ANNALS OF HUMAN BIOLOGY, 1990, VOL. 17, NO. 6, 475–481

# Non-parametric estimation of age-related centiles over wide age ranges

HUI QI PAN

Department of Mathematics and Biostatics, Shanghai Second Medical University, Shanghai, People's Republic of China

HARVEY GOLDSTEIN

Department of Mathematics, Statistics and Computing, Institute of Education, London University

QI YANG

Department of Mathematics and Biostatistics, Shanghai Second Medical University, Shanghai, People's Republic of China

*Received October 1989; revised December 20 1989*

**Summary.** A new method for estimating age-related centile curves has been developed, which is suitable for measurement covering a wide age range. The method was used to calculate weight centile curves of 8995 children from birth to 6 years obtained by the Collaborating Centre for Physical Growth and Psychosocial Development of Children in Shanghai, China.

## 1. Introduction

Healy, Rasbash and Yang (1988) presented a method for distribution-free estimation of a set of age-related centiles. The method is satisfactory for weight, circumferences, etc., whose distributions at a fixed age are often non-Gaussian. In practice, when the age span is wide, for example for childrens' stature from birth to 6 years old, or when the velocity of growth changes markedly over the age range, this method tends not to be satisfactory. This has led us to propose a new method by extending the procedure of Healy *et al.*

## 2. Procedure

The method of Healy *et al.* (1988) involves two steps. The first step is to obtain 'raw' centile estimates as follows. We define an age 'window', typically containing at least 50 data points. Starting from one end of the age scale this window is moved along the scale, one point at a time, to give overlapping sets of points. For each such set a ranking method estimates the centiles, giving an initial set of 'raw' centiles. Using these initial estimates, a polynomial of sufficiently high order is fitted to each centile. The intercept, linear, quadratic, etc., polynomial coefficients are then separately 'smoothed', effectively by regressing them on the centile normal equivalent deviates (NED). The final smoothed centiles are predicted from these values.

*Smoothing the initial estimates*

The raw estimates are as obtained by Healy *et al.* (1988) for the whole age range. These estimates are irregular and need to be smoothed to provide centile curves.

It is assumed that the curves can be fitted by a polynomial of degree $p$. If $t$ denotes age and $Y_{i,t}$ the smoothed value of the $i$th centile, we have

$$Y_{i,t} = a_{0,i} + a_{1,i}\, t + \ldots + a_{p,i} t^p + a_{p+1,i}(t-c_1)_+^p + \ldots + a_{p+m-1,i}(t-c_{m-1})_+^p \quad (1)$$

where

$$(t-c_l)_+^p = (t-c_l)^p \quad when \quad t>c_l$$
$$= 0 \quad when \quad t \leq c_l \quad l=1,2,\ldots,m-1$$

and $c_l$ is the $l$th cut or join point between the age groups with $c_1 < c_2 \ldots < c_{m-1}$. Of course, when $m=1$, the new method is the same as that presented by Healy *et al.* For example, with one join point and polynomials of degree 3 we obtain

$$Y_{i,t} = a_{0,i} + a_{1,i}t + a_{2,i}t^2 + a_{3,i}t^3 \qquad \text{when } t \leq c_1$$

$$= a_{0,i} + a_{1,i}t + a_{2,i}t^2 + a_{3,i}t^3 + a_{4,i}(t-c_1)^3 \qquad \text{when } t > c_1$$

It is clear that the two components of the curve join smoothly at the join point, and this is generally true for curves of the form (1). The value of $m$ and $c_l$ can be chosen after inspection of the data and using existing knowledge (see Discussion).

The coefficients in equation (1) are now regressed on a polynomial function of the NEDs $Z_i$ of degree $q$.

$$a_{j,i} = b_{j,0} + b_{j,1}Z_i + \ldots + b_{j,q_j}Z_i^{q_j} + e_{j,i} \quad j=0,1,\ldots p+m-1 \qquad (2)$$

Combining (1) and (2) leads to a composite linear model for the initial centile curves whose coefficients can be estimated using ordinary least-squares (OLS) analysis. The values of $p$ and $q_l$ still have to be determined and in the next section we give some examples. In general, the value of $q_l$ will usually be higher for the low-order coefficients and may be zero for high-order ones. Some exploration will generally be needed to obtain optimum values for $p$ and $q_l$. A test of fit can be obtained by comparing the percentages of data points that fall between the centile curves for sub-ranges, with their expected values.

## 3. Examples

*Data*

In general, more than three centiles, typically the 3rd, 10th, 25th, 50th, 75th, 90th and 97th, would be estimated. Here, for the purpose of illustration, only the 3rd, 50th and 97th weight centile curves were calculated for 4690 male and 4305 female children from birth to 6 years old. These children were random samples from Shanghai and five other provinces in the southeast of China, from rural and urban districts. The data were obtained by the WHO Collaborating Centre for Physical Growth and Psychosocial Development of Children in Shanghai, China, the staff having been specially trained to ensure validity under the guidance of WHO, Maternal and Child Health Division, Geneva.

## 4. Results

Many combinations of $p$, $q$, $m$ and $c$ have been tried and four examples for discussion are presented.

*Example 1:* $p=3$, $m=2$, $c_1=12$ months, $q=2, 2, 1, 0, 0$, that is, $q_0=2$, $q_1=2$, $q_2=1$,

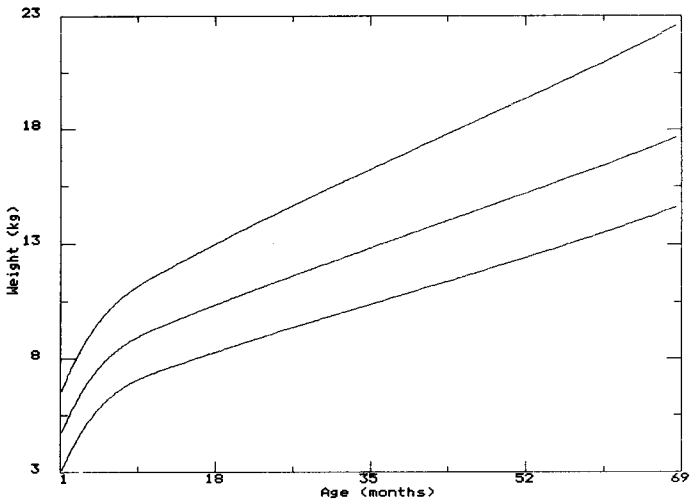


Figure 1. Example 1: 3rd, 50th and 97th centiles for males with $p=3$, $m=2$, $c_1=12$, $q=2, 2, 1, 0, 0$.

Table 1. Example 1: percentage for each subrange of age.

| | Age (months) | | | | | |
|---|---|---|---|---|---|---|
| | ≥1·5 | ≥12 | ≥24 | ≥36 | ≥48 | 60–67 |
| <p3 | 2·70 | 3·78 | 3·23 | 1·69 | 2·22 | 3·61 |
| p3–p97 | 94·49 | 92·86 | 92·90 | 95·78 | 95·15 | 92·78 |
| >p97 | 2·82 | 3·37 | 3·87 | 2·53 | 2·63 | 3·61 |
| Total | 100·0 | 100·0 | 100·0 | 100·0 | 100·0 | 100·0 |
| *N* | 1669 | 980 | 465 | 474 | 495 | 277 |

$q_3=0$, $q_4=0$, for 4690 male children, and there is one join point at 12 months. The full model is

$$Y_{i,t}=(b_{00}+b_{01}Z_i+b_{02}Z_i^2)+(b_{10}+b_{11}Z_i+b_{12}Z_i^2)t+(b_{20}+b_{21}Z_i)t^2+b_{30}t^3+b_{40}(t-12)_+^3$$

The smoothed curves are shown in figure 1. The number of data points that fall between the centile curves are given in table 1.

*Example 2:* $p=3$, $m=2$, $c_1=12$ months, $q=2, 1, 1, 0, 0$, for 4305 female children. The model is

$$Y_{i,t}=(b_{00}+b_{01}Z_i+b_{02}Z_i^2)+(b_{10}+b_{11}Z_i)t+(b_{20}+b_{21}Z_i)t^2+b_{30}t^3+b_{40}(t-12)_+^3$$

The smoothed curves with initial centile estimates are shown in figure 2. The counts in each category are given in table 2.

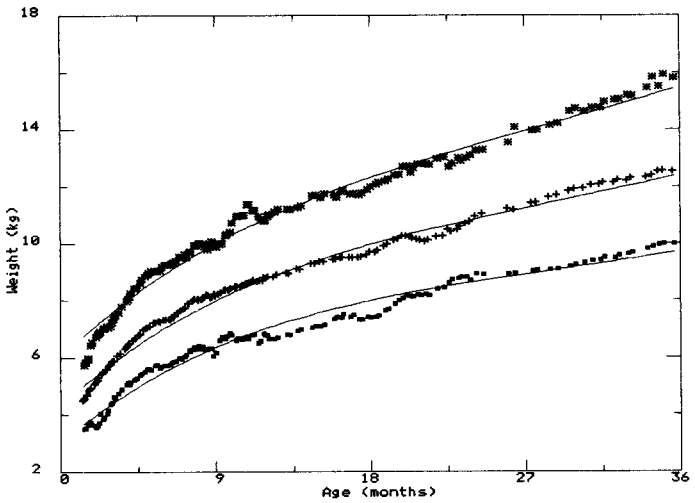


Figure 2. Examples 2: 3rd, 50th and 97th centiles for females with $p=3$, $m=2$, $c_1=12$, $q=2, 1, 1, 0, 0$.

Table 2. Example 2: percentage for each subrange of age.

| | Age (months) | | | | | |
|---|---|---|---|---|---|---|
| | ≥1·5 | ≥12 | ≥24 | ≥36 | ≥48 | 60–67 |
| <p3 | 2·59 | 3·57 | 2·17 | 3·32 | 1·84 | 2·55 |
| p3–p97 | 94·43 | 93·43 | 95·22 | 94·08 | 95·16 | 94·91 |
| >p97 | 2·98 | 3·00 | 2·61 | 2·61 | 3·00 | 2·55 |
| Total | 100·0 | 100·0 | 100·0 | 100·0 | 100·0 | 100·0 |
| *N* | 1543 | 868 | 460 | 422 | 434 | 275 |

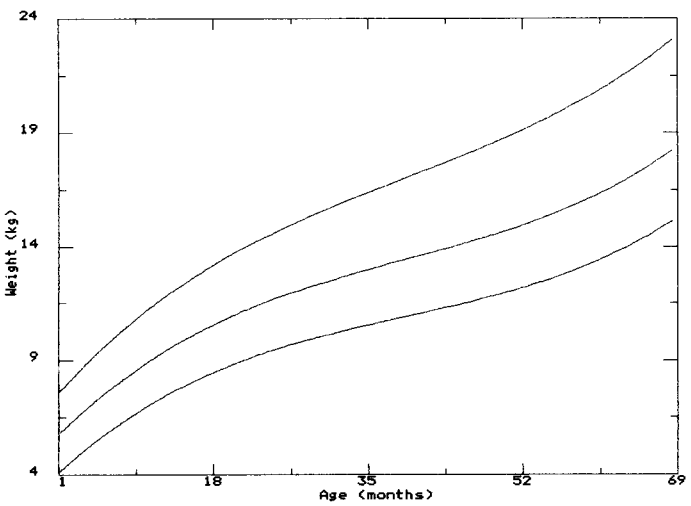


Figure 3. Example 3: 3rd, 50th and 97th centiles for males with $p=3$, $m=1$, $q=2, 2, 1, 0$.

Table 3. Example 3: percentage for each subrange of age.

| | Age (months) | | | | | |
|---|---|---|---|---|---|---|
| | ≥1·5 | ≥12 | ≥24 | ≥36 | ≥48 | 60–67 |
| <p3 | 2·22 | 5·10 | 4·95 | 1·48 | 1·62 | 4·69 |
| p3–p97 | 93·59 | 91·84 | 91·83 | 95·99 | 95·56 | 92·06 |
| >p97 | 4·19 | 3·06 | 3·23 | 2·53 | 2·83 | 3·25 |
| Total | 100·0 | 100·0 | 100·0 | 100·0 | 100·0 | 100·0 |
| *N* | 1669 | 980 | 465 | 474 | 495 | 277 |

*Example 3:* $p=3$, $m=1$, $q=2, 2, 1, 0$, for 4690 male children. Here, $m=1$ means there is no join point. The model is

$$Y_{i,t}=(b_{00}+b_{01}Z_i+b_{02}Z_i^2)+b_{10}+b_{11}Z_i+b_{12}Z_i^2)t+(b_{20}+b_{21}Z_i)t^2+b_{30}t^3$$

The smoothed curves are shown in figure 3 and the counts are given in table 3.

*Example 4:* $p=4$, $m=1$, $q=2, 1, 1, 0, 0$, for 4305 female children. The model is

$$Y_{i,t}=(b_{00}+b_{01}Z_i+b_{02}Z_i^2)+(b_{10}+b_{11}Z_i)t+(b_{20}+b_{21}Z_i)t^2+b_{30}t^3+b_{40}t^4$$

The curves with initial centile estimates are given in figure 4 and the counts are in table 4.

The curves for males and females with one join point appear to fit the data well. This is clear from figures 1 and 2 and from tables 1 and 2 where the percentages in the centile bands do not deviate appreciably from their expectations and a formal test of fit yields a non-significant result at the 5% level. By contrast, in model 3 with no join point, figure 3 shows an upward curvature at the oldest ages and an overall poor fit as is seen in table 3. Example 4 also fits a single curve, but with a fourth-order term in order to introduce an extra parameters, thereby giving the same number of parameters as models 1 and 2. It provides a better fit than model 3 at the older ages but a poor fit at the younger ones, which is also apparent from table 4.

## 5. Discussion

We have demonstrated how one of the limitations of the Healy *et al.* (1988) method for centile estimation can be overcome by introducing extra parameters into the polynomial smoothing. However, the approach of simply increasing the order of the polynomial does not work well in general as is illustrated in our examples. The proposed method allows a wide age range to be fitted using relatively low-order polynomials. The choice of the number and placing of join points in this method is a matter for experimentation. In our examples several ages were tried for a single join point in the range 9–12 months. This range was chosen because it is where the growth velocity decreased rapidly. The results were not sensitive to the precise choice of age.

A generalization of the procedure described here (Goldstein and Pan 1990) allows different order polynomials to be fitted within each age range and also allows simultaneous estimation for different population subgroups. Both procedures can be fitted using the GROSTAT II program (Rasbash 1989).

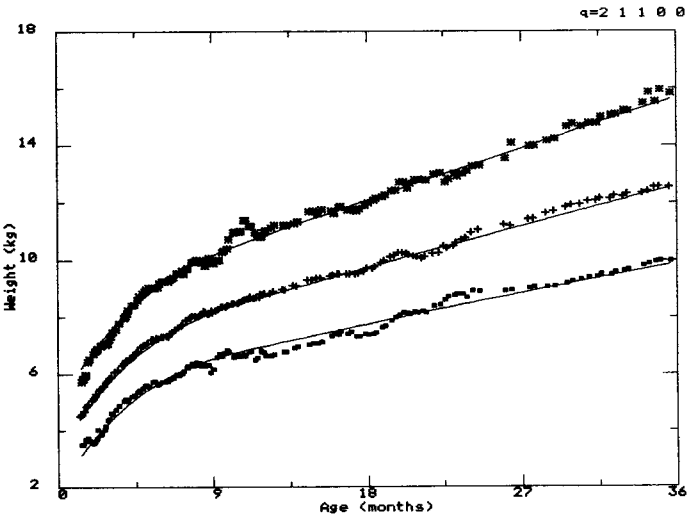


Figure 4. Example 4: 3rd, 50th and 97th centiles for females with $p=4$, $m=1$, $q=2, 1, 1, 0, 0$.

Table 4. Example 4: percentage for each subrange of age.

| | Age (months) | | | | | |
|---|---|---|---|---|---|---|
| | ≥1·5 | ≥12 | ≥24 | ≥36 | ≥48 | 60–67 |
| <p3 | 2·20 | 5·07 | 1·96 | 2·84 | 2·07 | 2·91 |
| p3–p97 | 93·84 | 93·09 | 94·57 | 94·31 | 95·16 | 94·91 |
| >p97 | 3·95 | 1·84 | 3·48 | 2·84 | 2·76 | 2·18 |
| Total | 100·0 | 100·0 | 100·0 | 100·0 | 100·0 | 100·0 |
| *N* | 1543 | 868 | 460 | 422 | 434 | 275 |

## Acknowledgements

We are most grateful to Dr Guo Di, and Professor Michael Healy who have provided invaluable comments on the paper. Many thanks are due to Mr Jon Rasbash and Mr Robert Prosser who assisted in this effort. We are also grateful to the staff of the WHO Collaborating Centre in Shanghai for the data. This work was partly carried out while the first author was in receipt of a WHO fellowship at the Institute of Education, University of London.

## References

HEALY, M. J. R., RASBASH, J. and YANG, M., 1988, Distribution-free estimation of age-related centiles. *Annals of Human Biology*, **15**, 17–22.

DEPARTMENT OF PROBABILITY AND STATISTICS AT THE CHINESE SCIENCE IINSTITUTE, 1979, *Probability, Statistics and Computation*. (Beijing: Science Press), 175–181.

GOLDSTEIN, H. and PAN, H., 1990, Percentile smoothing using piecewise polynomials, with covariates. (Submitted for publication).

RASBASH, J., 1989, *GROSTAT II: A Program for the Construction of Age Related Centiles*. WHO Collaborating Centre on Growth and Development, University of London, London WC1, England.

Address for correspondence: Hui Qi Pan, Department of Mathematics Statistics and Computing, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL.

**Zusammenfassung.** Es wurde eine neue Methode zur Schätzung der altersverknüpften Perzentilkurven entwickelt, die für Maße geeignet sind, die eine breite Altersspanne abdecken. Sie wird benutzt für die Berechnung von Gewichtsperzentilkurven von 8.995 Kindern von der Geburt bis 6 Jahre aus dem Kooperationszentrum für Körperwachstum und psychosoziale Entwicklung von Kindern in Shanghai in China.

**Résumé.** Une nouvelle méthode d'estimation des courbes de centile en fonction de l'âge a été élaborée, afin de convenir à des mensurations échelonnées sur une longue gamme d'âge. Elle est utilisée pour le calcul des courbes de centile du poids de 8995 enfants, depuis leur naissance jusqu'à 6 ans, qui ont été rassemblées par le Centre de Collaboration pour la Croissance Physique et le Développment Psychosocial de Shangaï (Chine).