Published in final edited form as:

Stat Med. 2010 February 28; 29(5): 558–572. doi:10.1002/sim.3810.

Application of Nonparametric Quantile Regression to Body Mass Index Percentile Curves from Survey Data

Yan Li¹, Barry I. Graubard¹, and Edward L. Korn²

¹Biostatistics Branch, National Cancer Institute, Bethesda MD 20892

²Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892

Summary

Increasing rates of overweight among children in the U.S. stimulated interest in obtaining national percentile curves of body size to serve as a benchmark in assessing growth development in clinical and population settings. In 2000, the U.S. Centers for Disease Control and Prevention (CDC) developed conditional percentile curves for Body mass index (BMI) for ages 2–20 years. The 2000 CDC BMI-for-age curves are partially parametric and only partially incorporated the survey sample weights in the curve estimation. As a result, they may not fully reflect the underlying pattern of BMI-for-age in the population. This motivated us to develop a nonparametric double-kernel-based method and automatic bandwidth selection procedure. We include sample weights in the bandwidth selection, conduct median correction to reduce small-sample smoothing bias, and rescale the bandwidth to make it scale-invariant. Using this procedure we re-estimate the national percentile BMI-for-age curves and the prevalence of high-BMI children in the U.S.

Keywords

Bandwidth selection; Conditional percentile; Kernel estimator; Local linear regression; National Health and Nutrition Examination Survey; National Health Examination Survey

1. INTRODUCTION

Determination of the relative growth of children with respect to a standard reference population is important for assessing their health and development [1]. National growth curves for height, weight, and head circumference measurements by age have been developed by the U.S. Centers for Disease Control and Prevention (CDC) in 1976 and 2000 from data from the National Health Examination Surveys (NHES) and the National Health and Nutrition Examination Surveys (NHANES) that were collected by the U.S. National Center for Health Statistics (NCHS) [2,3]. These curves were developed to provide a reference for assessing the physical development of children residing in the U.S. and in other countries [4]. Smoothed curves by age are available for selected percentiles of each body measurement, e.g., 5th, 10th, 25th, 50th, 75th, 85th, 90th, and 95th, which are used to identify poorly developing children when they reach an extreme percentile level. The reported distribution of the first edition of the growth curves was about 12 million NCHS growth charts annually by pharmaceutical companies, which indicates their widespread use [5].

Because of increasing rates of overweight among children in the U.S. [6], there is great interest in obtaining national reference percentile curves of adiposity in children. An indirect but highly correlated measure of adiposity is body mass index (BMI) [7], which is defined as body weight (in kilograms) divided by the square of height (in meters). Because BMI is a simple function of body weight and height, it is readily obtained in population studies and surveys where it is computed from self-reported or measured body weight and height. Since the distribution of BMI in children varies greatly by age, CDC has pooled height and body weight data from a number of surveys conducted from the 1960's to the 1990's to produce percentile curves of BMI versus age [3]. Since overweight among children has been increasing over the past 10–15 years [8], the use of these older survey samples are based on children who were thinner and considered healthier. Thus, these curves are treated as “ideal” healthwise and are used as cutoffs to determine current and future prevalence estimates of overweight for various populations [9].

The construction of the most recent CDC percentile curves of BMI-for-age involved nonparametric local weighted regression smoothing and then further smoothing using polynomial regression [3]. Because of the polynomial regression the construction of these percentile curves are partially parametric.

In this paper we use another approach to construct BMI-for-age percentile curves. The percentiles plotted for the curves are the (weighted) percentile from the weighted empirical conditional cumulative distribution function of BMI given x using local-linear kernel weights [10]. In this case, a bandwidth on the age axis (i.e., x -axis) is required for specifying the local-linear kernel weights. The sample weights from the survey data are incorporated into these local-linear kernel weights, and a median correction is used to reduce the bias when estimating percentiles other than the median [11]. Roughness in the BMI-for-age percentile curves is reduced by using linear interpolation between the steps of the cumulative conditional (on age) distribution function before obtaining the percentiles [11]. This nonparametric approach will be called the ‘single-kernel’ method.

Another approach, which tends to be smoother than the single-kernel method, is a ‘double-kernel’ method proposed by Yu and Jones [12] for nonsurvey data. This approach smoothes the data along both the x and y -axes using kernel smoothing and, additionally, using local-linear weighting in the x -axis direction. The double-kernel method requires two bandwidths, one for kernel smoothing along the x -axis and the other for kernel smoothing along the y -axis. Yu and Jones [12] describe an automated method for determining the two bandwidths for a particular dataset for each percentile curve. Like the single-kernel method, the double-kernel method is nonparametric. There are other methods that are parametrically based for obtaining conditional percentile curves, which can be used for estimating BMI-for-age percentiles; see the Discussion section.

In Section 2 of this paper, we modify the double-kernel method and the bandwidth selection procedure of Yu and Jones [12] to incorporate (survey) sample weights. Simulations are provided in Section 3 that illustrate the reduction in bias when using a median correction in the estimation of conditional percentile curves and that examine the mean-squared-error properties of incorporating sample weights in the bandwidth selection procedure. In Section 4, we re-estimate the BMI-for-age percentile curves for the NCHS data for male and female children and adolescents that pooled data from five cross-sectional nationally representative health examination surveys. We compare these nonparametric percentile curves to the original partially parametric 2000 CDC curves and show how the nonparametric curves can change the estimation of the BMI-for-age relationship. In addition, using nonparametric percentile curves we re-estimate the prevalence of children in the U.S. who are above specified BMI-for-age percentiles and compare those estimates to published estimates based

on the 2000 CDC percentile curves. We end with a discussion in Section 5 that includes areas of future research.

2. METHODS

In this paper, we describe the two kernel estimators of conditional percentile curves for sample-weighted survey data that we will use to construct BMI-for-age percentile curves. Before describing these estimators, we first introduce kernel estimation of a conditional mean, which extends naturally to conditional percentile estimation.

In the nonsurvey setting (which means without survey sample weights), let $(x_1, y_1), \dots, (x_n, y_n)$ be a sample of n independent and identically distributed observations where y_i 's are scalar response variables and the x_i 's are scalar predictor variables (more generally the x_i 's can be of higher dimension, see Ruppert and Wand[13]). Let the conditional mean of y given x be denoted by $m(x)$. The local degree t least-squares kernel estimator of $m(x)$ is given by $\widehat{m}(x; h, t) = \widehat{\beta}_0$ where

$$(\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_t)' = \arg \min_{\beta} \sum_{i=1}^n \{y_i - \beta_0 - \beta_1(x - x_i) - \dots - \beta_t(x - x_i)^t\}^2 K((x - x_i)/h).$$

Here $h > 0$ is the bandwidth and the kernel function $K\{\cdot\}$ is a nonnegative symmetric kernel function that integrates to one. If $t = 1$, the standard weighted least-squares theory leads to the local-linear kernel estimator of $m(x)$, given by

$$\widehat{m}^L(x) = \sum_{i=1}^n w_i^L(x) y_i, \quad (1)$$

where $w_i^L(x) = w_i^k(x) \left\{ 1 + \frac{(x_i - x)(x - x_i)}{\sum_{j=1}^n w_j^k(x) (x_j - x)^2} \right\}$ are local-linear kernel weights, $\bar{x} = \sum_{i=1}^n w_i^k(x) x_i$ and $w_i^k(x) = K((x - x_i)/h) / \sum_{j=1}^n K((x - x_j)/h)$ are kernel weights. The dependency of the mean estimators (e.g., $\widehat{m}^L(x)$) and weights (e.g., $w_i^L(x)$ and $w_i^k(x)$) on the bandwidth h is suppressed for notational convenience.

In the survey setting (which allows for survey sample weights), let w_i be the sample weight corresponding to the observation (x_i, y_i) , $i = 1, \dots, n$. To account for the sample weights in the estimation, we let

$$w_i^{ks}(x) = w_i K\{(x - x_i)/h\} / \sum_{j=1}^n w_j K\{(x - x_j)/h\}, \text{ and}$$

$$w_i^{LS}(x) = w_i^{ks}(x) \left\{ 1 + \frac{(x_i - \bar{x})(\bar{x} - x)}{\sum_{j=1}^n w_j^{ks}(x) (x_j - \bar{x})^2} \right\},$$

where $\bar{x} = \sum_{i=1}^n w_i^{ks}(x) x_i$. The sample-weighted local-linear conditional mean estimator is given by

$$\widehat{m}^{LS}(x) = \sum_{i=1}^n w_i^{LS}(x) y_i. \quad (2)$$

The use of the sample weights in (2) implies that it is estimating what (1) would be estimating if all of the population values were used for the estimation. For further details about estimating the conditional mean with survey data, see Korn and Graubard [14].

2.1 Weighted Single-Kernel Conditional Percentile Estimator

The representation (2) of the conditional mean estimator as a weighted mean of y_i 's leads naturally to obtaining percentile estimators for the conditional distribution of Y given X . In the survey setting, to estimate the conditional percentiles for each $X = x$, we can use the percentile estimated from the weighted empirical cumulative distribution function of Y given $X=x$ using the sample-weighted local-linear kernel weight $w_i^{LS}(x)$ [11,14]. Let $F(r | X = x) = P(Y \leq r | X = x)$, and its weighted local-linear kernel estimator be

$$\widehat{F}(r|x) = \sum_{i=1}^n w_i^{LS}(x) I(y_i \leq r) / \sum_{i=1}^n w_i^{LS}(x).$$

For Y with a continuous distribution, the p^{th} conditional percentile given x is defined by $F^{-1}(p | x)$. One would like to estimate the percentile by $\widehat{F}^{-1}(p|x)$, but since $\widehat{F}^{-1}(\cdot|x)$ is a step function, it is not defined. By using linear interpolation, we can define $\widehat{F}^{-1}(\cdot|x)$ as follows: Let $y_{(1)} < y_{(2)} < \dots < y_{(n)}$ be the ordered Y data. Then the estimator of the p^{th} percentile is given by:

$$\widehat{q}^{SW}(p|x) \equiv \widehat{F}^{-1}(p|x) = y_{(j)} + \frac{p - \widehat{F}(y_{(j)}|x)}{\widehat{F}(y_{(j+1)}|x) - \widehat{F}(y_{(j)}|x)} (y_{(j+1)} - y_{(j)}),$$

with j chosen such that $\widehat{F}(y_{(j)}|x) \leq p < \widehat{F}(y_{(j+1)}|x)$. We will refer to $\widehat{q}^{SW}(p|x)$ as the (sample) weighted single-kernel conditional percentile estimator of the percentile p because it uses one kernel weight along the x -axis requiring a single bandwidth. In the nonsurvey setting, this single-kernel estimator is similar to the one proposed by Yu and Jones [12].

2.2 Weighted Double-Kernel Conditional Percentile Estimator

The double-kernel percentile estimator of Yu and Jones [12] is obtained by estimating the conditional distribution function of the Y given x using local-linear weighting and a kernel smoother along the x -axis with a specified bandwidth, and simultaneously using a kernel smoother along the y -axis with its own bandwidth. The estimated percentile is obtained from the estimated conditional distribution. Because this estimator of the distribution function is not in closed form, iteration is required to locate the percentile of interest; see below. In spite of the extra computation, the double-kernel estimator was preferred by Yu and Jones [12] to the single-kernel estimator because of its smoother appearance and improved mean-squared-error properties. In what follows, we modify the double-kernel percentile estimator to incorporate sample weights so that it is applicable to survey data.

In the nonsurvey setting, the double-kernel conditional percentile estimator of Yu and Jones [12] is the solution to

$$p = \sum_j w_j^L(x) \Omega\left(\frac{Q(x) - y_j}{h_2}\right) \quad (3)$$

for $Q(x)$, where $w_j^L(x)$ are local-linear kernel weights along the x -axis with bandwidth h_1 (see previous section) and Ω is a specified distribution function associated with the kernel weights along the y -axis with bandwidth h_2 .

Following Yu and Jones [12] we use a uniform $[-1, 1]$ distribution function Ω . Equation (3) for percentile p at x can then be written as

$$\begin{aligned} Q(x) = & \widehat{m}^L(x) + (2p - 1)h_2 \\ & + \sum_j \left(Q(x) - y_j - h_2 \right) w_j^L(x) I[y_j \leq (Q(x) - h_2)] \\ & + \sum_j \left(Q(x) - y_j + h_2 \right) w_j^L(x) I[y_j \geq (Q(x) + h_2)], \end{aligned} \quad (4)$$

Where $I[\cdot]$ is an indicator function (see the Appendix for derivation of (4) from (3)). By solving (4) iteratively for $Q(x)$, we obtain the double-kernel percentile estimator, denoted by $\widehat{q}^D(p|x)$.

In order to incorporate the sample weights, we replace the $w_j^L(x)$ in (3) by $w_j^{LS}(x)$. We can see how this enters into the double-kernel percentile estimation by first expressing equation (4) without the sample weights as

$$\begin{aligned} Q(x) = & \frac{\sum_{i=1}^n (x - x_i)^2 K\{(x - x_i)/h_1\} \times \sum_{i=1}^n s_i(x) K\{(x - x_i)/h_1\}}{\sum_{i=1}^n (x - x_i)^2 K\{(x - x_i)/h_1\} \times \sum_{i=1}^n K\{(x - x_i)/h_1\} - \left(\sum_{i=1}^n (x - x_i) K\{(x - x_i)/h_1\} \right)^2} \\ & - \frac{\sum_{i=1}^n (x - x_i) K\{(x - x_i)/h_1\} \times \sum_{i=1}^n s_i(x) (x - x_i) K\{(x - x_i)/h_1\}}{\sum_{i=1}^n (x - x_i)^2 K\{(x - x_i)/h_1\} \times \sum_{i=1}^n K\{(x - x_i)/h_1\} - \left(\sum_{i=1}^n (x - x_i) K\{(x - x_i)/h_1\} \right)^2}, \end{aligned} \quad (5)$$

where

$$\begin{aligned} s_i(x) = & (2p - 1)h_2 + y_i + [Q(x) - y_i - h_2] I[y_i \leq (Q(x) - h_2)] \\ & + [Q(x) - y_i + h_2] I[y_i \geq (Q(x) + h_2)] \end{aligned}$$

(Yu, K., personal communication, 2006). Inclusion of the sample weights (w_i) in (3) results in changing each sum in (5) to a (sample) weighted sum, that is,

$$Q(x) = \frac{\sum_{i=1}^n w_i(x-x_i)^2 K((x-x_i)/h_1) \times \sum_{i=1}^n w_i s_i(x) K((x-x_i)/h_1)}{\sum_{i=1}^n w_i(x-x_i)^2 K((x-x_i)/h_1) \times \sum_{i=1}^n w_i K((x-x_i)/h_1) - \left(\sum_{i=1}^n w_i(x-x_i) K((x-x_i)/h_1) \right)^2} - \frac{\sum_{i=1}^n w_i(x-x_i) K((x-x_i)/h_1) \times \sum_{i=1}^n w_i s_i(x) K((x-x_i)/h_1)}{\sum_{i=1}^n w_i(x-x_i)^2 K((x-x_i)/h_1) \times \sum_{i=1}^n w_i K((x-x_i)/h_1) - \left(\sum_{i=1}^n w_i(x-x_i) K((x-x_i)/h_1) \right)^2}. \quad (6)$$

By solving (6) iteratively for $Q(x)$, we obtain the weighted double-kernel percentile estimator $\hat{q}^{DW}(p|x)$.

2.3 Bandwidth Selection

The choice of the bandwidth is critical in determining how smooth the resulting conditional mean/percentile curve will be. We will first consider modifying a bandwidth procedure of Ruppert et al. [15] for estimating the conditional mean. We do this because this allows us to follow Yu and Jones [12] and use this bandwidth to choose a bandwidth for the conditional percentile.

2.3.1 Bandwidth for Conditional Mean—In the nonsurvey setting several automatic methods for selecting a bandwidth for a conditional mean smoother have been proposed by Ruppert et al. [15]. To derive an asymptotically optimal bandwidth they use the global loss criterion of the conditional mean integrated squared error (MISE) of $\hat{m}^L(x)$:

$$MISE\{\hat{m}^L(x)|X_1, \dots, X_n\} = E\left\{\int [\hat{m}^L(x) - m(x)]^2 dF(x)|X_1, \dots, X_n\right\}, \quad (7)$$

where $F(x)$ is the distribution function of X and the conditional expectation in (7) is with respect to distribution of $Y_1, \dots, Y_n$. By interchanging the conditional expectation and integration, we can re-express the MISE in (7) as a sum of the integrated variance and bias-squared of $\hat{m}^L(x)$ by

$$\begin{aligned} MISE\{\hat{m}^L(x)|X_1, \dots, X_n\} &= \int E\left\{\left[\hat{m}^L(x) - m(x)\right]^2 | X_1, \dots, X_n\right\} dF(x) \\ &= \int \text{var}\{\hat{m}^L(x)|X_1, \dots, X_n\} dF(x) + \int \left[E\{\hat{m}^L(x)|X_1, \dots, X_n\} - m(x)\right]^2 dF(x). \end{aligned}$$

For the local-linear kernel estimator $\hat{m}^L(x)$, it follows from theorem 4.1 of Ruppert and Wand [13], with appropriate regularity conditions holding, that

$$MISE\{\hat{m}^L(x)|X_1, \dots, X_n\} \approx h^{-1} R(K_1) \int [v(x)/nf(x)] dF(x) + h^4 \{A_2/2\}^2 \int m''(x)^2 dF(x), \quad (8)$$

where

$$K_1(u) = \frac{A_2 - uA_1}{A_2 - A_1^2} K(u),$$

$$A_j = \int u^j K(u) du, j=1, 2,$$

$v(x) = \text{var}(Y|X=x)$ is assumed finite,

$f(x)$ is the density of X ,

$$R(K_1) = \int K_1^2(u) du,$$

$m''(x)$ is the 2nd derivative of $m(x)$ with respect to x ,

and the integrals are taken over the real line. The two terms on the right-hand side of (8) approximate the integrated variance and integrated bias-squared, respectively. By maximizing the right-hand side of (8) with respect to h , one obtains the MISE approximately asymptotically optimal bandwidth

$$h_{MISE}(n, \theta) \approx \left[\frac{R(K_1) \int v(x) dx}{nA_2^2 \theta} \right]^{\frac{1}{5}}, \quad (9)$$

where

$$\theta \int m''(x)^2 dF(x).$$

In the survey setting, we modify (9) by considering the effect of the sample weights on both the variance and bias-squared terms of (8). In what follows we assume that the sample weights (w_i) may depend on the values of x but, conditional on x , the w_i are independent of Y . The variance term of (8) involves the integration of the ratio, $v(x)/nf(x)$ over the distribution of X times the term $h^{-1}R(K_1)$ that does not involve the data. In the nonsurvey setting, $nf(x)$ can be viewed as roughly the sample size of the observations used to estimate $\widehat{m}^L(x)$ at x , and $v(x)/nf(x)$ times $h^{-1}R(K_1)$ approximates the variance of $\widehat{m}^L(x)$ at x . In the survey setting, where we use the estimator $\widehat{m}^{LS}(x)$ to estimate the conditional mean, the sample weighting will tend to inflate the variance of $\widehat{m}^{LS}(x)$. To account for this inflation we replace n by an effective sample size, n^* , which will be less than n when there is inflation. Then $n^*f(x)$ becomes an effective sample size for the estimation at x with $v(x)/n^*f(x)$ multiplied by $h^{-1}R(K_1)$ approximating the variance for $\widehat{m}^{LS}(x)$ at x .

We obtain n^* by dividing n by a design-effect factor. This factor is obtained by first considering a design effect at an x , $deff[\widehat{m}^{LS}(x), \widehat{m}^L(x)]$, which is the variance of the conditional mean estimator with sample weighting divided by the variance without sampling weighting. We estimate $deff[\widehat{m}^{LS}(x), \widehat{m}^L(x)]$ by estimating

$$\frac{\text{Var} \left[\sum_{i=1}^n w_i^{\circ LS}(x) y_i | x_i \right]}{\text{Var} \left[\sum_{i=1}^n w_i^{\circ L}(x) y_i | x_i \right]}.$$

Here the superscript “ $\circ$ ” is used to denote that a method is used for choosing a bandwidth for $w_i^{\circ LS}(x)$ and $w_i^{\circ L}(x)$ that does not take account of the sample weighting (e.g., bandwidth selection procedure $\widehat{h}_{DPI}$; see Ruppert et al.[15]). If we assume that the variance for the y_i 's is constant, i.e., independent of x , and ignore any effects from possible stratification and cluster sampling of the sample design on the variance of the conditional mean estimators,

then we can estimate $\text{deff}[\widehat{m}^{LS}(x), \widehat{m}^L(x)]$ by

$\text{deff}[\widehat{m}^{LS}(x), \widehat{m}^L(x)] = \frac{\sum_{i=1}^n [w_i^{\circ LS}(x)]^2}{\sum_{i=1}^n [w_i^{\circ L}(x)]^2}$. (An alternative estimation of the design effect that takes into account the stratification and clustering of the sample design can be obtained by using a replication method (e.g., jackknife variance estimation) for estimating the variance of the conditional mean [14]. However, replication methods could entail considerable computing in this application.) An estimate of the effective sample size n^* is then obtained by first choosing a set G of equally spaced x values distributed over the range of x 's to evaluate the estimates of $\text{deff}[\widehat{m}^{LS}(x), \widehat{m}^L(x)]$. We use the median of the $\text{deff}[\widehat{m}^{LS}(x), \widehat{m}^L(x)]$ over G , $\text{med}_G\{\text{deff}[\widehat{m}^{LS}(x), \widehat{m}^L(x)]\}$, to represent the design effect for estimating $\widehat{m}^{LS}(x)$ over the range of x . The effective sample size is then estimated by

$$n^* = \frac{n}{\text{med}_G\{\text{deff}[\widehat{m}^{LS}(x), \widehat{m}^L(x)]\}}. \quad (10)$$

An implicit assumption in this calculation is that the design effect is approximately constant over the x 's so that the density $f(x)$ does enter into the computation of n^* .

In the survey setting there is also a choice as to what distribution to integrate $v(x)/nf(x)$ over for the variance term in (8). We choose to integrate over the unweighted $f(x)$ ($dF(x)$), i.e., the density for the sample not the population. (We do the same for integrating the bias-squared term that is discussed next.) By doing the integration in this way we minimize the MISE for our sample not for the population. In other words, our objective is to use a bandwidth to obtain a conditional mean curve that is optimally smooth for the sample at hand not for the population that we do not fully observe.

Next we considered the effect of the sample weights on estimating the bias-squared term in expression (8), in particular the term $\theta = \int m''(x)^2 dF(x)$. The sample weights can be incorporated in the estimation of $m''(x)$ and/or of $F(x)$, but we chose not to do this for the following reason. As we explained earlier, we chose to use the unweighted density of X so we will focus only on the effect of using the sample weights in estimating $m''(x)$. We explore this by incorporating the sample weights in the estimation of θ by using $\widehat{m}^{LS}(x)$ to estimate $m(x)$ for computing $m''(x)$, and compared this to the estimate of θ using $\widehat{m}^L(x)$ to estimate $m(x)$ for computing $m''(x)$. We used a Monte Carlo simulation (not shown) to make the comparison. In this simulation, we used as our bandwidth for computing $\widehat{m}^L(x)$ the bandwidth selection procedure $\widehat{h}_{DPI}$ recommended by Ruppert et al. [15] that is based on (9).

We modified $\widehat{h}_{DPI}$ for computing $\widehat{m}^{LS}(x)$ by substituting n^* (given by (10)) for n in the computation of $\widehat{h}_{DPI}$, denoted by $\widehat{h}_{DPI}^*$; see Ruppert et al. [15] for further details about computing $\widehat{h}_{DPI}$. We found that there was little improvement in the estimation of $MISE$ when using the weighted estimator of θ compared to the unweighted one. This appeared to be due to the variability of the estimator of $m''(x)$ employed by Ruppert et al. [15] in their $\widehat{h}_{DPI}$ bandwidth selection procedure. This variability swamped any effect of the sample weighting in the estimation of $m''(x)$. Therefore, in what follows we will use unweighted estimates of the $m''(x)$ in the estimation of θ .

2.3.2 Bandwidth for Conditional Percentiles—In the nonsurvey setting, Yu and Jones [12] proposed an automatic bandwidth selection strategy for estimating conditional percentiles using single and double-kernel methods. Their bandwidth selection procedure for the single-kernel conditional percentile estimator requires a bandwidth for the conditional mean be selected first. As suggested by Yu and Jones [12], we use $\widehat{h}_{DPI}$ as the bandwidth selection procedure for the conditional mean. Next the conditional mean bandwidth is modified according to the percentile being estimated. To obtain a bandwidth in the x direction to be used for the conditional p^{th} percentile single-kernel estimator, we follow Yu and Jones [12] and use

$$\widehat{h}_1(p) = \widehat{h}_{DPI} \left\{ p(1-p) / \phi(\Phi^{-1}(p))^2 \right\}^{\frac{1}{5}}, \quad (11)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the normal density and distribution function, respectively. For the double-kernel estimator of the conditional percentile, Yu and Jones [12] use the same $\widehat{h}_1(p)$ for smoothing in the x direction as was used for the single-kernel estimator and then they suggest a method that depends only on p and $\widehat{h}_1(p)$ for obtaining $\widehat{h}_2(p)$, the bandwidth estimator for smoothing in the y direction:

$$\widehat{h}_2(p) = \max \left(\frac{\widehat{h}_1^5(0.5)}{\widehat{h}_1^3(p)}, \frac{\widehat{h}_1(p)}{10} \right) \text{ if } \widehat{h}_1(0.5) < 1$$

and

$$\widehat{h}_2(p) = \frac{\widehat{h}_1^4(0.5)}{\widehat{h}_1^3(p)} \quad \text{otherwise.} \quad (12)$$

A conditional percentile methodology should be scale invariant for both the x and y data. The automatic bandwidth selection procedure for bandwidth $\widehat{h}_1(p)$ in the x -axis direction used for the single- and double-kernel procedures is scale invariant since $\widehat{h}_{DPI}$ is scale invariant. Unfortunately, the automatic bandwidth selection procedure for obtaining the bandwidth $\widehat{h}_2(p)$ in the y -axis direction, as suggested by Yu and Jones [12], is not scale invariant. An implicit assumption made by Yu and Jones [12] is that the x and y data have the about the same scale (i.e., the x and y data have approximately equal standard deviation). Limited simulations that we have conducted appear to support this assumption (not shown). To overcome this scale problem for obtaining the bandwidth $\widehat{h}_2(p)$, we rescale the y data by

the simple transformation $\frac{SD(x)}{SD(y)}^y$, where $SD(x)$ and $SD(y)$ denote the estimated standard deviation of the x and y data, respectively. The estimate for the p^{th} conditional percentile for the untransformed scale of the y data can be obtained by back-transforming the double-kernel estimate, i.e., $\frac{SD(y)}{SD(x)} \hat{q}^{DW}(p|x)$.

2.4 Median Correction

The single or double-kernel methods have a serious drawback for estimating percentiles other than the median. The larger the bandwidth associated with the x -axis the more the percentiles will be biased away from the median, even if the relationship of the percentiles and x was linear (but not horizontal). This is because the changing values of the conditional percentiles as a function of x causes the spread of y values weighted by the kernel to be larger when a larger bandwidth is considered and thereby increasing and decreasing the estimated percentiles that are greater and less than the median, respectively [14].

To avoid this bias when estimating conditional percentiles other than the median, we apply the similar approach of Korn and Graubard [11] for “centering” the y data about the median before estimating the conditional percentiles: We first estimate the conditional median using either the single or double-kernel estimator. For example, in case of the double-kernel estimator the conditional median estimator is $\hat{q}^{DW}(0.5|x)$. Let $z_i = y_i - \hat{q}^{DW}(0.5|x_i)$. To estimate a conditional percentile greater than the median, say the 90th percentile, we use the same conditional percentile method as used to estimate the median to estimate the conditional 90th percentile of the z 's given x as $\hat{q}_z^{DW}(0.9|x)$, then the desired conditional 90th percentile is estimated by $\hat{q}^{DW}(0.5|x) + \hat{q}_z^{DW}(0.9|x)$. This modification applies to conditional percentiles less than the median in the obvious fashion and similarly applies to the single-kernel method.

3. SIMULATION STUDIES

In the Methods section, we describe a median correction to reduce the bias of conditional percentile estimates and the proposed design-effect adjusted bandwidth in the survey setting. In this section, two simulation studies evaluate these techniques. In these simulations we use a normal kernel in the x -direction and a uniform kernel in the y -direction.

3.1 Median Correction

The datasets are generated using a simple quadratic model with

$$Y_i = 10 - \frac{2}{5}(X_i - 5)^2 + b_i, \quad (13)$$

where $X_i \sim UNIF(-.5, 5.5)$ and $b_i = (\frac{1}{5} + 1) \times \frac{\exp(e_i) - \exp(.02)}{\exp(.04)(\exp(.04) - 1)}$ with $e_i \sim N(0, .2)$. The function of b_i is specified such that the curves for the true percentile values of Y are similar to the observed growth curves of the 2000 CDC BMI-for-age, whose BMI distribution is skewed to the right for a given age and has increasing spread for older ages. We generated 1,000 datasets of sample sizes $n=500$. The mean of the 1,000 double-kernel percentile estimates $\hat{q}^D(p|x)$ with and without median correction were compared to the true percentile values for percentiles $p=.05$, $p=.5$ and $p=.95$. The results for X ranging from 0 to 5 are shown in Figure 1. The curve for the true percentile values of Y is close to the median

corrected curve while the uncorrected conditional percentile curve tends to be higher or lower than the true percentile curves for the 95th and 5th percentiles especially for X less than 3.5. As expected, without the median correction the bias appears to be greater when the slope of the true percentile curve is steeper.

3.2 Design-Effect Adjusted Bandwidth

A second simulation study was conducted to evaluate our proposed approach for selecting a bandwidth, $\hat{h}_{DPP}^*$ for a local-linear kernel conditional mean estimator for sample weighted survey data as described in section 2.3.1. This bandwidth is important because we follow the approach of Yu and Jones [12] to use it to subsequently select the bandwidths for the conditional percentile estimators. In this simulation, the quadratic model (13) was employed to generate 10,000 datasets of sample sizes $n=500$ of (X, Y) observations. In addition, we generated noninformative sample weights for the observations in each dataset by assigning sample weights with value 1 to a random nine-tenths of the observations of each generated dataset and assigning sample weights with value 50 to the remaining one-tenth of the sample. This results in a design effect of approximately 7 for estimating the conditional mean. We computed the conditional mean estimator for sample weighted data in two ways:

We computed $\hat{m}^{LS}(x)$ using $\hat{h}_{DPI}^*$ (which uses the sample weights) for determining the bandwidth and compared it to using $\hat{h}_{DPI}$ (which does not use the sample weights).

The mean square errors for the two conditional mean estimators are displayed for a selected set of x 's in Table 1. The values for x were chosen closer together at the ends of the interval to evaluate better the edge effects. The MSE of $\hat{m}^{LS}(x)$ at each x was lower for bandwidths determined by $\hat{h}_{DPI}^*$ than for $\hat{h}_{DPP}$, particularly near the edges. Using these same simulated datasets, we also compared the MSE of the single-kernel conditional percentile estimators $\hat{q}^{SW}(p|x)$ with bandwidths $\hat{h}_1(p)$ and $\hat{h}_1^*(p)$. The results are shown in Table 2 for the 25th, 50th, 75th, and 95th percentiles. For most values of x , the single-kernel conditional percentile estimators $\hat{q}^{SW}(p|x)$ with $\hat{h}_1^*(p)$ have lower MSE than that with $\hat{h}_1(p)$ across different percentile values.

4. BODY MASS INDEX-FOR-AGE PERCENTILE CURVES FOR U.S. CHILDREN

U.S. national growth charts consisting of a series of percentile curves of various measures of body size (e.g., height and body weight) in children and adolescents were first constructed in 1976 by NCHS [2]. These charts provided reference values that are an important clinical tool for health professionals for assessing the appropriate development of children in the U.S. and throughout the world. In 2000 CDC revised the growth charts for U.S. boys and girls and included additional charts for BMI for ages 2 to 20 years old [16]. These charts were created to replace the weight-for-age charts. The data used in the revision comes from the National Health Examination Survey (NHES) II (1963–65) and III (1966–70), and the National Health and Nutrition Examination Survey (NHANES) I (1971–74), II (1976–80), and III (1988–94). The sample designs of these surveys are stratified, multistage probability samples of the civilian, noninstitutionalized population in the 48 contiguous states (NHES II, NHES III, NHANES I) or all 50 states (NHANES II, NHANES III). Each of the surveys has sample weights that are a combination of the inverse of the rates of sample selection, and nonresponse and poststratification adjustments.

The construction of the most recent CDC percentile curves of BMI-for-age involved a three-step procedure [3]. In the first step, BMI values were grouped by age into 6-month intervals

for ages 2 to 20 years and sample-weighted empirical percentiles were computed for each age group for the percentiles of interest (3rd, 5th, 10th, 25th, 50th, 75th, 85th, 90th, and 97th). In the second step, each of these empirical percentiles was then smoothed across age using local weighted regression [17] with a tricubic kernel weight function. The bandwidths for the local weighted regression varied by age and sex; for details see Kuczmarski et al.[3]. Differential population sizes among the age groups were not accounted for in this smoothing. In the third step “The smoothed percentile curves obtained through LWR [local weighted regression] were then fit by a 4-degree polynomial to achieve parametric percentiles.” [3]

A public use data file called NHANES_GROWTHDATA was provided to us by NCHS. This data file corresponds to the data used to create the 2000 CDC Growth Charts. The file includes the BMI values along with the sample weights of 18,592 boys and 18,779 girls with ages 18–305 months. The curves of the conditional percentiles corresponding to the single and double-kernel methods with the median correction, and bandwidths $h_1^*(p)$ and $h_2^*(p)$ (that incorporate the sample weights in the bandwidth selection along with a scale adjustment for $h_2^*(p)$) are presented in Figures 2 and 3 along with the 2000 CDC BMI-for-age curves. (The kernel percentile curves are plotted beyond age 20 to age 23.) The design effect due to the sample weights was same for boys and girls,

$med_G \left\{ deff \left[\widehat{m}^{LS}(x), \widehat{m}^L(x) \right] \right\} = 1.79$, and resulted in effective sample sizes of $n^* = 10,387$ ($=18,592/1.79$) boys and 10,491 ($=18,779/1.79$) girls and in the bandwidth for the conditional mean curves of 8.48 and 9.27 months for boys and girls, respectively.

In general, the shape of the single-kernel curves is similar to the double-kernel ones, but the expected greater roughness of the single-kernel curves is evident (Figure 2). The CDC curves and the single and double-kernel curves for the conditional percentiles generally track together (Figures 2 and 3). However, there appears to be important systematic differences between the kernel and CDC curves. For example, the nadirs for CDC curves tend to be lower than those for the kernel curves where BMI decreases until approximately 4.5 to 6.5 years depending upon the percentile. The nadir discrepancy between the CDC curves and kernel curves appear to be larger for girls than for boys. Among the boys for ages beyond the nadir, the percentiles increase approximately linearly for the CDC growth curves whereas the kernel curves show nonlinear trajectories particularly in the higher and lower percentiles. Among the girls for ages beginning at about 11 to 13 years the kernel curves tend to be higher than the CDC curves across the different percentiles. Another major difference between the kernel and CDC curves appears near the oldest ages between 19 and 20 years that suggests the CDC curves may be oversmoothed. Among the boys, the 2000 CDC growth curves show an approximately linear increase in BMI in the older ages whereas the kernel curves show a plateauing of BMI starting at about 18 to 19 years. For the girls, the 2000 CDC growth curves for percentiles above the 50th percentile show an approximately linear increase in BMI in the older ages whereas the kernel curves show a plateauing of BMI starting at about 16–17 years. In order to further evaluate the differences in the patterns of growth in BMI between the curves, we plotted BMI-for-age using the kernel methods for ages 20–23 years. The increase of BMI for older ages, as expected, is not as fast as the earlier ages, especially for the extreme percentiles. Because of the upward trajectory of the CDC BMI-for-age curves up to age 20, these curves give the impression that the children at older ages will increase their BMI as they age beyond 20 years and do not show the leveling off of the curves at ages 18–20.

The standard errors in Figure 4 for the double-kernel percentile curve were estimated using 200 random half-sample replications where one PSU was randomly selected from the two (sample) PSUs in each (sample) stratum from each survey to form a half-sample replicate

[14]. The five national surveys used to develop the growth curves have two sample PSUs for each sample stratum except for the first ten strata in NHANES I where the entire stratum was a PSU (called a certainty stratum). In order to form two (pseudo-)PSUs in these strata, we followed a standard survey research approach by randomly dividing the next smaller sample units (segments) in these stratum into two groups per stratum [14]. The strata were combined across the surveys by concatenating them to form a total of 156 strata, each now with two PSUs. Each half-sample were used to estimate a replicate double-kernel smoothed percentile curve. Following the recommendation of Korn and Graubard [14], the bandwidth used for the replicate double-kernel estimates is the same as the bandwidth used for the kernel estimates for the original data. The variances were estimated from these 200 half-sample replicates [14]. Figure 4 presents the standard error of 85th and 95th double-kernel percentile estimates for age among U.S. children. In general, standard error increases with age, and is less than 0.4 for the 85th percentile and less than 1 for the 95th percentile.

To test statistically the difference between the CDC and the double-kernel growth curves, we need to estimate the variances for differences between corresponding points for the CDC and double-kernel growth curves. As described below, this requires re-estimating both sets of curves from half-sample replicates of the data. We were unable to do this because some of the exact details of the construction of the CDC growth curves were not described and the CDC computer programs for estimating the curves are not available. Therefore, in Figure 4, we provide estimates for the standard errors of the points on 85th and 95th percentile for only the double-kernel curves. Because both the CDC and the double-kernel curves have similar shapes and are estimated using the same data, the covariance between the two sets of estimates should be large resulting in quite small standard errors for the differences between the curves.

Recently the 2000 CDC percentile curves were used as reference curves with the recent NHANES 2003–2006 sample for obtaining the most current U.S. prevalence estimates of children and adolescents with a high BMI, i.e., a BMI at or above the 85th or 95th percentiles for age [9]. In this application we compare the prevalence estimates using our double-kernel percentile curves to those using the 2000 CDC percentile curves (Table 3). Prevalence estimates are presented for the same sex, racial/ethnic, and age groups as Ogden et al. [9], except that the last age group (12–19 years) was further partitioned into 12–15 years and 16–19 years so that there were approximately same sample sizes across age groups. Design-based standard errors (computed using Taylor linearization variance estimation, SUDAAN [18]) of the prevalence estimates and two-sided p-values for differences in the prevalence based on the two percentile curves are presented by taking into account the variability from the NHANES 2003–2006 data but assuming no variability in either the 2000 CDC or the double-kernel percentile curves. This was the same assumption used by Ogden et al. [9] in their estimates of standard errors and is the assumption used whenever the 2000 CDC percentile curves are used as reference curves.

For 2–5 years olds, the CDC prevalence estimates for children at or above the 85th and 95th percentiles tend to be larger. For 6–11 years, the double-kernel estimates are close to the CDC estimates and most differences between the two estimates are within 1%. For 12–15 and 16–19 years, the boys and girls show different patterns. For boys the two sets of estimates are similar, but for the girls the CDC estimates for 12–15 years tend to estimate higher prevalence compared to the double kernel estimates, particularly for the 85th percentile; while the CDC estimates for girls in 16–19 years are consistently smaller than the double-kernel estimates. These findings are consistent with the appearance of the curves in Figure 3.

5. DISCUSSION

We modified the double-kernel method of Yu and Jones [12] for estimating conditional percentile curves in order to re-estimate the U.S. BMI-for-age percentiles curves for children and young adults. Because the data used to compute the BMI percentile curves came from national surveys with sample weighting, the estimation of the percentile curves was modified to incorporate these sample weights so that the percentiles curves would closely approximate percentile curves of the populations from which the samples were selected. The advantage of our percentile curves is that they are estimated using a nonparametric method that does not make modeling assumptions, as opposed to the partially parametric method used in the generation of the 2000 CDC percentile curves. An attribute of the double kernel method is that its estimated percentiles curves will not cross or touch each other (Yu and Jones [12]). Another advantage is that the procedure is automatic so that standard errors can be estimated using re-sampling methods such as bootstrap.

The emphasis of this paper is on nonparametric estimation. However, there are parametric-based methods with the LMS method of Cole and Green [19] being one of the more popular ones. For growth curves analyses, the LMS method relies on the age-specific measurements being transformed to normality using Box-Cox transformation, an assumption that may be questionable. However, the LMS method can provide very smooth age-specific percentile curves using a penalized likelihood estimation method along with age-specific curves for the Box-Cox power parameter, mean, and the coefficient of variation used in the data transformation that can be of substantive interest. Reference percentile growths curves are expected to be visually smooth [19]. The double kernel method produces smooth curves but can be made even smoother by either overriding the automatic bandwidth to widen it or by fitting splines through the points on the percentile curves. Gannoun et al. [20] and Wei et al. [21] provide comprehensive reviews of conditional percentile estimation, in general, and for use in estimating growth curves. A recent parametric method of Jones and Yu [22] uses a general class of distributions with exponential tails to produce quite percentile curves. Another useful alternative approach to kernel methods for developing percentile curves uses cubic B-spline basis functions [21]. This approach is quite flexible for capturing particular growth features such as the pubertal growth spurt, but it requires subjective placement of knots in the appropriate parts of the curves. Koenker [23] provides a comprehensive review of the development and application of parametric and nonparametric quantile regression methods.

Different characteristics were observed between our curves and those generated by CDC. The nadirs of the CDC curves tend to be lower than they are for the kernel curves, particularly for the higher percentiles. With increasing age beyond the age of the nadirs, the kernel curves show more nonlinearity than the CDC curves. At the upper end of the age range, the CDC curves for the boys and the CDC curves at or above the 50th percentile for the girls increase at a near constant rate whereas the kernel curves show a leveling off. This leveling off also appears to be evident from the CDC's empirical BMI percentiles [3]. We applied our 85th and 95th percentile curves as reference values for estimating the 2003–2006 prevalence of children who have a high BMI-for-age by gender, age and race/ethnicity groups. Many of our prevalence estimates differed significantly from those based on the using the 2000 CDC 85th and 95th percentile curves as reference values, indicating how differences in the shapes of these curves can affect their use in overweight research.

We made several simplifying assumptions about the survey sample design in our modification of the bandwidth selection procedure, $\hat{h}_{DPR}^*$. First, we assumed that the sample weights were noninformative for Y given x . Also, the estimation of the design effect,

deff $\left[\widehat{m}^{LS}(x), \widehat{m}^L(x)\right]$ by $\sum_{i=1}^n \left[w_i^{LS}(x)\right]^2 / \sum_{i=1}^n \left[w_i^L(x)\right]^2$ further assumes that the conditional variances of the y_i 's are constant and that the effect of any stratification or clustering in the sample design can be ignored. In practice for biological processes such as growth, stratifying by gender or other variables and conditioning on an important predictor such as age will considerably diminish the informativeness of the weights and other aspects of the sample design. Under these assumptions, simulations showed that our modified bandwidth procedure usually attained smaller MSE than choosing a bandwidth without regard to sample weighting. As with Yu and Jones [12], our modified bandwidth selection method provides a fixed bandwidth for smoothing along the x -axis that is applied to the entire range of x values for the single and double-kernel methods (i.e., bandwidths do not vary by the x). It would be useful to consider bandwidth selection procedures that allow the bandwidths to change with x in order to adjust them for differential precision of the estimated conditional mean/percentile with respect to x . For example, one would expect the precision to vary with x when the numbers of observations changes with x . This type of generalization in the bandwidth selection procedure is an area of further research.

In conclusion, we have re-estimated conditional percentile curves for BMI for U.S. boys and girls who are 2 to 20 years of age using data from combining five national health examination surveys as was used by the CDC to produce their percentile curves. These re-estimated curves show differences from the partially parametric curves published by the CDC that significantly alters estimates of the prevalence of children with high BMI-for-age in the U.S.

APPENDIX

Derivation of equation (4) from equation (3)

Using a uniform $[-1, 1]$ kernel density, the expression (4):

$$p = \sum_j w_j^L \Omega\left(\frac{Q(x) - y_j}{h_2}\right) \text{ becomes}$$

$$2p = \sum_j w_j^L \left\{ 1 + \left(\frac{Q(x) - y_j}{h_2} \right) \right\} I\{Q(x) - h_2 < y_j < Q(x) + h_2\} + \sum_j w_j^L \times 2 \times I\{y_j \geq Q(x) + h_2\}$$

Because $I\{Q(x) - h_2 < y_j \geq Q(x) + h_2\} = 1 - I\{y_j \geq Q(x) + h_2\} - I\{y_j \leq Q(x) - h_2\}$,

$$2p = 1 + \sum_j w_j^L \left(\frac{Q(x) - y_j}{h_2} \right) - \sum_j w_j^L \left\{ 1 + \left(\frac{Q(x) - y_j}{h_2} \right) \right\} \{I\{y_j \geq Q(x) + h_2\} + I\{y_j \leq Q(x) - h_2\}\} + \sum_j w_j^L \times 2 \times I\{y_j \geq Q(x) + h_2\},$$

$$(2p-1)h_2 = Q(x) - \sum_j w_j^L y_j - \sum_j w_j^L \{h_2 + Q(x) - y_j\} \{I\{y_j \geq Q(x) + h_2\} + I\{y_j \leq Q(x) - h_2\}\} + h_2 \sum_j w_j^L \times 2 \times I\{y_j \geq Q(x) + h_2\},$$

and

$$(2p-1)h_2 = Q(x) - \sum_j w_j^L y_j - \sum_j w_j^L \{-h_2 + Q(x) - y_j\} I\{y_j \geq Q(x) + h_2\} - \sum_j w_j^L \{-h_2 + Q(x) - y_j\} I\{y_j \leq Q(x) + h_2\}.$$

Therefore,

$$Q(x) = \widehat{m}^L(x) + (2p-1)h_2 + \sum_j (Q(x) - y_j - h_2) w_j^L I(y_j \leq (Q(x) - h_2)) + \sum_j (Q(x) - y_j + h_2) w_j^L I(y_j \geq (Q(x) + h_2)),$$

where $\widehat{m}^L(x) = \sum_j w_j^L y_j$.

REFERENCES

1. Garza C, de Onis M, WHO Multicentre Growth Reference Study Group. Rationale for Developing a New International Growth Reference. Food and Nutrition Bulletin. 2004; 25(Suppl 1):S5–14. [PubMed: 15069915]
2. Hamill PV, Drizd TA, Johnson CL, Reed RB, Roche AF. NCHS growth curves for children birth—18 years, United States. Vital Health Statistics. 1977; 11(165)
3. Kuczmarski RJ, Ogden CL, Guo SS, Grummer-Strawn LM, Flegal KM, Mei Z, Wei R, Curtin LR, Roche AF, Johnson CL. 2000 CDC Growth Charts for the United States: Methods and Development. National Center for Health Statistics. Vital Health Statistics. 2002; 11(246)
4. De Onis M, Garza C, Onyango AW, Borghi E. Comparison of the WHO Child Growth Standards and the CDC 2000 Growth Charts. The Journal of Nutrition. 2007; 137(1):144–148. [PubMed: 17182816]
5. Roche, AF. Executive Summary of the Growth Chart Workshop. National Center for Health Statistics; Hyattsville, Maryland: 1994. December 1992
6. Troiano RP, Flegal KM. Overweight Children and Adolescents: Description, Epidemiology, and Demographics. Pediatrics. 1998; 101:497–504. [PubMed: 12224656]
7. Freedman DS, Wang J, Ogden CL, Thorton JC, Mei Z, Pierson RN, Dietz WH, Horlick M. The Prediction of Body Fatness by BMI and Skinfold Thickness Among Children and Adolescents. Annals of Human Biology. 2007; 34:183–194. [PubMed: 17558589]
8. National Center for Health Statistics. With chartbook on trends in the health of Americans. Hyattsville, MD: 2007.
9. Ogden CL, Carroll MD, Flegal KM. High body mass index for age among US children and adolescents, 2003–2006. Journal of the American Medical Association. 2008; 299(20):2401–2405. [PubMed: 18505949]
10. Owen, AB. Technical Report No. 265. Department of Statistics, Stanford University; Stanford, CA: 1987. Nonparametric Conditional Estimation.
11. Korn EL, Graubard BI. Scatterplots with Survey Data. American Statistician. 1998; 52:58–69.

12. Yu K, Jones MC. Local Linear Quantiles Regression. *Journal of the American Statistical Association*. 1998; 93:228–237.
13. Ruppert D, Wand MP. Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics*. 1994; 22:1346–1370.
14. Korn, EL.; Graubard, BI. *Analysis of Health Surveys*. John Wiley and Sons; New York: 1999.
15. Ruppert D, Sheather SJ, Wand MP. An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*. 1995; 90:1257–1270.
16. Kuczmarski, RJ.; Ogden, CL.; Guo, SS.; Grummer-Strawn, LM.; Flegal, KM.; Guo, SS.; Wei, R.; Mei, Z.; Curtin, LR.; Roche, AF.; Johnson, CL. *Advance data from vital and health statistics; no 314*. National Center for Health Statistics; Hyattsville, Maryland: 2000. CDC Growth Charts: United States. Available from <http://www.cdc.gov/growthcharts>
17. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*. 1979; 74:829–836.
18. Research Triangle Institute. SUDAAN Language Manual, Release 9.0I. 2004. Research
19. Cole TJ, Green PJ. Smoothing reference centile curves: The LMS methods and penalized likelihood. *Statistics in Medicine*. 1992; 11:1305–1319. [PubMed: 1518992]
20. Gannoun A, Girard S, Guinot C, Saracco J. Reference Curves Based on Non-parametric Quantile Regression. *Statistics in Medicine*. 2002; 21:3119–3135. [PubMed: 12369086]
21. Wei Y, Pere A, Koenker R, He X. Quantile regression methods for reference growth charts. *Statistics in Medicine*. 2006; 25:1369–1382. [PubMed: 16143984]
22. Jones MC, Yu K. Improved double kernel local linear quantile regression. *Statistical Modelling*. 2007; 7:377–389.
23. Koenker, R. *Quantile Regression*. Cambridge University Press; Cambridge: 2005.

Figure 1.

Mean of the double-kernel percentile estimates $\widehat{q}^D(p|x)$ with and without median correction, comparing with the true percentile curve

Figure 2.
Comparison between the single-kernel smoothing curves and the 2000 CDC BMI-for-age curves for boys and girls

Figure 3.
Comparison between the double-kernel smoothing curves and the 2000 CDC BMI-for-age curves for boys and girls

Figure 4. Standard Error of the 85th and 95th double-kernel conditional percentile curve estimates by age among US boys and girls that are computed using half sample replication.

Table 1

Mean square error (MSE) of the conditional mean estimators $\widehat{m}^{LS}(x)$ with $\widehat{h}_{DPI}$ and $\widehat{h}_{DPI}^*$

x	MSE using	
	$\widehat{h}_{DPI}$	$\widehat{h}_{DPI}^*$
0	0.37	0.28
.2	0.16	0.15
.4	0.11	0.10
.6	0.11	0.08
.8	0.10	0.07
1	0.09	0.07
1.5	0.09	0.07
2	0.10	0.08
2.5	0.13	0.09
3	0.17	0.12
3.5	0.20	0.14
4	0.26	0.19
4.2	0.28	0.22
4.4	0.36	0.29
4.6	0.49	0.44
4.8	0.88	0.76
5	2.28	1.52

Table 2

Mean square error (MSE) of the single-kernel percentile estimators $\hat{q}_1^w(p|x)$ with $\hat{h}_1(p)$ and $\hat{h}_1^*(p)$ when percentile values $p=(.25, .5, .75, .95)$

x	p=.25 MSE using		p=.5 MSE using		p=.75 MSE using		p=.95 MSE using	
	$\hat{h}_1(p)$	$\hat{h}_1^*(p)$	$\hat{h}_1(p)$	$\hat{h}_1^*(p)$	$\hat{h}_1(p)$	$\hat{h}_1^*(p)$	$\hat{h}_1(p)$	$\hat{h}_1^*(p)$
0	0.51	0.49	0.47	0.57	0.76	0.73	1.34	1.05
.2	0.45	0.34	0.33	0.28	0.77	0.64	1.31	1.01
.4	0.29	0.28	0.20	0.20	0.51	0.50	0.95	0.86
.6	0.31	0.27	0.25	0.22	0.49	0.47	0.89	0.80
.8	0.26	0.24	0.21	0.20	0.43	0.40	0.83	0.74
1	0.25	0.21	0.20	0.18	0.38	0.34	0.83	0.66
1.5	0.20	0.16	0.17	0.13	0.30	0.24	0.71	0.48
2	0.21	0.16	0.17	0.14	0.33	0.26	0.77	0.54
2.5	0.24	0.18	0.20	0.16	0.36	0.27	0.93	0.61
3	0.30	0.21	0.26	0.18	0.50	0.34	1.16	0.74
3.5	0.36	0.26	0.30	0.21	0.59	0.39	1.28	0.88
4	0.41	0.31	0.35	0.27	0.71	0.54	1.67	1.35
4.2	0.49	0.37	0.40	0.32	0.86	0.65	1.87	1.53
4.4	0.66	0.50	0.49	0.41	1.13	0.84	2.13	1.75
4.6	0.95	0.78	0.66	0.60	1.65	1.27	2.82	2.33
4.8	1.37	1.41	1.21	1.08	2.59	2.11	4.38	3.47
5	2.36	2.41	3.37	2.31	5.32	4.25	9.22	6.08

Table 3

¹Comparison of Percentage of Children with high BMI by Age, Gender and Race based on Double-Kernel versus 2000 CDC Percentile Curves among US Children, 2003–2006 (SE's are in parentheses)

	All Races				Non-Hispanic White			Non-Hispanic Black			Mexican-American		
	DK ²	CDC ³	DK-CDC ⁴		DK	CDC	DK-CDC	DK	CDC	DK-CDC	DK	CDC	DK-CDC
BMI ≥ 95th percentile of the Growth Charts													
Boys													
2~5	12.0 (1.2)	12.8 (1.2)	-0.8 (0.3)*		10.5 (2.2)	11.1 (2.2)	-0.6 (0.4)	11.6 (2.4)	13.3 (2.5)	-1.6 (0.8)*	18.4 (2.8)	18.8 (2.8)	-0.4 (0.6)
6~11	18.2 (1.6)	18.1 (1.7)	0.2 (0.4)		16.3 (2.6)	15.5 (2.8)	0.7 (0.7)	17.5 (2.5)	18.6 (2.6)	-1.1 (0.6)	27.0 (2.1)	27.5 (2.1)	-0.5 (0.6)
12~15	17.4 (2.0)	17.0 (2.0)	0.5 (0.3)		16.2 (2.5)	15.6 (2.5)	0.6 (0.4)	19.2 (2.0)	19.2 (2.0)	0.0 (0.0)	19.6 (2.9)	20.0 (2.7)	-0.4 (0.5)
16~19	19.0 (2.2)	19.6 (2.3)	-0.6 (0.3)*		18.5 (2.8)	19.1 (2.9)	-0.6 (0.4)	17.0 (1.5)	17.7 (1.6)	-0.7 (0.4)	23.7 (2.3)	24.7 (2.5)	-1.0 (0.6)
Girls													
2~5	9.6 (1.1)	12.1 (1.4)	-2.5 (0.7)**		8.0 (1.7)	10.2 (1.9)	-2.2 (1.0)*	13.3 (1.6)	16.6 (2.3)	-3.3 (1.4)*	11.4 (2.1)	14.5 (2.7)	-3.1 (1.0)**
6~11	15.6 (1.4)	15.8 (1.4)	-0.2 (0.1)		14.4 (2.1)	14.4 (2.1)	0.0 (0.0)	24.3 (1.8)	24.0 (2.0)	0.2 (0.6)	17.9 (2.6)	19.7 (2.6)	-1.8 (0.8)*
12~15	16.9 (2.1)	17.9 (2.2)	-1.0 (0.3)**		15.0 (3.2)	15.6 (3.3)	-0.6 (0.4)	23.4 (1.9)	24.9 (1.9)	-1.4 (0.6)*	20.1 (1.8)	21.4 (1.8)	-1.3 (0.7)
16~19	16.3 (2.0)	15.8 (1.9)	0.5 (0.3)*		13.6 (2.1)	13.4 (2.0)	0.2 (0.2)	32.9 (4.5)	30.8 (3.6)	2.2 (1.2)*	18.9 (2.1)	17.9 (2.2)	1.0 (0.5)
BMI ≥ 85th percentile of the Growth Charts													
Boys													
2~5	23.9 (1.9)	25.5 (2.0)	-1.6 (0.7)*		23.9 (2.9)	25.4 (2.9)	-1.6 (1.0)	22.1 (3.2)	23.2 (3.3)	-1.1 (0.5)*	28.9 (3.1)	32.4 (3.2)	-3.5 (1.3)**
6~11	34.6 (2.3)	33.9 (2.3)	0.7 (0.4)		32.6 (3.7)	31.7 (3.6)	1.0 (0.6)	34.1 (3.2)	33.8 (3.6)	0.4 (0.7)	47.6 (2.4)	47.1 (2.5)	0.5 (0.3)
12~15	35.0 (2.7)	34.9 (2.6)	0.1 (0.4)		34.6 (4.0)	34.6 (3.9)	0.0 (0.6)	33.9 (2.2)	33.5 (2.2)	0.4 (0.0)**	39.0 (3.0)	39.0 (3.0)	0.0 (0.0)
16~19	35.1 (2.2)	35.0 (2.3)	0.1 (0.1)		34.4 (2.9)	34.4 (2.9)	0.0 (0.0)	31.5 (2.3)	30.6 (2.5)	0.9 (0.4)*	41.9 (3.2)	42.3 (3.2)	-0.4 (0.4)
Girls													
2~5	21.6 (1.7)	23.3 (1.8)	-1.6 (0.5)**		19.3 (2.5)	20.9 (2.6)	-1.6 (1.0)	26.0 (2.0)	26.4 (2.1)	-0.4 (0.4)	24.1 (2.7)	27.3 (2.8)	-3.3 (0.9)**
6~11	33.3 (2.2)	32.6 (2.4)	0.6 (0.4)		32.4 (3.2)	31.5 (3.6)	0.9 (0.8)	40.0 (2.3)	40.1 (2.3)	0.0 (0.4)	38.9 (3.2)	38.1 (3.3)	0.8 (0.7)
12~15	32.2 (2.3)	35.4 (2.4)	-3.2 (0.8)**		31.9 (3.4)	35.5 (3.5)	-3.7 (1.2)**	37.4 (2.4)	40.2 (2.7)	-2.8 (1.1)*	34.5 (2.8)	37.6 (2.9)	-3.1 (1.0)**
16~19	33.7 (1.9)	31.2 (2.0)	2.5 (0.6)**		30.4 (2.4)	27.8 (2.5)	2.6 (0.9)**	51.6 (3.5)	49.4 (3.5)	2.2 (0.8)**	38.6 (3.0)	36.5 (3.0)	2.1 (0.7)**

¹ Discrepancies of the differences between the columns of DK and CDC from the third column (DK-CDC) are due to rounding error.

² Double-Kernel

³ Centers for Disease Control and Prevention

⁴ Difference in prevalences between those based on the double-kernel and those based on the 2000 CDC percentile curves.

* indicate that the double-kernel estimate is statistically different from the CDC estimate with $p\text{-value} < .05$ and $< .01$, respectively.

** indicate that the double-kernel estimate is statistically different from the CDC estimate with $p\text{-value} < .05$ and $< .01$, respectively.