Growth Charts of Body Mass Index (BMI) with Quantile Regression

Colin Chen
SAS Institute Inc. Cary, NC, U.S.A. *

Abstract

Growth charts of body mass index (BMI) are constructed from the recent four-year national cross-sectional survey data (1999−2002) using parametric quantile regression methods, which are implemented with a newly developed SAS procedure (http://www.sas.com/statistics) and SAS macros.

KEY WORDS: Body mass index, growth charts, quantile regression, smoothing algorithm, simplex, interior point.

1   Introduction

Overweight has become a common problem in public health, especially for children. Obesity has been related to numerous health risks, both physical and psychological. Body mass index, defined as the ratio of weight (kg) to squared height (m^2), has been popularly used as a measure of overweight and obesity.

The percentiles of BMI for a specified age is of particular interest in light of public health concerns. Not only are the upper percentiles closely watched for overweight and obesity, the lower percentiles are also observed for underweight. The empirical percentiles with grouped age provide a discrete approximation for the population percentiles. However, continuous percentile curves are both more accurate and attractive.

There have been several methods used to construct such age-dependent growth charts. Early methods fit smoothing curves on sample quantiles of segmented age groups. However, such methods are not robust to outliers. Large sample size is needed in order to estimate the percentiles in each age group with appropriate precision. The segmentation may lose information from nearby groups. To avoid segmentation, Cole and Green (1992) developed a Box-Cox transformation-based semiparametric approach from the LMS (Lamda-Mu-Sigma) method introduced by Cole (1988). The semiparametric LMS method solves penalized likelihood equations. Because of the lack of finite expectation for some of the derivatives of the penalized log-likelihood, solutions of these equations could be sensitive to a start point. There is also the question whether the Box-Cox transformation is good enough for the specified distributional assumption, e.g., normality.

Quantile regression, which was introduced by Koenker and Bassett (1978), is an alternative way to create growth charts. It does not put any distributional assumption beforehand. It is also relatively easy to accommodate other covariates besides age. Computationally, it is fast and stable.

For a random variable Y with probability distribution function

$$F(y) = \text{ Prob } (Y \leq y), \tag{1}$$

the τth quantile of Y^1 is defined as the inverse function

$$Q(\tau) = \text{ inf } \{y : F(y) \geq \tau\}, \tag{2}$$

where $0 < \tau < 1$. In particular, the median is $Q(1/2)$.

---

*Phone: 919-531-6388. E-mail: Lin.Chen@sas.com

[1]Recall that a student's score on a test is at the τth quantile if his (or her) grade is better than $100\tau\%$ of the students who took the test. The score is also said to be at the 100τth percentile.

The τth sample quantile $\hat{\xi}(\tau)$, which is an analogue of $Q(\tau)$, may be formulated as the solution of the optimization problem

$$\min_{\xi \in \boldsymbol{R}} \sum_{i=1}^{n} \rho_\tau (y_i - \xi), \tag{3}$$

where $\rho_\tau(z) = z(\tau - I(z < 0))$, $0 < \tau < 1$, is usually called the check function.

When covariates X (e.g., age) are considered, the linear conditional quantile function, $Q(\tau|X = x) = x'\beta(\tau)$, can be estimated by solving

$$\hat{\beta}(\tau) = \mathrm{argmin}_{\beta \in \boldsymbol{R}^p} \sum_{i=1}^{n} \rho_\tau (y_i - x_i'\beta) \tag{4}$$

for any quantile $\tau \in (0, 1)$. The quantity $\hat{\beta}(\tau)$ is called the *regression quantile*. The case $\tau = 1/2$, which minimizes the sum of absolute residuals, is usually known as L_1 (median) regression.

As the unconditional sample quantile $\hat{\xi}(\tau)$, for a given τ, τ percent of the observed values of the continuous response variable Y (i.e., BMI) is expected to fall below the conditional quantile hyperplane $x'\hat{\beta}(\tau)$. Given $X = x$, as an unbiased estimator of the 100τth percentile of the conditional distribution of Y, $x'\hat{\beta}(\tau)$ should be close to the 100τth sample percentile at $X = x$. This indicates that the percentile curves constructed with the quantile regression method fit the local empirical percentiles well.

BMI has been known to be skewed on the right. Departure from the normality assumption after the Box-Cox transformation in the LMS method was reported in Flegal (1999). Departure from normality, especially in the tails, can affect estimates of underweight or overweight. This further promotes the quantile regression method for constructing growth charts of BMI.

Historically, quantile regression, which solves the optimization problem of (5) with a general simplex algorithm, was known to be computationally expensive. Barrodale and Roberts (1973) developed a faster simplex algorithm according to the special structure of the design matrix for median regression. Koenker and d'Orey (1993) extended this special version of the simplex algorithm to quantile regression for any τ. However, in large statistical applications, the simplex algorithm is regarded as computationally demanding. In theory, the worst-case performance of the simplex algorithm shows an exponentially increasing number of iterations with sample size.

Since the general quantile regression fits nicely into the standard primal-dual formulations of linear programming, the powerful interior point algorithm can be applied. The worst-case performance of the interior point algorithm is proven to be better than that of the simplex algorithm. More important, experience has shown that the interior point algorithm is advantageous for larger problems (Portnoy and Koenker 1997).

Besides the interior point method, various heuristic approaches have been provided for computing L_1-type solutions. Among these approaches, the finite smoothing algorithm of Madsen and Nielsen (1993) is the most useful. By approximating the L_1-type objective function with a smoothing function, the Newton-Raphson algorithm can be used iteratively to obtain the solution after finite loops. The smoothing algorithm extends naturally to general quantile regression (Chen 2003). It turns out to be significantly faster for problems with a large number of covariates.

These three algorithms represent the most advanced algorithms for computing regression quantiles. Comparisons of these algorithms and other computational issues for quantile regression are described in Chen and Wei (2005). All these three algorithms have been implemented in the newly developed QUANTREG procedure in SAS, which can be downloaded from http://www.sas.com/statistics. The QUANTREG procedure also computes three types of confidence intervals for regression quantiles and conducts diagnostics and other statistical inferences. In the following sections, some usages of the procedure will be demonstrated. More details about this new procedure can be found in Chen (2005) and the documentation from the download site.

The purpose of this paper is to show how to use the quantile regression method to construct BMI growth charts, to introduce the new QUANTREG procedure, and to compare the BMI growth charts constructed using the recent four-year national cross-sectional survey data (1999−2002) with the CDC 2000 BMI reference growth charts.

2

Table 1: Summary Statistics

Variable	Q1	Median	Q3	Mean	SD	MAD
Men						
WEIGHT	48.2000	70.3000	85.8000	67.0428	29.0151	26.0938
HEIGHT	153.9	169.9	176.9	160.7	24.9267	13.0469
AGE	12.5000	19.6667	46.0000	29.5916	21.9532	18.4090
BMI	19.1400	23.9000	28.3800	24.3102	6.5572	6.8496
Women						
WEIGHT	45.8000	60.4000	75.8000	60.2485	25.7518	22.2390
HEIGHT	149.1	158.1	163.8	151.4	20.3807	9.9334
AGE	12.5000	19.6667	47.0000	29.6765	22.0103	18.9032
BMI	19.2800	23.7200	29.4150	24.9298	7.5863	7.4278

2　Data

Since 1999, the National Center for Health Statistics has conducted a national health and nutrition examination (NHANES) survey annually. The survey data are released in a two-year cycle. The recent releases are NHANES 1999−2000 and NHANES 2001−2002. Each release includes several data files. To construct the growth charts of BMI, two data files are needed. One contains demographic variables, such as age, sex, race, income, etc. The other contains variables related to body measurements, such as height, weight, BMI, head circumference, etc. Each data file includes the respondent sequence number (SEQN), which identifies each individual. Different files in the same survey can be merged by this variable. The data files are in the binary XPT format and can be easily read and edited using the SAS editor.

After merging the two data files for each survey, the two merged files are combined to form the four-year (1999−2002) data set. Records for pregnant women are deleted. Then the following variables, WEIGHT(kg), HEIGHT(cm), BMI(kg/m^2), AGE(year), GENDER, and SEQN, are kept and the others are dropped. AGE was recorded in the best months for younger respondents (<20 years old) and in the best years for elders. Months are transfered to years in decimal. Records with missing values in any of the six variables are also deleted. The data set *bmi* contains all the remaining records with ages from 2 years to 80 years. There are 8,250 men and 8,053 women in *bmi*. Since men and women have different growth patterns, *bmi* is split into two data sets, *bmimen* and *bmiwomen*, by gender. Parallel growth charts of BMI are constructed for men and women with these two data sets.

It should be pointed out that the sample weight variable in the demographic data files is not included in this study. The sample weight is used to balance oversampling and undersampling in the surveys to make the data more representative for the national pupulation. This is certainly an important factor when national reference growth charts are constructed. In this paper, individuals in the two surveys are the target population and sample weight is not considered.

3　Preliminary Analysis

Table 1 displays the summary statistics of weight, height, age, and BMI for men and women in the two data sets *bmimen* and *bmiwomen*, respectively. Q_1 and Q_3 are the first and third quantiles. SD is the standard deviation and MAD is the median absolute deviation, which is a robust measure of the univariate scale. Large differences between SD and MAD indicate that there may be some outliers in that variable. Among the four variables, height shows the greatest differences between SD and MAD for both men and women.

These statistics are univariate. To explore the growth pattern of BMI, a preliminary median regression is used. The logarithm of BMI is used as the response. Although the logarithm transformation might not help the quantile regression fit, it might help the statistical inference on regression quantiles. The response is fitted with a parametric model, which involves six powers of AGE. These power terms are tested with the Wald and likelihood ratio tests for significance.

The median regression can be done with the QUANTREG procedure:

Table 2: Parameter Estimates with Median Regression: Men

Parameter	DF	Estimate	95% Confidence Limits	
Intercept	1	6.41816705	5.28206683	7.55426727
inveage	1	-1.1339904	-1.7752615	-0.49271930
sqrtage	1	-3.7649349	-4.7275936	-2.8022763
age	1	1.46718520	1.13480543	1.79956496
sqrtage*age	1	-0.24610559	-0.30024265	-0.19196854
age*age	1	0.01643716	0.01279510	0.02007923
age*age*age	1	-0.00003114	-0.00003836	-0.00002392

```
proc quantreg data=bmimen algorithm=interior ci=resampling;
    model  logbmi = inveage sqrtage age sqrtage*age age*age age*age*age
                 / diagnostics cutoff=4.5 quantile=.5;
    id SEQN age weight height bmi;
    test_age_cubic: test age*age*age / wald lr;
run;
```

where $inveage = age^{-1}$, $sqrtage = age^{\frac{1}{2}}$.

Table 2 displays the estimated parameters and their 95% confidence intervals based on the covariance computed by the resampling method. The resampling method implements the Markov chain marginal bootstrap (MCMB) of He and Hu (2002), which greatly promotes the bootstrap method for quantile regression. More details about confidence intervals of regression quantiles by MCMB and other methods can be found in Kocherginsky, He, and Mu (2005) and Chen and Wei (2005). All confidence intervals do not include 0, which indicates that these parameters are significant at 5% test level.

Median regression (and general quantile regression) is robust to extremes of the response variable. The QUANTREG procedure provides a diagnostic table of outliers. With CUTOFF=4.5, 22 respondents (0.27% of 8,250 men) are identified as outliers.

Median regression with *bmiwomen* can be done in parallel fashion. The results are similar with *bmimen*. With CUTOFF=4.5, only 1 female respondent is identified as an outlier. With CUTOFF=3.5, which corresponds to a probability of less than $0.25E-3$ under normality, 34 female respondents (0.42% of 8,053), compared with 83 (1.01% of 8,250) male respondents, are identified as outliers.

The same six power terms are tested as significant in median regression with *bmiwomen*. This model used for median regression is used for other quantiles in order to construct the growth charts for both males and females.

4   BMI Growth Charts

The parametric method with the six powers in AGE identified in the preliminary analysis is used to construct the growth charts of BMI. This can be done with the following SAS macro.

```
%macro quantiles(NQuant, Quantiles);
  %do i=1 %to &NQuant;
    proc quantreg data=bmimen ci=none algorithm=interior;
        model logbmi = inveage sqrtage age sqrtage*age age*age age*age*age
                    / quantile=%scan(&Quantiles,&i,'','');
        output out=outp&i pred=p&i;
    run;
  %end;
%mend;
```

The macro can be used for any number of quantiles. The following statements request fitted values for 10 quantiles ranging from 0.03 to 0.97.

```
%let quantiles = %str(.03,.05,.10,.25,.5,.75,.85,.90,.95,.97);
%quantiles(10,&quantiles);
```

For each quantile, the regression quantile is computed as in median regression and fitted values of BMI are computed at each observed age and saved in an output data set. These fitted values are plotted against age to create the percentile curves.
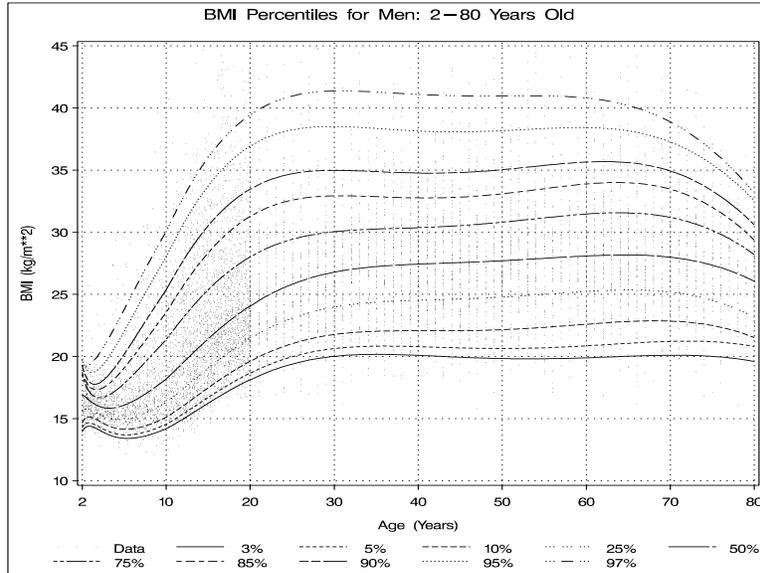

Figure 1: BMI Growth Percentile Curves with Polynomials

The 10 percentile curves together with the scatter plot of BMI for men are shown in Figure 1, and correspondingly for women. For both men and women, starting from age 2, BMI has a small drop followed by a quick growth until age 20. For men, BMI becomes stable at around age 30, while for women the growth continues to around age 40. Beginning from age 10, girls show a faster growth rate of BMI for all upper quantiles and the median. This prevalence continues to around age 40 and results in a larger proportion of women with higher BMI. This is also shown by the stronger concavity of the 95th and 97th percentile curves for women. For both men and women, BMI decreases from age 70. The BMI growth patterns for both men and women suggest that an effective way to control overweight in a population is to start in childhood.

The 10 quantiles selected here are the 10 quantiles used in the CDC 2000 BMI reference charts; see Kuczmarski et al. (2002). The CDC 2000 BMI reference charts were constructed from a much larger data pool, which consists of data from five previous national surveys between 1963 to 1994. The percentile curves in these reference charts are developed from two stages. In the smoothing stage, smoothed empirical percentiles are computed with a locally weighted regression on the empirical percentiles of segmented age intervals. In the transformation stage, these smoothed empirical quantiles are used to fit the LMS model with a set of 10 equations, which corresponds to the 10 selected quantiles. The three parameters λ, μ, and σ in the LMS model are solved for each age interval. The CDC 2000 BMI reference charts were constructed only for young people of 2 to 20 years old. The age intervals have a length of one month each. Interpolation of these parameters is suggested for finer age intervals.

This smoothing-transformation-based method requires large and balanced data, such that there are enough data in each segmented age interval. For the four-year survey data used here, age was recorded in the best months for young people and in the best years for elderly people. To construct BMI growth charts for all ages, such an interval segmentation is hard to achieve. However, quantile regression, which does not involve any age segmentation, uses all data to fit each percentile curve.

Although different methods are used for the BMI growth charts with the recent four-year survey data here from the CDC 2000 BMI reference charts, these charts are comparable. This comparison reveals important nationwide health information, especially for children. Growth pattern is more easily changed for young people by growth-related environmental factors, such as nutrition, dietary habits, exercise, and so on.

The comparison is summarized as follows:

(a) For boys of 2 to 20 years old, the upper percentile curves in Figure 1 have large lifts compared to the upper percentile curves in the CDC 2000 BMI reference charts. The 97th percentile for 10-year-old boys in Figure 1 is about 6.4 BMI units higher (an increase of 27%); and the 85th percentile is 4.4 BMI units higher (an increase of 23%), respectively.

(b) For girls of 2 to 20 years old, the upper percentile curves (not diaplayed) also show large lifts compared to the upper percentile curves in the CDC 2000 BMI reference charts. The 97th percentile for 10-year-old girls is about 5.5 BMI units higher (an increase of 22.5%); and the 85th percentile is about 4.0 BMI units higher (an increase of 20%), respectively.

(c) For boys of 2 to 20 years old, the median and lower percentile curves in Figure 1 do not have too much lift compared to the upper percentile curves in the CDC 2000 BMI reference charts. The median for 10-year-old boys in Figure 1 is about 1.5 BMI units higher (an increase of 9%); and the 3rd percentiles in these two charts are almost equal.

(d) Similarly, for girls of 2 to 20 years old, the median and lower percentiles do not show much difference in the two charts. The median for 10-year-old girls is about 1.4 BMI units higher (an increase of 8.5%); and again the 3rd percentiles in these two charts are almost equal.

(e) For both boys and girls, the growth pattern for the 95th and 97th percentile curves in Figure 1 looks different from that of the CDC BMI reference charts. The curves in are concave in the neighborhood of age 20, but the corresponding two curves in the CDC 2000 BMI reference charts are not.

(f) The percentile curves in Figure 1 have some boundary effect at age 2 because there are not enough data at age 2 or closeby.

Results from (a) and (b) can be interpreted as a warning of overweight or obesity for boys and girls in the four-year survey. These results are consistent with the empirical results of Hedley et. al. (2004). Similar warnings have been issued in other countries, e.g., in Switzerland by Zimmermann et al. (2004). (c) and (d) indicate that BMI does not increase for the entire population over time. That indicates that overweight and obesity can be controlled.

The cause for (e) is not so clear. It might be due to the different models used for the two sets of charts, or it might be due to the different data used. The boundary effect in (f) might be reduced by using some smoothing techniques in nonparametric quantile regression, which will be discussed in the next section.

5 Discussion

We have introduced an approach to construct growth charts of BMI using parametric quantile regression. This approach can also be used to construct growth charts for other age-related (time-related) variables, such as height, weight, head circumference, and other medical measurements. If the target data is used as reference data, such growth charts constructed using quantile regression can be used as reference growth charts.

Instead of using polynomials to fit the percentile curves, cubic B-splines based on age can be used in quantile regression as a method of nonparametric quantile regression. Wei et. al. (2004) have used such a nonparametric quantile regression method to construct reference growth charts for Finnish children. For the four-year national survey data, we tried the nonparametric quantile regression method with B-splines. For the BMI data of men, we selected the knots $\{2.5, 5, 10, 20, 35, 55, 75\}$. Based on these knots, the basis B-spline functions can be generated. Refer to Chen (2005) for how to generate the basis B-spline functions in SAS. By replacing the powers of age in the macro with these basis variables, percentile curves can be drawn.

Compared to the parametric quantile regression method, the nonparametric quantile regression method with B-splines needs extra effort. First, it needs a method to select the knots, or some other methods to fix the basis B-spline functions. Too many knots or too many basis functions provide a good fit of the percentile curves to the data, but make the curves too erratic. Second, a large amount of data is usually needed to make the constructed percentile curves stable to some changes of parameters.

The nonparametric quantile regression method with B-splines does not reduce the boundary effect. Eilers and Marx (1996) introduced a new smoothing technique using B-splines, which is called the P-spline approach. We are investigating the performance of the P-spline approach in nonparametric quantile regression, especially in constructing growth charts.

References

[1] Barrodale, I. and Roberts, F.D.K. (1973), An improved algorithm for discrete l_1 linear approximation, *SIAM J. Numer. Anal.,* 10, 839-848.

[2] Chen, C. (2003), A finite smoothing algorithm for quantile regression, Preprint.

[3] Chen, C. (2005), An introduction to quantile regression and the QUANTREG procedure, *Proceedings of the Thirtieth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

[4] Chen, C. and Wei, Y. (2005), Computational issues on quantile regression, *Special Issue on Quantile Regression and Related Methods*, *Sankhya*, forthcoming.

[5] Cole, T.J. (1988), Fitting smoothed centile curves to reference data (with discussion), *JRSS A*, 151, 385-418.

[6] Cole, T.J. and Green, P.J. (1992), Smoothing reference centiles curves: The LMS method and penalized likelihood, *Statistics in Medicine*, 11, 1305-1319.

[7] Eilers P.H.C. and Marx, B.D. (1996), Flexible smoothing using B-spline and panelized likelihood, *Statistical Science*, 11, 89-121.

[8] Flegal, K.M. (1999), Curve smoothing and transformations in the development of growth curves, *American Journal of Clinical Nutrition*, 70, 163S-5S.

[9] He, X. and Hu, F. (2002), Markov chain marginal bootstrap, *Journal of the American Statistical Association*, 97, 783-795.

[10] Hedley, A.A., Ogden, C.L., Johnson, C.L., Carroll, M.D., Curtin, L.R., Flegal, K.M. (2004), Prevalence of overweight and obesity among US children, adolescents, and adults, 1999-2002. *JAMA*, 291(23), 2847-50.

[11] Kocherginsky, M., He, X., and Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14, 41-55.

[12] Koenker, R. and Bassett, G. (1978), Regression quantiles, *Econometrica*, 46, 33-50.

[13] Koenker, R. and d Orey, V. (1993), Computing regression quantiles, *Applied Statistics*, 43, 410-414.

[14] Kuczmarski R.J., Ogden C.L., Guo S.S., et al. (2002), 2000 CDC growth charts for the United States: Methods and development. *Vital Health Stat.*, 11, 246, 1-190.

[15] Madsen, K. and Nielsen, H.B. (1993), A finite smoothing algorithm for linear L_1 estimation, *SIAM J. Optimization*, 3, 223-235.

[16] Portnoy, S. and Koenker, R. (1997), The Gaussian Hare and the Laplacian Tortoise: Computation of Squared-error vs. Absolute-error Estimators, *Statistical Science*, 12, 279-300.

[17] Wei, Y., Pere, A., Koenker, R., and He, X. (2004), Quantile regression methods for reference growth charts. Preprint.

[18] Zimmermann, M.B., Gübeli, C., Püntener, C., and Molinari, L. (2004), Overweight and obesity in $6-12$ year old children in Switzerland, *Swiss Med. Wkly.*, 134, 523-528.