not depend on X is plausible for these data. Comparing the empirical distribution of the standardized residuals with a standard normal distribution reveals some right skewness, as seen in the QQ-plot in Fig. 2(b).

Fig. 3 shows the fitted percentiles of the BMI assuming either a constant base-line distribution or a variable base-line distribution. These two semiparametric estimates are in good agreement for all the percentiles although there are slight differences for the 99th percentile.

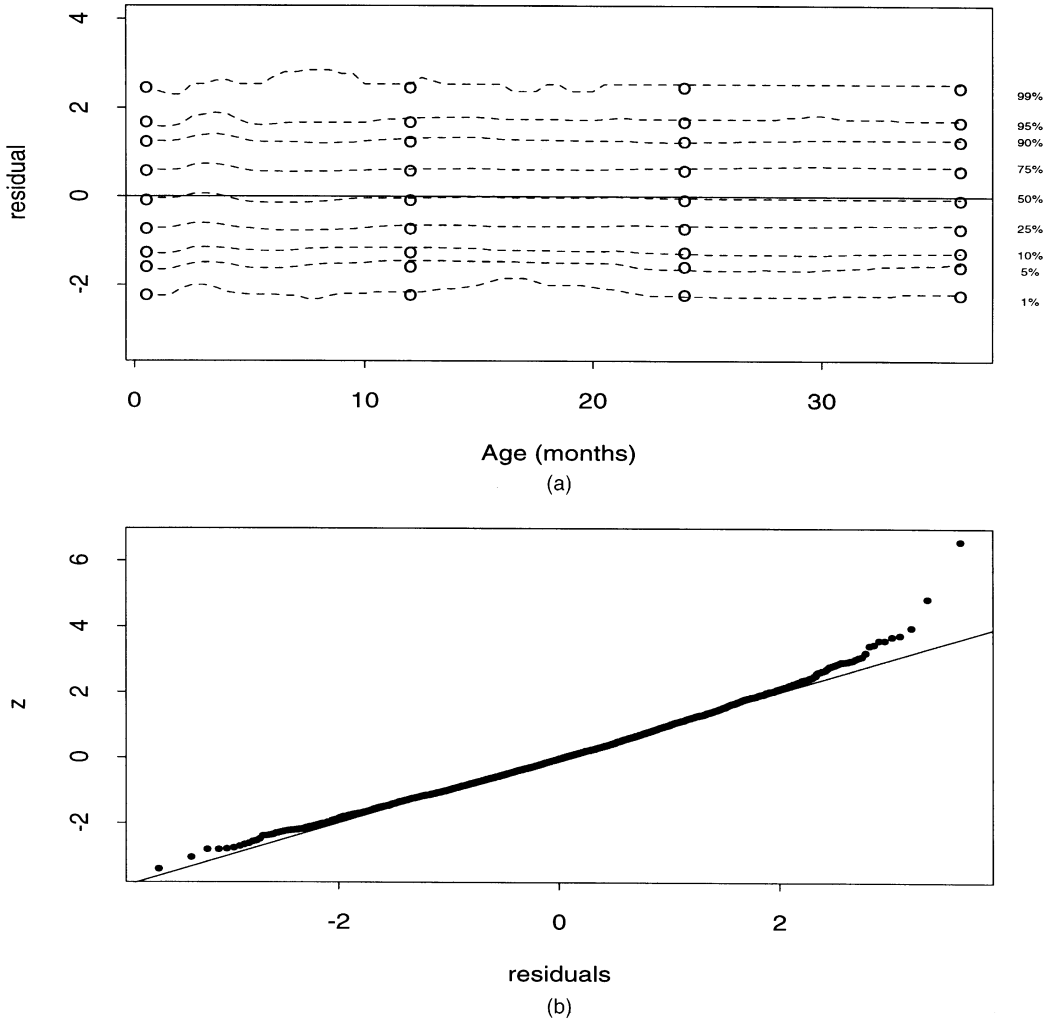

Fig. 2. (a) Estimated quantiles of F_0 based on the kernel estimation method using a Gaussian weight function with standard deviation $\lambda_1(X)$, selected to contain 15% of the observations (- - - - -), and the quantiles of the empirical distribution of the standardized residuals, $\hat{e}_i = \{Y_i - \hat{\mu}(X)\}/\hat{\sigma}(X)$ (○), and (b) normal *QQ*-plot for the residuals

To compare our method with alternative approaches, we used adaptations of Koenker and Bassett (1978) and Cole and Green (1992) that also used natural spline functions. Fig. 4(a) shows the fitted quantiles using the *LMS* method where we model

$$M(X) = \sum_{j=1}^{p} \beta_j R_j(X),$$

$$S(X) = \sum_{j=1}^{q} \gamma_j S_j(X)$$

and

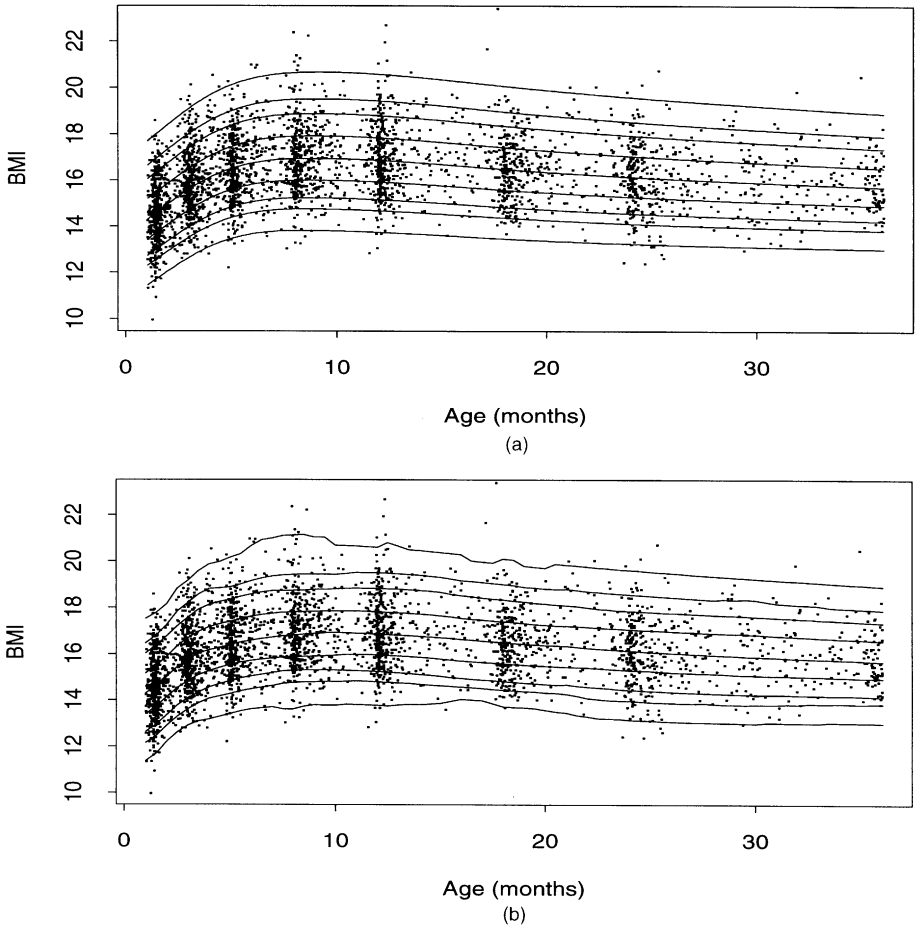

Fig. 3. Estimated percentiles of BMI as a function of age based on the semiparametric quasi-likelihood method (shown from the bottom to the top are the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles, and the raw data): (a) F_0 assumed to be constant as a function of X; (b) F_0 allowed to depend on X

$$L(X) = \sum_{j=1}^{r} \delta_j T_j(X)$$

where $\mathcal{R}(X) = \{R_j(X); j = 1, 2, \ldots, p\}$, $\mathcal{S}(X) = \{S_j(X); j = 1, 2, \ldots, q\}$ and $\mathcal{T}(X) = \{T_j(X); j = 1, 2, \ldots, r\}$ are natural spline basis functions. We used natural splines with knots at 4, 9 and 18 months for each of $\mathcal{R}(X)$, $\mathcal{S}(X)$ and $\mathcal{T}(X)$. In this example the mean function $\hat{\mu}(X)$ and the function $\hat{M}(X)$ are quite similar, and the function $\hat{S}(X)$ approximates $\hat{\sigma}(X)/\hat{\mu}(X)$ (data not shown). The major difference in terms of summaries provided by the *LMS* method and the semiparametric method is that the *LMS* method provides distribution information via the transformation function $\hat{L}(X)$ whereas we provide the estimated base-line distribution function $\hat{F}_0(z, X)$ (Fig. 2(a)). For this example the two approaches provide similar quantile estimates, seen by comparing Fig. 3(a) with Fig. 4(a).

Fig. 4(b) shows the results of fitting individual quantiles by using the method of Koenker and Bassett (1978) adapted for use with parametric splines (Hendricks and Koenker, 1992). For each regression quantile $Y^{\alpha}(X)$, we used a natural spline function

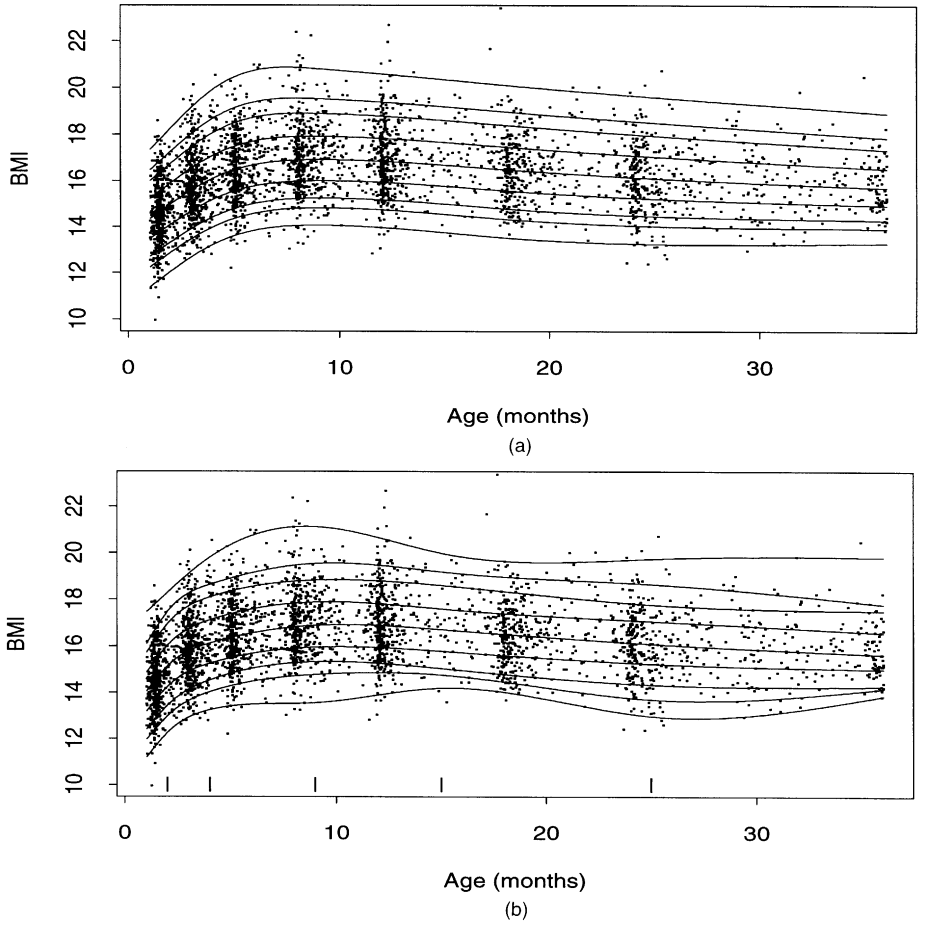

Fig. 4. Estimated percentiles of BMI as a function of age based on (a) the *LMS* method of Cole (1988) and (b) the method of Koenker and Bassett (1978) (shown from the bottom to the top are the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th and 99th percentiles, and the raw data)

$$Y^{\alpha}(X) = \sum_{j=1}^{p} \beta_j^{\alpha} R_j(X)$$

where $\mathcal{R}(X) = \{R_j(X); j = 1, 2, \ldots, p\}$ is a natural spline basis with knots at 2, 4, 9, 15 and 25 months. Using this method requires that we obtain a separate coefficient $\widehat{\beta^{\alpha}}$ for each quantile of interest. The fitted quantiles in Fig. 4(b) are quite similar to those obtained from the semiparametric approach that allows the base-line distribution F_0 to depend on X (Fig. 3(b)). However, we also see in Fig. 4(b) that the M-estimation approach may result in crossing quantiles as is nearly the case for later ages.

In summary, for quantile estimation with a single covariate, the semiparametric method that we propose shares features of both the *LMS* method (Cole, 1988) and the distribution-free method of Koenker and Bassett (1978). The advantage of the semiparametric approach is that all quantiles are obtained from the three summaries mean, standard deviation and base-line distribution, yet the functional form of the base-line distribution need not be

specified parametrically. Flexibility is controlled by separate model choices for $\mu(X)$, $\sigma(X)$ and $F_0(z, X)$.

5. Standardizing weight for height and age

In the CPAO study, we used a subject's age-specific BMI percentile (relative to an external reference data set) as a measure of age- and height-adjusted weight. Details of how these calculations were performed can be found in Whitaker *et al.* (1997). In this section we calculate a potentially better measure of age- and height-adjusted weight, namely age- and height-specific percentiles of weight. Again, we illustrate using data for girls 1–36 months of age in the CPAO study.

One elegant approach to standardizing weight for both height and age is given by Cole (1979) where weight is first converted into weight-for-age z-scores and then regressed on height-for-age z-scores. The result is that only methods for a single covariate are needed. Our first analysis of these data used this approach but we were confronted with a multidimensional covariate when we chose to investigate whether the relationship between weight-for-age and height-for-age varied with age (i.e. use of the main effects for height-for-age and interactions with age). For simplicity we present results that use weight directly, using the height-for-age transformation as a convenient method for reducing covariate collinearity.

5.1. Methods and results

A technically challenging aspect of this analysis is that there are two continuous components to the covariate (age and height), and these covariates are highly correlated. Such correlation is known to be problematic in regression analysis. Moreover, an implication of this association in our setting is that it makes it difficult to define an appropriate set of regression spline basis functions on the bivariate domain $X = (\text{age}, \text{height})$. If the basis is taken to be the set of cross-products of basis functions for height and basis functions for age, many of the cross-products will have their non-zero values only outside the range of the observed data and hence will be irrelevant.

Our solution for this is to transform height to height-for-age specific percentiles or z-scores. We calculated regression quantiles for height as a function of age using the CPAO study data and by applying the semiparametric method. The resultant percentiles of height have approximately a uniform distribution in (0, 100%) at each age, thus transforming the domain of (height, age) to a rectangular region. We chose to use the z-score transformation

$$X_2 = \{\text{height} - \mu_h(\text{age})\}/\sigma_h(\text{age})$$

so that the boundaries of the rectangular region would be more diffuse. We refer to the transformation X_2 as the 'height-for-age z-score'. Thus with $Y =$ weight, $X_1 =$ age and $X_2 =$ height-for-age z-score, we calculated the conditional distributions of Y given (X_1, X_2) by using the semiparametric method.

The model for Y is that

$$\epsilon(X_i) = \{Y - \mu(X_{1i}, X_{2i})\}/\sigma(X_{1i}, X_{2i})$$

has an unspecified distribution F_0 with mean 0 and variance 1. The parameter functions μ and σ are modelled as

$$\mu(X_1, X_2) = \beta_0 + \sum_{j=1}^{p} \beta_j^{(1)} R_j^{(1)}(X_1) + \sum_{k=1}^{q} \beta_k^{(2)} R_k^{(2)}(X_2) + \sum_{jk} \beta_{jk}^{(12)} R_j^{(1)}(X_1) R_k^{(2)}(X_2),$$

$$\log\{\sigma(X_1, X_2)\} = \gamma_0 + \sum_{j=1}^{p'} \gamma_j^{(1)} S_j^{(1)}(X_1) + \sum_{k=1}^{q'} \gamma_k^{(2)} S_k^{(2)}(X_2)$$

where the functions $\mathcal{R}^{(1)} = \{R_j^{(1)}(\cdot), j = 1, 2, \ldots, p\}$ are natural cubic regression spline basis functions for age and $\mathcal{R}^{(2)} = \{R_k^{(2)}(\cdot), k = 1, 2, \ldots, q\}$ are basis functions for height-for-age z-score. We used knots at 4, 9 and 18 months for age ($\mathcal{R}^{(1)}$) and knots at -1 and 1 for height-for-age z-score ($\mathcal{R}^{(2)}$). The basis functions $\mathcal{S}^{(1)} = \{S_j^{(1)}(\cdot), j = 1, 2, \ldots, p'\}$ and $\mathcal{S}^{(2)} = \{S_k^{(2)}(\cdot), j = 1, 2, \ldots, q'\}$ were chosen to be the same as $\mathcal{R}^{(1)}$ and $\mathcal{R}^{(2)}$ respectively.

For $\hat{F}_0$ we chose to use the empirical distribution function of the standardized residuals

$$\hat{e}_i = \{Y_i - \hat{\mu}(X_{1i}, X_{2i})\}/\hat{\sigma}(X_{1i}, X_{2i})$$

after exploring possible variation as a function of the covariates. We used the locally weighted empirical distribution function $\hat{F}_0\{z, X; \lambda_1(X)\}$, where we considered distance based on age only, height only and a multivariate measure $\hat{\mu}(X_1, X_2)$, the estimated mean weight for age and height. Each of these resulted in quantiles Z^{α} that appeared constant (Fig. 5).

Using the empirical distribution function of the standardized residuals to estimate F_0 yields fitted quantiles of the form $\hat{\mu}(X_1, X_2) + \hat{\sigma}(X_1, X_2)\hat{Z}^{\alpha}$. Figs 6(a)–6(c) display quantiles of weight *versus* age for subjects at the 25th, 50th and 75th percentiles of height-for-age. Figs 6(d)–6(f) display quantiles of weight *versus* age-specific height percentiles when age is fixed at 6, 12 and 24 months. These charts can be used to assess a child's weight for age and height status as follows. Consider a 12-month-old female with a height of 70.9 cm and a weight of 8.64 kg. This child is short, being at the sixth percentile of height for her age (height-for-age z-score -1.50). However, given her height and age, she is not underweight, being at the 50th percentile of weight on the basis of Fig. 6(e).

To determine whether age-specific BMI percentiles standardize weight appropriately for age and height, we investigated the correspondence between BMI for age percentiles and weight for height and age (WHA) percentiles for subjects in the CPAO study (Fig. 7). The percentiles appear to be reasonably close, with the difference between the two being less than 10% for 98% of the observations. Larger differences were observed but only for children who were particularly short (below the fifth percentile of height for age) or particularly tall (above the 95th percentile of height for age). Among tall children the bias is always in the direction that the BMI percentile is less than the WHA percentile, i.e. BMI indicates that a tall child is more underweight for her height and age than she really is. For short children bias in either direction appears to be possible, although on average the BMI percentile is larger than the WHA percentile. There was no trend with age in the discrepancy between the BMI and WHA percentiles (data not shown). We conclude therefore that in the CPAO cohort the age-specific BMI percentile appears to be a reasonable measure of height- and age-adjusted weight in females 1–36 months of age, although for some very tall and some very short children it performs less well.

6. Discussion

We have proposed a new approach to regression quantile estimation which bears some relation to a method proposed recently by He (1997). However, our representation,

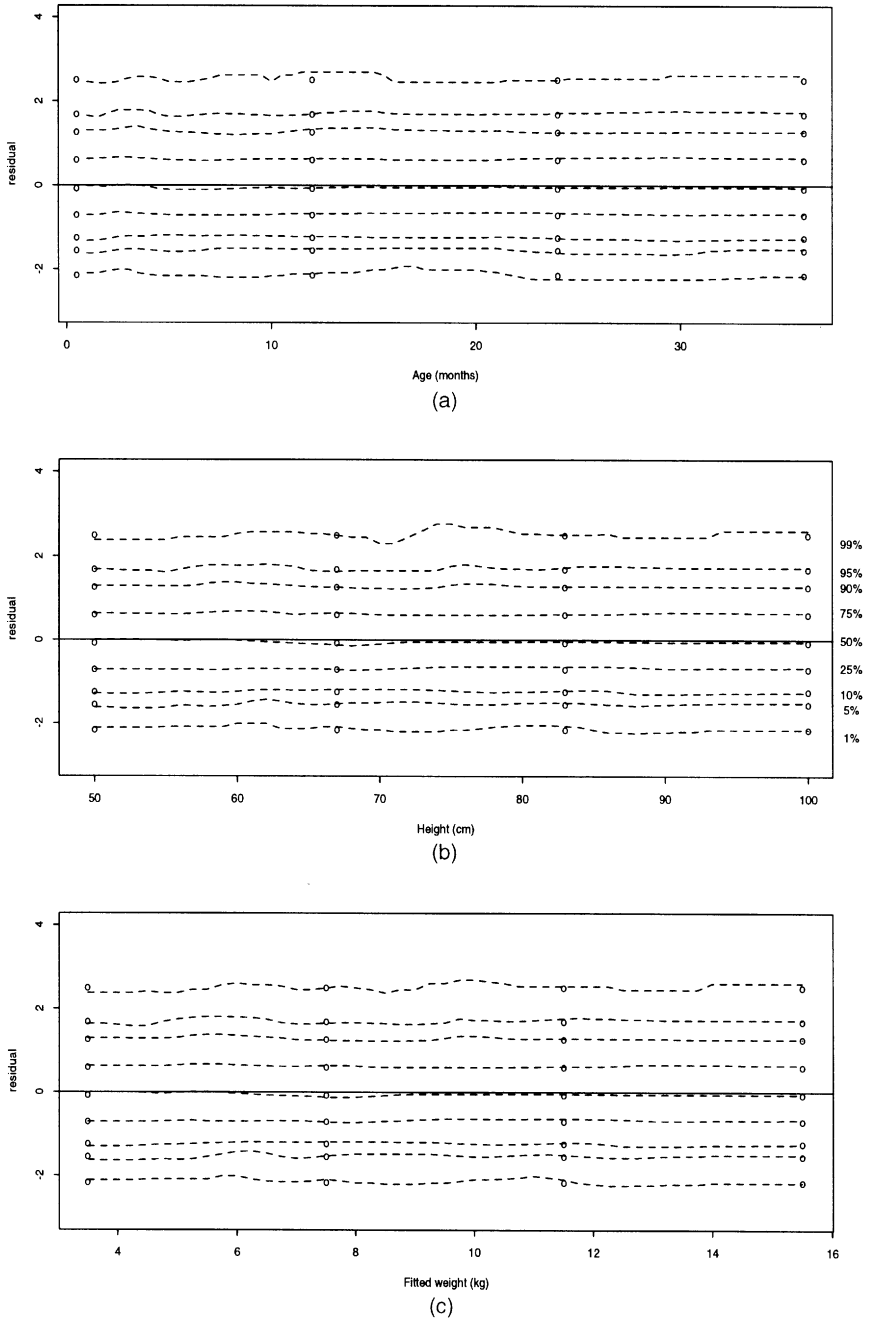

Fig. 5. Weight for age and height — base-line distribution: the locally weighted empirical distribution function $\hat{F}_0\{z, X; \lambda_1(X)\} = \Sigma_i \{1/W(X)\}\, w_{\lambda_1(x)}(X, X_i)\, \mathbf{1}(\hat{e}_i \leqslant z)$, where $\lambda_1(X)$ is chosen so that 15% of the observations are within $\lambda_1(X)$ units, is presented; the metric used by the weight function w_{λ_1} is distance based on (a) $X =$ age, (b) $X =$ height and (c) $X = \hat{\mu}(X_1, X_2)$ ($\bigcirc$, quantiles of the empirical distribution function)

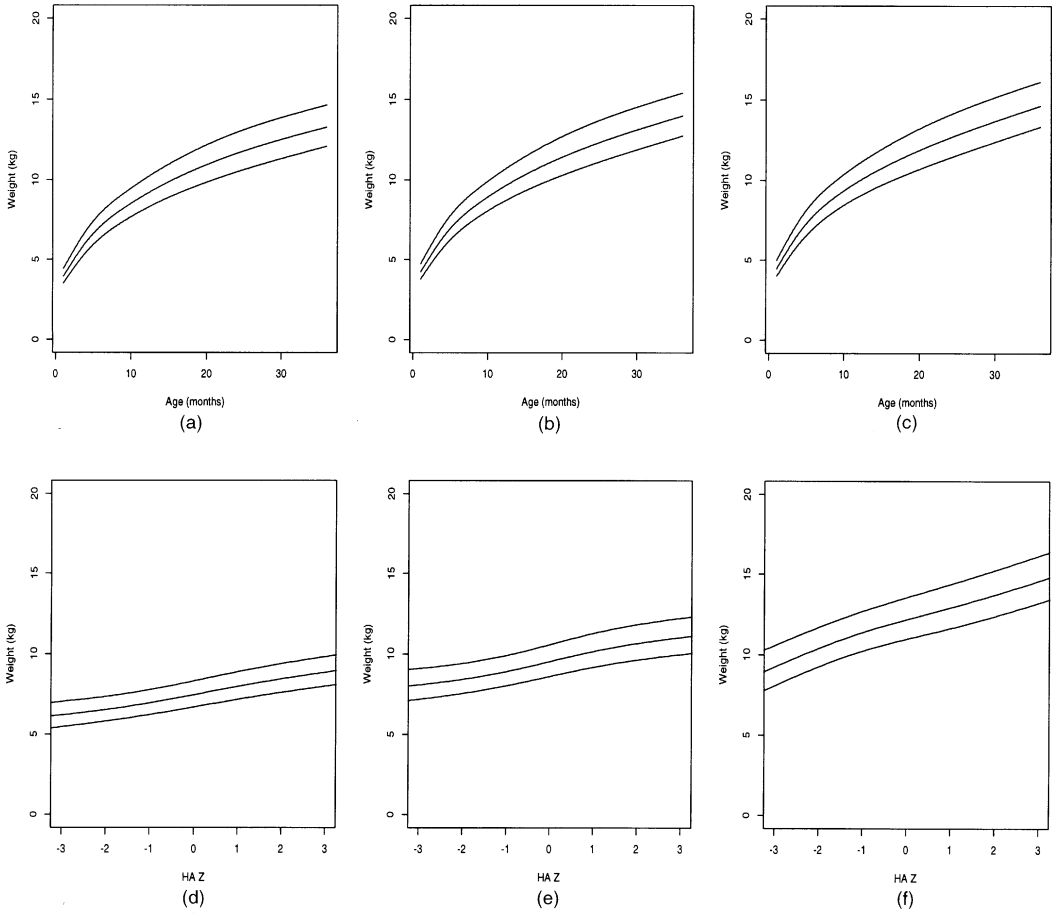

Fig. 6. Estimated age- and height-specific percentiles of weight: (a)–(c) percentiles of weight *versus* age for fixed height percentiles ((a) first quartile of height for age; (b) median of height for age; (c) third quartile of height for age); (d)–(f) percentiles of weight *versus* percentiles of height, using the height-for-age *z*-score (HAZ) for fixed ages ((d) 6 months; (e) 12 months; (f) 24 months) (in each part, from the bottom to the top, the 10th, 50th and 90th percentiles are shown)

$Y = \mu(X) + \sigma(X)\,\epsilon(X)$, generalizes to a broader class of models by allowing the base-line distribution $F_0(z, X) = P\{\epsilon(X) \leqslant z|X\}$ to depend also on covariates.

Although we have focused on natural spline models for the mean and the standard deviation functions, our general approach of decomposition into location, scale and base-line distribution can be used with other model specifications or other function estimators. For example, $\mu(X)$ could be defined as the median and estimated by using linear programming methods. Alternatively, smoothing splines could be used in place of the natural splines. Such choices would lead to additional computational effort but would be feasible modifications to our general semiparametric approach. Further research into the selection of the smoothing parameters and methods for formal inference regarding the base-line distribution is warranted.

It is appropriate to point out that direct kernel smoothing of Y (Ducharme *et al.*, 1995) may be difficult for regression quantile estimation because of the potential for large bias

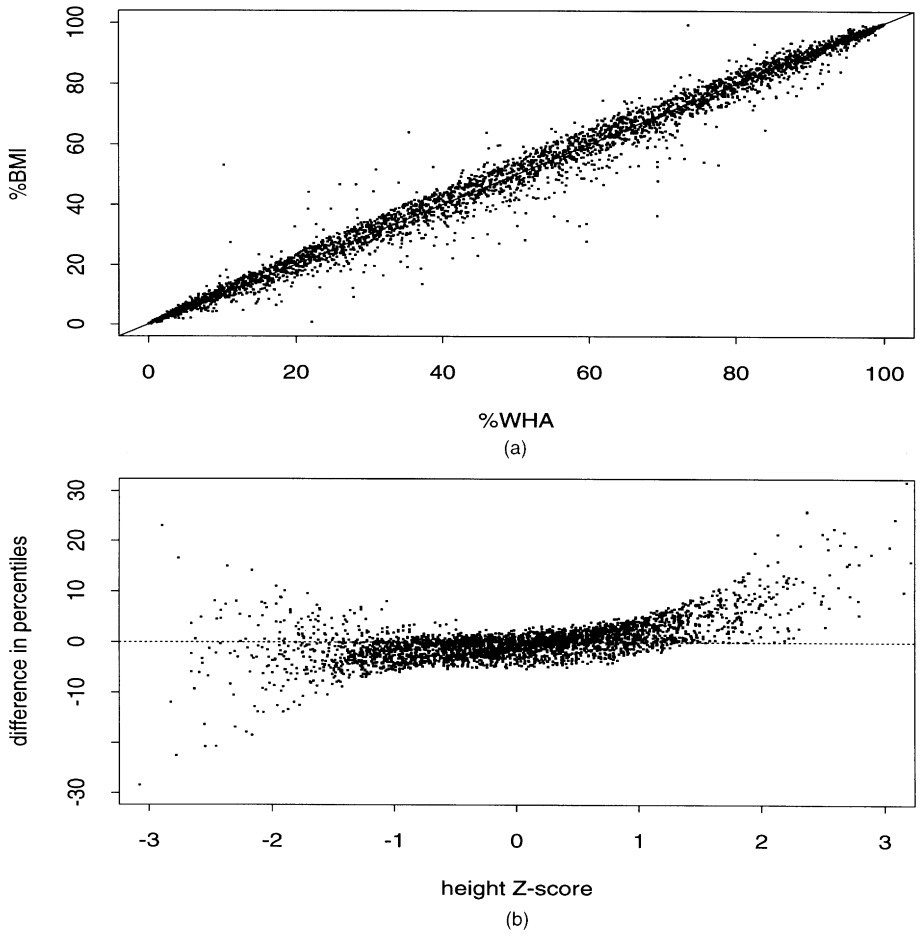

Fig. 7. Height- and age-specific percentile of weight (%WHA) and age-specific percentile of BMI (%BMI): (a) %BMI *versus* %WHA; (b) difference in percentiles (%WHA − %BMI) *versus* height-for-age *z*-score

introduced by local trends. The same concern arises with the bin and smooth methods where small bin sizes are required to avoid distortion in the estimation of percentiles due to increasing or decreasing trends within bins. For example, if the 0.05 percentile of weight for female infants, age 15 months and height 80 cm, was of interest then the empirical 0.05 percentile for a large number of female infants with age 15 ± 1 months and height 80 ± 5 cm may be adequate. However, if a larger bin was used, such as age 15 ± 10 months and height 80 ± 20 cm, then the empirical 0.05 percentile would probably be for a small young child (at the extreme of the interval: 5 months and 60 cm), providing a biased estimate for the desired age and height combination (15 months and 80 cm). By first estimating $\mu(X)$ and $\sigma(X)$ we greatly reduce the potential for this type of bias.

The approach of Healy and Rasbash (1988) avoids this potential bias by first detrending within intervals and then ranking the residuals to form quantile estimates. In this regard their method is similar to our semiparametric approach where we use the functions $\mu(X)$ and $\sigma(X)$ to centre and scale the responses, using the standardized residuals to estimate the base-line distribution.

We have compared our approach with two alternatives: the parametric approach of Cole (1988) and the nonparametric approach of Koenker and Bassett (1978). Both qualitative and quantitative comparisons were made. Quantitative comparisons are difficult because a direct comparison of the 'degrees of freedom' used by the different approaches is not possible. In our illustration in Section 3 (Figs 3 and 4), with the semiparametric method we used three knots for $\mu(X)$, two knots for $\sigma(X)$ and a nonparametric estimator for F_0, with Cole's method we used three knots for each of three functions and with the nonparametric method we used five knots for each of the regression quantiles estimated. All these approaches offer a high degree of model flexibility and each can be used in a parsimonious fashion by judicious restriction of this flexibility. The primary difference between these methods lies in the form of their model assumptions.

Though existing quantile regression methods are applicable to settings with multiple covariates, we are not aware of any previously published papers in which multiple covariates were involved. A methodological issue which arose in our application and which may arise in other applications involving multiple covariates concerns the restricted domain of the covariate space due to inherent correlations between the covariates age and height. A z-score transformation of height yielded a rectangular space which was more amenable to models using regression splines. This strategy might be useful in other applications also.

Our methodology might be applied to an array of biological parameters which are known to vary with height and age in children. These include, for example, measures of pulmonary function and blood pressure. A comparison of existing methods of standardization with age- and height-specific percentiles would be of interest. For example, pulmonary function is currently standardized by taking the ratio of observed pulmonary function to the median in a reference population with the same age, height and gender. Our methods would allow the calculation of age-, gender- and height-specific percentiles of pulmonary function and the evaluation of whether the per cent of median standardization algorithm has a close correspondence to the percentile approach.

Acknowledgements

We would like to thank Dr Bob Whitaker of the Children's Hospital Medical Center, Cincinnati, for insights on the clinical relevance of our work and for permission to use data from the CPAO study. Thanks are due to Linda Weistaner and Gary Longton for help in preparing the manuscript. Funding for this research was provided by National Institutes of Health grants R01 HL57288-01 and R01 GM54438, and by grant CF R565 from the Cystic Fibrosis Foundation.

References

Carroll, R. J. (1982) Two examples of transformations when there are possible outliers. *Appl. Statist.*, **31**, 149–152.

Cole, T. J. (1979) A method for assessing age-standardized weight-for-height in children seen cross-sectionally. *Ann. Hum. Biol.*, **6**, 249–268.

——— (1988) Fitting smoothed centile curves to reference data (with discussion). *J. R. Statist. Soc.* A, **151**, 385–418.

——— (1990) The LMS method for constructing normalized growth standards. *Eur. J. Clin. Nutrn*, **44**, 45–60.

Cole, T. J., Freeman, J. V. and Preece, M. A. (1995) Body mass index reference curves for the UK, 1990. *Arch. Dis. Childhd*, **73**, 25–29.

Cole, T. J. and Green, P. J. (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Statist. Med.*, **11**, 1305–1319.

Davidian, M. and Carroll, R. J. (1987) Variance function estimation. *J. Am. Statist. Ass.*, **82**, 1079–1091.

Dierckx, P. (1995) *Curve and Surface Fitting with Splines*. Oxford: Clarendon.
Ducharme, G. R., Gannoun, A., Guertin, M.-C. and Jéquir, J.-C. (1995) Reference values obtained by kernel-based estimation of quantile regressions. *Biometrics*, **51**, 1105–1116.
Efron, B. (1991) Regression percentiles using asymmetric squared error loss. *Statist. Sin.*, **1**, 93–125.
Hamill, P. V., Drizd, T. A., Johnson, C. L., Reed, R. B. and Roche, A. F. (1977) NCHS growth curves for children birth-18 years. In *Vital and Health Statistics*, ser. 11, no. 165. Washington DC: US Government Printing Office.
Härdle, W. (1991) *Applied Nonparametric Regression*. New York: Cambridge University Press.
He, X. (1997) Quantile curves without crossing. *Am. Statistn*, **51**, 186–192.
Healy, M. J. R. and Rasbash, J. (1988) Distribution-free estimation of age-related centiles. *Ann. Hum. Biol.*, **15**, 17–22.
Hendricks, W. and Koenker, R. (1992) Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Am. Statist. Ass.*, **87**, 58–68.
Heyde, C. C. (1997) *Quasilikelihood and Its Applications: a General Approach to Optimal Parameter Estimation*. New York: Springer.
Huber, P. (1981) *Robust Statistics*. New York: Wiley.
Koenker, R. and Bassett, G. (1978) Regression quantiles. *Econometrica*, **46**, 33–50.
Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995) Hazard regression. *J. Am. Statist. Ass.*, **90**, 78–94.
Liang, K.-Y and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
MathSoft (1996) *S-PLUS, Version 3.4*. Cambridge: MathSoft.
National Center for Health Statistics (1987) Anthropometric reference data and prevalence of overweight: United States, 1976–1980. In *Vital and Health Statistics*, ser. 11, no. 238. Washington DC: US Government Printing Office.
National Institutes of Health Consensus Development Panel on the Health Implications of Obesity (1985) Health implications of obesity: National Institutes of Health Consensus Development conference statement. *Ann. Intern. Med.*, **103**, 147–151.
Smyth, G. K. (1989) Generalized linear models with varying dispersion. *J. R. Statist. Soc.* B, **51**, 47–60.
Whitaker, R. C., Wright, J. A., Pepe, M. S., Seidel, K. D. and Dietz, W. H. (1997) Predicting adult obesity from childhood and parent obesity. *New Engl. J. Med.*, **337**, 869–873.