Functional Analysis of Childhood BMI based on Simplicial Band Depth

Yutong Ren[1]

[1]Beijing City International School, Beijing China

Abstract. This paper mainly studied the relationship between obesity and BMI. BMI is the ratio of height to weight and measures a person's health. In this paper, height and weight data of obese children from 2 to 18 years old in local hospitals were collected. The linear regression method was used to find the correlation between BMI and children's health degree, and the regression curve between age and BMI was plotted. This paper supplemented the vacancy of BMI study on children, and a complete BMI curve could be obtained for each child according to the curve, so as to predict possible health problems of children in advance.

1 Introduction

Obesity is a common global health disease [1], it can cause a lot of health risks, such as coronary heart disease, cancer, stroke and premature death. In clinical treatment, the importance of identifying obesity as a disease is very important for the treatment of such patients. Accurate diagnosis of obesity in the elderly is an important first step in providing effective treatment for high-risk groups.

Health problems highly related to obesity include high blood glucose, diabetes, high blood pressure, hypertension, high blood cholesterol, triglycerides, dyslipidemia, coronary heart disease, heart failure, and stroke. Another bone and joint problems due to the pressure that weight puts on to the bones and joints, which can lead to osteoarthritis, and daytime fatigue and sleepiness, poor attention, and problems at work due to stopping breathing during sleep.

BMI is short of Body Mass Index, which is a value derived from the mass and height of a person. The BMI is defined as the weight divided by the square of height. It is universally expressed in units of kg/m^2. A high BMI can be an indicator of upper body fatness. BMI cannot diagnose personal obesity or personal health.

1.1 Exploring the association between BMI and obesity

People usually Table 1 as a reference to illustrate the association between BMI and obesity. The biological causes of childhood obesity seem to be multifactorial. Eating junk foods and overeating is easy in nowadays society. Emotions, habits, food acquisition and many other factors can affect eating behavior. Modern conveniences such as elevators, cars and TV remote controls paralyze our lives.

Table1. The BMI and Obesity Categories

BMI	Obesity Categories
<18.5	Underweight
18.5 – 24.9	Normal
25 – 29.9	Overweight
30+	Obese

Genetic composition has a major influence on body weight. It will affect the energy consumption rate of your body in a static state, that is, the basal metabolic rate. Some people are born with a higher basic metabolic rate than others. So they burn more calories than others. Regular physical exercise can increase a person's metabolic rate. The lower metabolic rate makes it easier to gain weight. Fat distribution also has an impact. For example, men have fat stored in the abdomen, while women have fat stored in the hips and thighs. Some studies also show a strong association between birth weight and childhood obesity.

1.2 BMI curves and functional data analysis

The curve is a smooth, increasing line as the kids grow up. Some kids may have specific time points, but not all, so a powerful statistical tool is essential for our analysis. Functional data analysis (FDA) deals with the analysis and theory of data that are in the form of functions, images, and shapes or more general objects. Essentially, functional data is infinite dimensional. The high inherent dimensionality of these data poses considerable theoretical and computational challenges. These challenges change due to the way the functional data is sampled, and at the same time bring many opportunities for research and data analysis.

The FDA's method and model can be flexibly modeled because it is basically non-parametric. FDA's statistical tools include smoothing based on sequence expansion,

jenny.ding@ebcn.co, 2018326080@bcis.cn

penalty spline or partial polynomial smoothing, and functional principal component analysis. The smoothing method is different from the FDA. Smoothing is usually used in the following situations: you want to obtain an estimate of a non-random object (here the object is a function or surface) from noisy observations, while the FDA aims to analyze random samples Objects, it can be assumed that they are observed completely without noise or sparsely observed with noise and many interesting scenes can be found from them. Application areas that have been emphasized in the statistical literature include growth curves [2], econometrics and ecommerce [3], evolutionary biology [4], and genetics and genomics [5].

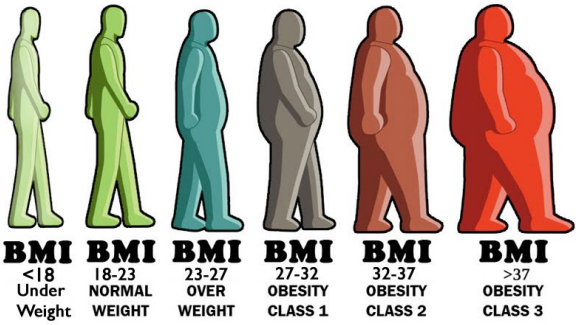

Fig1. BMI and Obesity

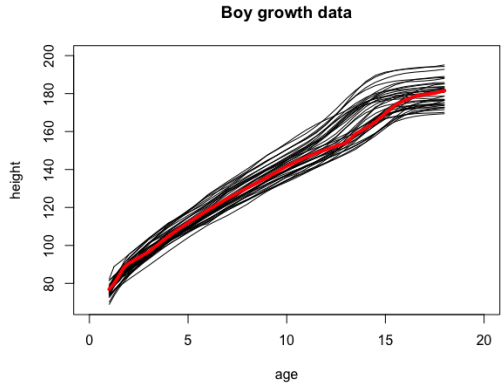

Fig2. The Boy Growth Data with functional median

2 Method

2.1 BAND DEPTH FOR FUNCTIONAL DATA

In the field of functional analysis, each subject is represented by $y_i(t), i = 1, \dots, n, t \in \Gamma$, where Γ is an interval in R. The band depth for functional data provides a method to order all the sample curves. Indeed, we can compute the band depths of all the sample curves and order them according to decreasing depth values. Let $y_{[i]}(t)$ denote the sample curve associated with the i th largest band depth value. We view $y_{[1]}(t), \dots, y_{[n]}(t)$ as order statistics, with $y_{[1]}(t)$ being the deepest (most central) curve or simply the median curve, and $y_{[n]}(t)$ being the most outlying curve. This indicates that a smaller level is associated with a higher center position compared to the sample curve. The order statistics caused by the belt depth start from the centermost sample curve and move outward in all directions.

Under this basic idea, Oldford introduced the band depth concept through a graph based approach [6]. The graph of a function $y(t)$ is the subset of the plane $G(y) = (t, y(t)) : t \in \Gamma$. The band in R^2 delimited by the curves $y_{i1}, \dots, y_{ik}$ is $B(y_{i1}, \dots, y_{ik})$, which is

$$\big(t, x(t)\big): t \in \Gamma, \min_{r=1,\dots,k} y_{i_r}(t) \le x(t) \le \max_{r=1,\dots,k} y_{i_r}(t) \quad (1)$$

Let J be the number of curves determining a band, where J is a fixed value with $2 \le J \le n$. If $Y_1(t), \dots, Y_n(t)$ are independent copies of the stochastic process $Y(t)$ generating the observations $y_1(t), \dots, y_n(t)$, the population version of the band depth for a given curve $y(t)$ with respect to the probability measure P is defined as

$$BD_j(y, P) = BD^{(2)}(y, p) = P\{G(y) \subset B(y_1, y_2)\} \quad (2)$$

where $B(Y_1, \dots, Y_j)$ is a band delimited by j random curves. The sample version of $BD^{(j)}(y, P)$ is obtained by computing the fraction of the bands determined by j different sample curves containing the whole graph of the curve $y(t)$. In other words,

$$BD_n^{(j)} = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} I\{G(y) \subseteq B(y_{i1}, \dots, y_{ij})\} \quad (3)$$

where $I\{\cdot\}$ denotes the indicator function. The implication is that by computing the fraction of the bands containing the curve $y(t)$, the bigger the value of band depth, the more central position the curve has. Then, the sample band depth of a curve $y(t)$ is

$$BD_{n,j}(y) = \sum_{j=2}^{B} D^{(j)}(y) \quad (4)$$

Instead of considering the indicator function, López-Pintado and Romo also proposed a more flexible definition [7], the modified band depth (MBD), by measuring the proportion of time that a curve $y(t)$ is in the band:

$$MBD_n^{(j)} = \binom{n}{j}^{-1} \sum_{1 \le i_1 \le i_2 < \cdots < i_j \le n} \lambda_r\{A(y; y_{i1}, \dots, y_{ij})\} \quad (5)$$

where

$$A_i(y) \equiv A\big(y; y_{i_1}, \dots, y_{i_1}\big) \equiv \{t \epsilon \Gamma : min_{r=i_1,\dots,i_j} y_r(t) \le y(t) \le max_{r=i_1,\dots,i_j} y_r(t)\} \quad (6)$$

and

$$\lambda_r(y) = \lambda(A_i(y))/\lambda(y) \quad (7)$$

if λ is the Lebesgue measure on Γ. If $y(t)$ is always inside the band, the modified band depth degenerates to the band depth in (2.1).

After considering the ratio of the curve in the band, the band depth is modified, so that it avoids a lot of depth constraints, and it is more convenient to obtain the most representative curve in amplitude. The shape of the curve that is often linked determines the depth of the band, and therefore can be used to obtain the most representative curve in terms of shape. So there are two types of outliers: quantity outliers and shape outliers. In general, when the amplitude outlier is far away from the mean, the pattern of the shape outlier is different from other curves.

A sample median function is a curve from the sample with largest depth value, defined by arg $max_{y\epsilon y_1,\ldots,y_n} BD_n, J(y)$. If there are ties, the median will be the average of the curves maximizing depth.

Although the number of curves determining a band, j, could be any integer between 2 and J, the order of curves induced by band depth is very stable in J. To avoid computational issues, we use $J = 2$, and for simplicity, we write $BD_n^{(2)}$ as BD and $MBD(2)$ as $MBD_N^{(2)}$ as MB in the sequel.

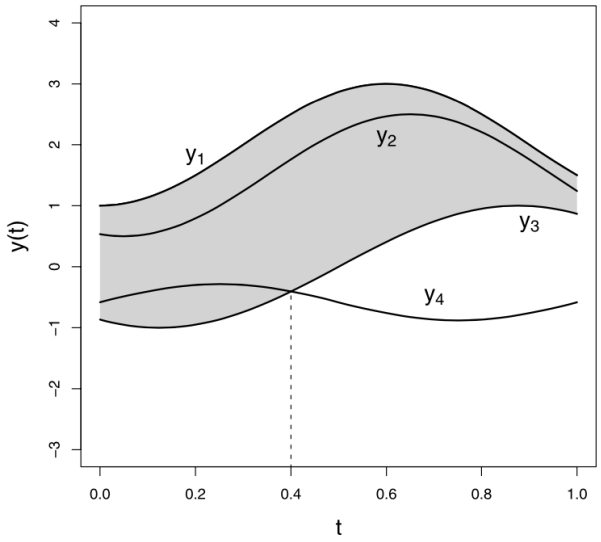

Fig3. Illustration of modified band depth, cited by L**ó**pez-Pintado and Romo

Figure 3 provides a simple example with $n = 4$ curves on how to compute BD and MBD in practice. When $J = 2$, there are six possible bands delimited by two curves. For instance, the gray area in Figure 3 is the band delimited by $y_1(t)$ and $y_3(t)$. We can see that the curve $y_2(t)$ completely belongs to the band, but $y_4(t)$ only partly does. We define that a curve is contained in a band even if this curve is on the border of the band. Then $BD(y_2) = 5/6 = 0.83$ since only the band delimited by $y_3(t)$ and $y_4(t)$ does not completely contain the curve $y_2(t)$ and $BD(y_4) = 3/6 = 0.5$ as it is only completely contained in the bands delimited by itself and another curve. Similarly, we could compute $BD(y_1) = 0.5$ and $BD(y_3) = 0.5$. To compute MBD, note that the curve $y_2(t)$ is always contained in the five bands, hence $MBD(y_2) = 0.83$, the same value as BD. In contrast, the curve $y_4(t)$ only belongs to the band in gray 40/100 of the time, thus $MBD(y_4) = (3 + 0.4 + 0.4)/6 = 0.63$ by definition. For the other two curves, $MBD(y_1) = 0.5$ and $MBD(y_3) = 0.7$.

2.2 CONSTRUCTION OF FUNCTIONAL BOXPLOTS

In the classical boxplot, the box itself represents the middle 50% of the data. An interesting idea that can be extended to functional data is the concept of central region introduced by Liu et al. [8]. The band delimited by the α proportion ($0 \leq \alpha \leq 1$) of deepest curves from the sample is used to estimate the central region. In particular, the sample 50 central region is

$$C_{0.5} = \big(t, y(t)\big): \min_{r=1,\ldots,[n/2]} y_{[r]}(t) \leq y(t) \leq \max_{r=1,\ldots,[n/2]} y_{[r]}(t) \quad (8)$$

where $n/2$ is the smallest integer not less than $n/2$. The border of the 50% central region is defined as the envelope representing the box in a classical boxplot. Thus, this 50% central region is the analog to the "inter-quartile range" (IQR) and gives a useful indication of the spread of the central 50% of the curves. This is a robust range for interpretation because the 50% central region is not affected by outliers or extreme values, and gives a less biased visualization of the curves' spread. There is also a curve in the box that indicates the median $y_{[1]}(t)$, Or the most central curve with the largest band depth value. The median curve can also be used to measure centrality. The "whisker" of the box plot is the vertical line of the plot, which extends from the box and represents the maximum envelope of the data set except for outliers. Therefore, we first need to identify outliers. Similarly, we extend the 1.5 times IQR experience outlier criterion to the functional box plot. The fence is obtained by expanding the envelope of the central area by 50% by 1.5 times the range of the central area. And mark all the curves outside the fence as potential outliers, as shown in Figure 4. It should be noted that when each curve is just one point, the functional box plot will degenerate into a classic box plot. We recommend using a constant coefficient of 1.5 in the classic boxplot, but the user can modify it.

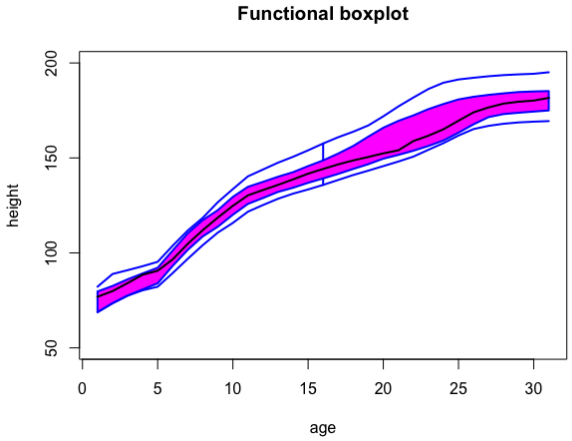

Fig4. The Boy Growth Data and Functional Boxplot Visualization

Now that the various parts of the function box diagram have been determined, we begin to explain its structure on the data set used.

3 Applications

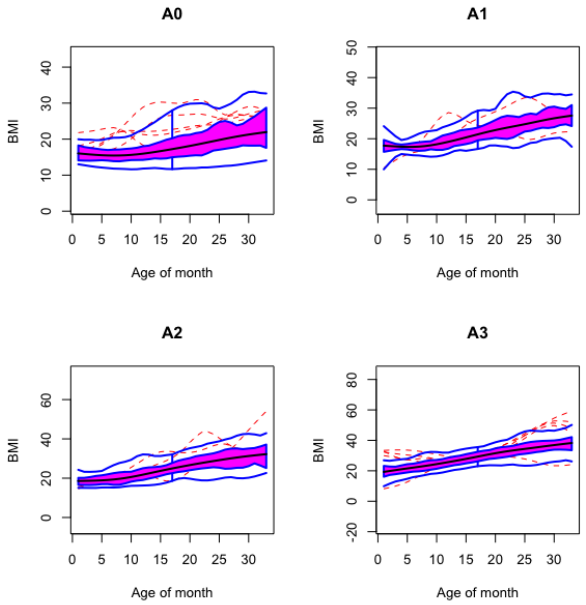

Fig5. The Functional Boxplot Visualization for Four Obesity Categories

We conducted a secondary data analysis of the BMI records collected from Boston Children Hospital. Our data contains healthy and obese children from 2 years to 18 years old. We record their corresponding height and weight in order to get their BMI over time. The original dataset is stored in a long-format. We use R to transform them into wide-format, which contains 21150 children. However, our dataset is sparse; in other words, not all the kids have contact records from 2 to 18 years. Therefore, a necessary statistical smoothing technique is applied to simulate curves for the sparse BMI data.

The commonly used methods are linear regression (linear, Quadratic, or Cubic regression) and spline model. In our setting, we an advanced cubic smoothing spline model that simulates the new curves of BMI by borrowing information from nearby curves. After getting the fitted curve, we do the following in-depth analysis.

Experts often rely on BMI to determine if a person is overweight. The BMI estimates the level of ones' body fat based on their height and weight. Starting at 25.0, the higher the BMI, the greater risk they have of developing obesity-related health problems. However, a functional curve instead of pointwise analysis provides a better measurement of obesity. The graphs above show four groups of people in different obesity categories. A0 describes the average normal population while A1, A2, and A3 described slightly overweight, obese, and severely obese people, respectively. By using the functional boxplot based on depth measurement, we have an excellent performance of visualizing their BMI trends. The function shows the median (black curve in the middle), 50% deepest curve (the magenta region), the whisker (the boundary for outlier detection), and outliers in magnitude (the dashed red line). A comparison of Figure 5 is shown in the following. In the spread perspective, A0 has the most significant variability, the others get narrower and narrower orderly, while A3 has the least variability. Another very worth noticing phenomenon is the rebound appears, especially in the A0 group. This means when the kids are about 5 months to 10 months old, they would learn to walk and do other exercises, which would cause a decrease in BMI accordingly. But children who are obese do not have this trend. The total trend is the BMI increases with age. A0 is growing relatively very slowly. However, the fatter people were, the faster their BMI increased. The A3 curve almost became a straight line that continued to rise with the steepest degree. The median of A0 is the smallest, but A3 is the greatest. Also, A0 has the most outliers compared to others.

4 Conclusion

In conclusion people who have obesity are at increased risk for many serious disease and health conditions. BMI is an estimate of body fat and a good measure of ones' risk for health problems. Based on this point-wise measure, the functional BMI is a fast and convenient method to represent the obesity conditions throughout the age.

However, in the real situation, many records for children's BMI are sparse, most of them have only a few periods of records but are lack of points in certain ages. The main contribution in this paper is that we use cubic spline model to estimate the whole throughout their age and get a smooth and intact curves for each child. After getting the estimated curve, we analyze the processed data based on depth measure, which is a powerful tool for ranking and outlier detection of functional data. We also construct boxplot based on our ranking results for different obesity groups and compare their summary statistic. A further and thoroughly analysis of association between obesity and spare BMI curve is also presented.

REFERENCES

1. Ogden, C. L., Carroll, M. D., Kit, B. K., and Flegal, K. M. (2011). Prevalence of childhood and adult obesity in the united states. Survey of Anesthesiology, 58(4):206.
2. Altman, N. S. and Casella, G. (1995). Nonparametric empirical bayes growth curve analysis. Publications of the American Statistical Association, 90(430):508–515.
3. Ramsay, J. O. and Silvermann, B. W. (1998). Functional data analysis. springer series in statistics. Biometrical Journal, 40(1):56–56.
4. Heckman, N. E. (2003). Functional data analysis in evolutionary biology. Recent Advances and Trends in Nonparametric Statistics, pages 49–60.
5. Salazar, M., Vongsangnak, W., Panagiotou, G., Andersen, M. R., and Nielsen, J. (2009). Uncovering transcriptional regulation of glycerol metabolism in aspergilli through genome wide gene expression data analysis. Molecular Genetics and Genomics, 282(6):571.

6. Oldford, R. W. (1998). Functional data analysis. International Encyclopedia of the Social Behavioral Sciences, 8(4):401–403.
7. L ó pez-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. Journal of the American Statistical Association, 104(486):718–734.Conference Name: ACM Woodstock conferenceConference Short Name: WOODSTOCK' 18Conference Location: El Paso, Texas USA
8. Liu, R. Y., Parelius, J. M., Singh, K., et al. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by liu and singh). The annals of statistics, 27(3):783–858.