A Linear Quantile Mixed Models Assessing the Relationship between Total Daily Energy Expenditure and Sex-adjusted BMI among Elementary School Students.

Xi Yang and Yuanyuan Luan

1. Introduction

Childhood obesity is a serious public health issue in the United States. It puts children at risk for poor health as well as poor health condition when they grow up to adults. Despite recent declines in the prevalence among preschool-aged children, obesity amongst all children is still too high. According to CDC, nearly 1 in 5 school age children and young people (6 to 19 years) in the United States has obesity from 2015-2016. Obesity could cause various health issues including type II debates, asthma, and osteoarthritis. About 90% of children diagnosed with type 2 diabetes are either overweight or obesity patients (Liu et al., 2010). Besides that, obesity is also considered to contribute to increase the risk for a number of cancers such as lung cancer and colon cancer. It is well known that obesity basically results from a chronic imbalance between energy expenditure, which is defined as the amount of energy used by the body to perform daily tasks such as breathing and digestion, and energy intake. Other important risk factors are environmental exposures and genetic predisposition. However, the exact role of reduced energy expenditure in obesity development is unclear (Bandini et al., 2004). In order to combat this growing health problem amongst children, researchers become increasingly interested in employing school-based interventions as targeted interventions designed to increase school day energy expenditure (SDEE), which is the total amount of energy or calories expended by the body to perform physical activity during the school day. A good example of such school-based intervention is the activity permissive learning environment (APLE) (Wechsler et al., 2000; Benden et al., 2014; Lanningham-Foster et al., 2008). Activity permissive learning environments introduce stand-biased desks into classrooms as a means of increasing physical activity among school-aged children.

The objectives of this paper are two-fold. First of all, we want to examine if measures of energy expenditure obtained at the prior time such as baseline can be used as screening standard for future obesity risk, such as one-year post-baseline. Secondly, we will describe the use of quantile mixed effect model to study the relationship between mean energy expenditure and the difference of BMI between the first semester and second semester.

2. Data and Materials

The data we used in this study was collected during a study named “stand-biased desk study”, which was conducted from 2012 to 2014 in three elementary schools within the College Station Independent School District (CSISD) (Benden et al., 2014). There were 374 students recruited at the beginning while only 193 of them have completed the study due to either graduate from elementary school or their parents retract their consent from the study. At the end the study, five students with large proportions of missing data were excluded from the final sample. Thus, the final analytic sample size was 188. The average school day energy expenditure (SDEE) was 2.48

per minute. Average age of the participants were about 7 years old. Half of the participants were boys (Table 1).

Each student's height and weight were obtained at the start of each semester by trained research assistants to calculate their BMI. All the participants of the study were required to wear calibrated BodyMedia SenseWearR Armband devices (BodyMedia, PA) during the school hours for a week for each semester from fall 2012 to spring 2014. The devices recorded subject-specific steps counts and caloric energy expenditure per minute while worn. If we plot the mean school day energy expenditure (SDEE) of each student at each day across the whole study period against the step counts (Figure 1.), the plots are very messy, which indicates a large data set. Therefore, instead of using data per minutes, we decide to use data per hour for calculation purpose. The distribution of BMI was concentrated within the first two quantiles (Figure 2) at baseline (first fall semester).

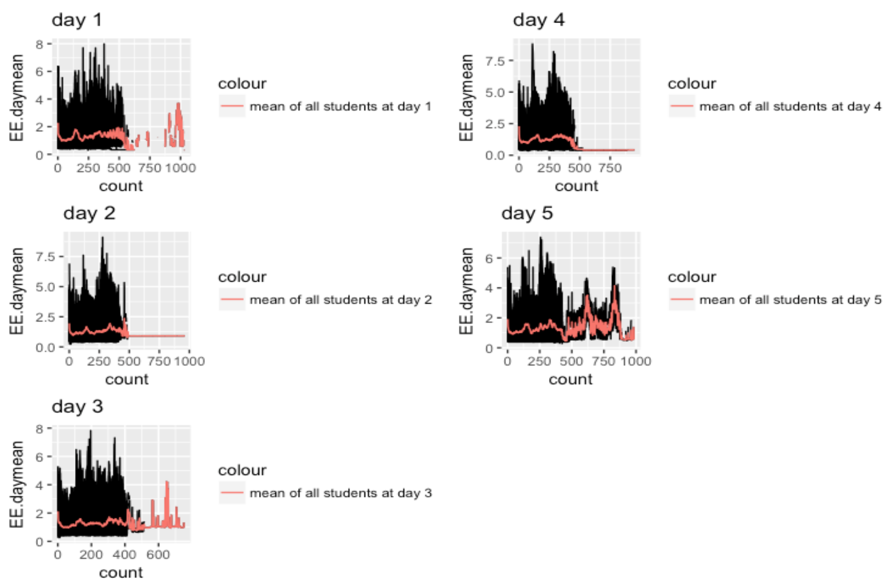


Figure 1.

Variable	Mean(s.d.)/ N(%)
BMI at baseline (kg/m^2)	17.18(2.79)
BMI in Spring Year 2 (kg/m^2)	17.55(3.23)
Average SDEE (cal/min)	2.48(1.34)
Age (years)	7.75(0.73)
Whites	141(75.00 %)
Hispanics	15(7.98 %)
Blacks	14(7.45 %)
Other	18(9.57 %)
Boys	94(50.00 %)
Girls	94(50.00 %)

Table 1. Descriptive data

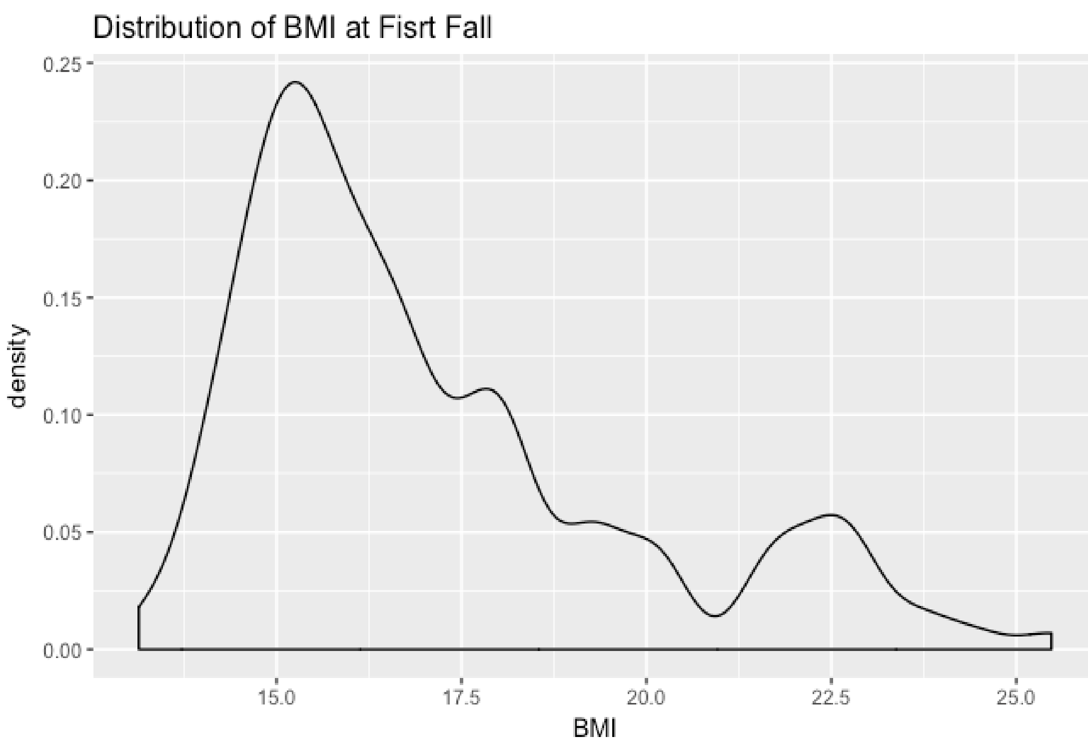


Figure 2. Distribution of BMI at the first semester.

3. Model Considered

3.1 Linear Mixed Effects Model

Since the both step counts and SDEE data were collected for one subject at multiple time points, we could consider the study as a longitudinal study which contain repeated measure data. Linear mixed effects models are often used for such studies. Mixed effects models are particularly applied when repeated measurements are made on the same statistical units such as longitudinal data. This model is written as:

$$Y_{ij} = X_{ij}\beta + Z_{ij}b_i + \epsilon_{ij}$$

where Y_{ij} is the jth response for the ith subject, β is a $p \times 1$ vector of fixed coeffeicients and X_{ij} is the $1 \times p$

vector of fixed variables, b_i is a $q \times 1$ for the random effects while Z_{ij} is a $1 \times q$ vector for the random variables included in the model. ϵ_{ij} is the random error term representing the random variation associated with the Y_{ij}^{th} response.

We assume that ϵ_{ij} and b_i have a normal distribution and the explanatory variables are related linearly to the response. However, smoothed scatter plot of the relationship between the mean SDEE and BMI difference between first and second semester (Figure 3) indicates the different. There is clearly not a linear relationship between SDEE and BMI difference. What's more, the gender-adjusted relationships between them are linear either (Figure 4).

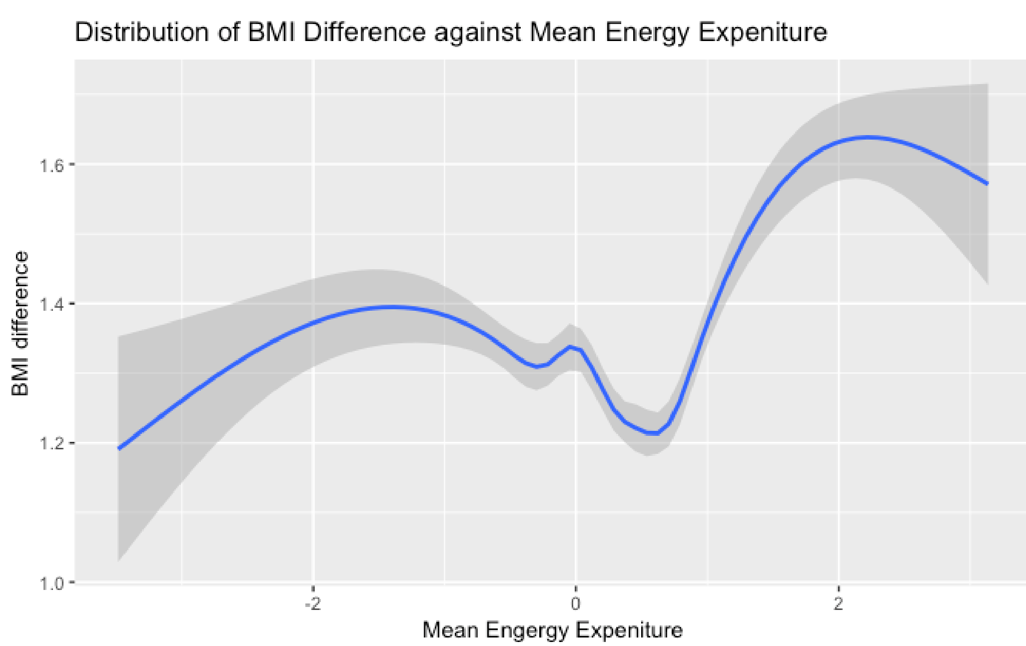


Figure 3. Plots of BMI difference between the first semester and second semester against Mean SDEE

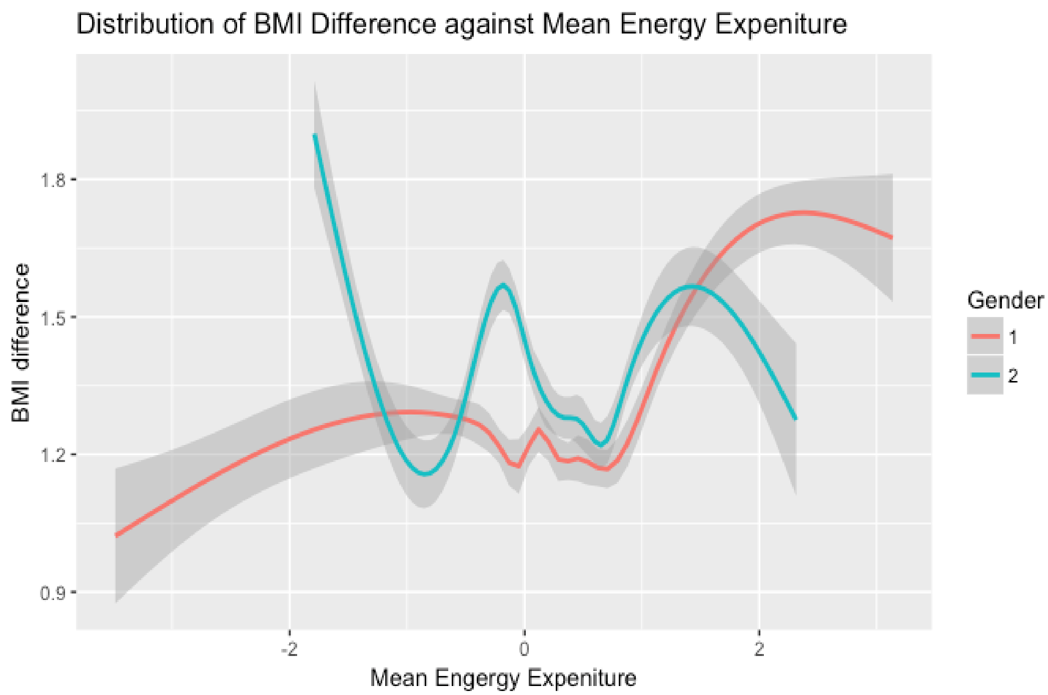


Figure 3. Plots of BMI difference between the first semester and second semester against Mean SDEE stratified by gender

3.2 Quantile Linear Mixed Effects Model

3.2.1 Quantile Regression

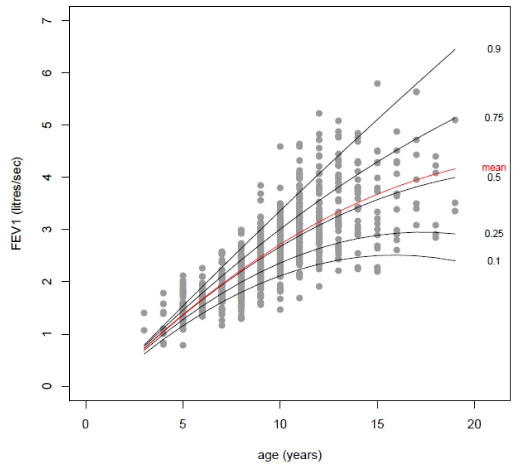
Overweight and obesity in children are defined based on age- and sex- adjusted body mass indexes (BMI) in the upper percentile ranges. There are studies on BMI using traditional linear regression model, which are designed to assess the impact of BMI within the "normal" percentile range and limit the analysis of the impact of BMI on people at higher risks for overweight and obesity. Therefore, statistical method that would allow assessing the effect of BMI across the full distribution of it is more preferred and desirable here. Quantile regression is a statistical approach that can be used to evaluate the effects of response variables on the outcome across the full distribution of it (Koenker and Bassett, 1978). Therefore, in this study, we choose to use quantile regression mixed model to analyze our data set for the relationship between BMI change and mean energy expenditure of students.

Quantile regression is the basis in the study. Generally, quantile regression is a type of regression analysis used in statistics and econometrics. We know that method of least squares estimates the conditional mean of the response variable with certain values of the predictor variables. However, quantile regression aims at estimating either the conditional median or other quantiles of the response variable. So, in some literature, quantile regression is described as an extension of linear regression but we can use it for some conditions that linear regression is not applicable.

Furthermore, quantile regression pertains to the unknown quantiles of an outcome as a function of a set of covariates and a vector of fixed regression coefficients.

After introducing what is quantile regression, we want to explain why we want to use quantile regression to analyze out dataset. Firstly, compared to ordinary mean regression problem, which uses least squares estimation), one of the biggest advantages of quantile regression is the robustness against outliers in the response measurements. We know that when there are severe outliers in the data, the mean value will be significantly affected. Thus, mean regression will be inappropriate for the data set with severe outliers. However, quantile regression, with analyzing different quantiles at the same time, can let us track the change of 10% or 90% quantiles and eliminates the effect of outliers. Furthermore, quantile regression is useful to discover more useful predictive relationships between variables in cases where there is no relationship or only a weak relationship between the means of such variables.

In order to explain the merits of quantile regression mentioned above, we show an example in the figure below. This is an example of 654 observations of FEV1 in individuals aged 3 to 19 years who were seen in the Childhood Respiratory Disease (CRD) study in Ease Boston, Massachusetts. FEV1, which is also known as Tiffeneau-Pinelli index, is a calculated ratio used in the diagnosis of obstructive and restrictive lung disease. In the figure below, we can see that the mean value increases when the individuals are getting older but the slope is smaller at older age. Thus, the mean value analysis gives us limited result regarding the relationship between FEV1 and age. However, if we take a look at the quantile regression results, we can see that the high quantiles (75% and 90%) increase linearly with age, especially 90% quantile. The low quantiles (10% and 25%) increase non-linearly with age. Also, if we take a look at the curves of low quantiles, the 10% quantile curve even drops with increasing age. So after performing quantile analysis of this small dataset, we can see draw a conclusion that if a child is ranked in lower range of FEV1, he or she will have a higher possibility of limited increase in FEV1 during grow up. On the other hand, if a child is ranked in higher ranger of FEV1, he or she will have a high possibility of fast increasing of FEV1 while growing up. Thus, the children who have low FEV1 may have a larger possibility of having respiratory or lung disease in the future. In a nutshell, quantile regression provides analysts more comprehensive understanding on the data set, and thus dig more useful results.



Regression quantiles (black) and mean fit (red) of FEV1 vs Age.

Figure 4 FEV1 versus age example of quantile regression.

3.2.2 Methodology of Quantile Regression

After introducing the general information and merits of quantile regression, we now need to introduce the methodology of quantile regression. As introduced above, quantile regression shares some similarities with linear regression. So the methodology or equation derivation will be similar of these two regression analysis. Firstly, we can consider a sample of observations $\left(\mathbf{x}_i^T, y_i\right)$ with $i = 1, \ldots, M$ drawn independently from a population with continuous distribution function $F_{y_i|\mathbf{x}_i}$. We assume the distribution function $F_{y_i|\mathbf{x}_i}$ to be unknown, and the corresponding quantile function is then given by its inverse as

$$Q_{y_i|\mathbf{x}_i} \equiv F_{y_i|\mathbf{x}_i}^{-1} \quad \ldots\ldots(1)$$

Similar as mean regression problem, for linear quantile regression problems, our goal is to estimate models of the following form of

$$Q_{y_i|\mathbf{x}_i}(p) = \mathbf{x}_i^T \beta(p) \quad \ldots\ldots(2)$$

In Eq. 2 above, p denotes the quantile level of interest and $0 < p < 1$. In some literature and the handbook of "lqmm" package, the quantile level of interest is denoted as τ. Also, in another word, we can index the quantiles Q of the continuous response y_i with p, that is

$$\Pr\left(y_i \le Q_{y_i}(p)\right) = p \quad \ldots\ldots(3)$$

In this case, the linear or conditional quantile function $Q_{y_i}\left(p \mid x_i\right) = x_i^T \beta(p)$ with $i = 1, \ldots, M$ can be estimated by solving the following minimum function

$$\min_{\beta} \sum_{i=1}^{M} g_p\left(y_i - \mathbf{x}_i^T \beta(p)\right) \qquad (4)$$

In Eq. 4 above, $\beta(p)$ is the regression coefficient vector indexed by p. Also, we may notice that the biggest difference between QR and mean regression appears. Compared to mean regression, which follows symmetrical normal distribution, QR follows asymmetric Laplace distribution. Thus, in Eq. 4, we have

$$g_p(v) = p \cdot \max(v, 0) + (1 - p) \cdot \max(-v, 0) \qquad (5)$$

Eq. 5 is asymmetrically weighted L1 loss function. In some of the literature, Eq. 5 can also be expressed as

$$g_p(v) = v\left[p - l_1(z < 0)\right] \qquad (6)$$

In Eq. 6, $l_1(z < 0)$ represents the gold standard loss. When the condition of this function is true, it equals one. Otherwise, $l_1(z < 0)$ equals zero. So, in our problem, when z is negative, $l_1(z < 0)$ equals one.

To show the difference between mean regression, which is least square estimation, and quantile regression, which is least absolute deviations estimation, we made the following figure. The difference between the minimum estimation of least squares and least absolute deviations, is the loss function g_p as mentioned above. Moreover, mean regression problem follows normal distribution so it has scale parameter, which defines the concentration level of the distribution, and location parameter, which defines the location of the mean value or the peak. The quantile regression problem, on the other hand, follows asymmetric Laplace distribution, which sometimes abbreviated as AL distribution. It also has scale parameter and location parameter but AL distribution has another unique skewness parameter. The skewness parameter controls whether the distribution skews to the left or right, and the severity of the skewed tail. Thus, AL distribution can be denoted as $AL(\mu, \sigma, p)$.

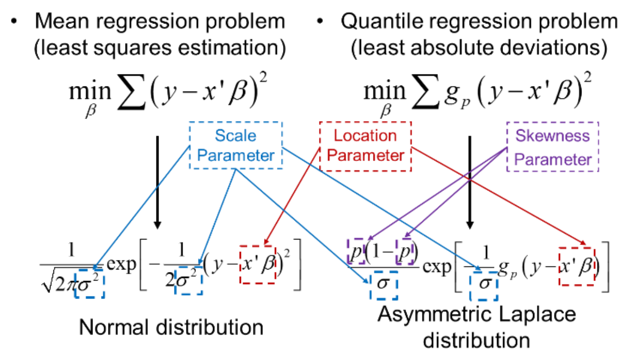


Figure 5 Least squares estimation versus least absolute deviations estimation.

3.2.3 Linear Quantile Mixed Model (LQMM)

We have introduced what is quantile regression and the basis methodology of it. Now, we need to research on how to apply quantile regression to our data set. Mixed effects models represent a common and well-known class of regression models used to analyze data coming from similar design. For clustered data, we need to use linear mixed models. We can consider a clustered data in the form of $\left(\mathbf{x}_{ij}{}^T, \mathbf{z}_{ij}{}^T, y_{ij}\right)$ for $j = 1, ..., n_i$, $i = 1, ..., M$, and $N = \sum_i n_i$. In the clustered data,

- $\mathbf{x}_{ij}{}^T$ is the j-th row of a known $n_i \times p$ matrix $\mathbf{X}_i$
- $\mathbf{z}_{ij}{}^T$ is the j-th row of a known $n_i \times q$ matrix $\mathbf{Z}_i$
- y_{ij} is the i-th observation of the i-th response vector $\mathbf{y}_i = \left(y_{i1}, ..., y_{in_i}\right)^T$

Then, a typical formulation of a linear mixed model (LMM) for clustered data can be expressed as

$$y_{ij} = \mathbf{x}_{ij}{}^T \beta + \mathbf{z}_{ij}{}^T u_i + \varepsilon_{ij} \quad \text{.........(7)}$$

Eq. 7 can be also written in the matrix form as

$$y = \mathbf{X}\beta + \mathbf{Z}u + \varepsilon \quad \text{.........(8)}$$

In Eq. 7 and 8, β is the fixed effect associated with p model covariates, and u is the random effects associated with q model covariates.

The linear quantile mixed model is similar as the linear mixed model, but the results to obtain are different quantiles. For LQMM, we need to introduce a joint AL model for $\mathbf{y}_i \mid \mathbf{u}_i$ with location

$$\boldsymbol{\mu}_i{}^{(\tau)} = \mathbf{X}_i \boldsymbol{\theta}_x{}^{(\tau)} + \mathbf{Z}_i \mathbf{u}_i \quad \text{.........(9)}$$

Also, the scale parameter is $\sigma^{(\tau)}$, respectively. In Eqn. 9, $\boldsymbol{\theta}_x{}^{(\tau)} \in \mathbb{R}^p$ is a vector of unknown fixed effects. We need to point out that in Eq. 9, the quantile is denoted as τ -th instead of p-th above. The reason is that while we applying LQMM model to our data set, we used lqmm package. In the lqmm package, the author kept using the notation as τ -th. The τ -th LQMM is given by

$$\mathbf{y} = \boldsymbol{\mu}^{(\tau)} + \boldsymbol{\varepsilon}^{(\tau)} \quad \text{...} (10)$$

In Eq. 10, $\boldsymbol{\mu}^{(\tau)} = \left(\boldsymbol{\mu}_1{}^{(\tau)}, ..., \boldsymbol{\mu}_M{}^{(\tau)}\right)^T$ and it can be compactly written in matrix form as

$$\boldsymbol{\mu}^{(\tau)} = \mathbf{X}\boldsymbol{\theta}_x{}^{(\tau)} + \mathbf{Z}\mathbf{u} \quad \text{...} (11)$$

Thus, the joint density of (y, u) based on M clusters in the τ -th quantile is expressed as

$$p\left(\mathbf{y}, \mathbf{u} \mid \boldsymbol{\theta}_x{}^{(\tau)}, \sigma^{(\tau)}, \boldsymbol{\Psi}^{(\tau)}\right) = \prod_{i=1}^{M} p\left(\mathbf{y}_i \mid \boldsymbol{\theta}_x{}^{(\tau)}, \sigma^{(\tau)}, \mathbf{u}_i\right) p\left(\mathbf{u}_i \mid \boldsymbol{\Psi}^{(\tau)}\right) \quad \text{.............................} (12)$$

After obtaining Eq. 12, the next step will be adopting an estimation strategy for the parameters of interest, which prompts considerations on how to deal with the unobserved random effects u. There are many trials on mixed model estimation. For LQMM, Monte Carlo EM procedure is implemented by Geraci and Bottai (2007), which is time consuming. In the **lqmm** R package, Gaussian quadrature approach has been implemented by Geraci and Bottai (2014).

4. R Package: lqmm

In this part, we want to briefly introduce the package **lqmm**. This package is a suite of commands for fitting linear quantile mixed models of the form shown in Eq. 8. The model is two –level nested like households within same postcode or repeated measurements on same subject. The error term follows the asymmetric Laplace distribution as $\varepsilon \sim AL(0, \sigma I, p)$. As mentioned above, in order to fit quantile regression models with random intercepts, Gaussian quadrature is implemented in this package.

```r
fit.lqmm <- lqmm(fixed = delta_yr1 ~ mean_ee * Gender,
                 random = ~ 1, group = Subject,
                 data = data_clean, tau = c(0.25,0.5,0.75),
                 nK =7, type = "robust",
                 control=lqmmControl(method = "gs", LP_tol_ll = 1e-3, LP_max_iter = 2000))

summary(fit.lqmm)
```

```
tau = 0.25

Fixed effects:
                   Value Std. Error lower bound upper bound Pr(>|t|)
(Intercept)     -0.18684    0.23384    -0.65676      0.2831   0.4281
mean_ee          0.09027    0.13956    -0.19018      0.3707   0.5208
Gender2          0.36184    0.27752    -0.19586      0.9195   0.1984
mean_ee:Gender2 -0.11339    0.13910    -0.39292      0.1661   0.4189

tau = 0.5

Fixed effects:
                    Value Std. Error lower bound upper bound Pr(>|t|)
(Intercept)     -0.002989   0.218154   -0.441386      0.4354   0.9891
mean_ee          0.129981   0.140739   -0.152844      0.4128   0.3602
Gender2          0.201494   0.256274   -0.313508      0.7165   0.4355
mean_ee:Gender2 -0.146919   0.143594   -0.435481      0.1416   0.3113

tau = 0.75

Fixed effects:
                     Value Std. Error lower bound upper bound Pr(>|t|)
(Intercept)     -0.0417498  0.2139864  -0.4717715      0.3883  0.84612
mean_ee          0.2592706  0.1324772  -0.0069523      0.5255  0.05605 .
Gender2          0.7389187  0.3001798   0.1356848      1.3422  0.01740 *
mean_ee:Gender2 -0.2894821  0.1536918  -0.5983374      0.0194  0.06557 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC:
[1]  -572.5 (df = 6) -1346.8 (df = 6)  -295.0 (df = 6)
```

Figure 6 An example of usage of lqmm package.

In Fig. 6, we show an example of applying lqmm package to a dataset. The main call lqmm is the core of this package. The first element in the main call is fixed is an object of class formula for fixed effects and a symbolic description of the model to be fitted. The random part is a one sided formula of the form of : $x_1 + x_2 + ... + x_n$ for random effects and a symbolic description of the model to be fitted. The third part is the grouping factor, and the fourth part data is the place to input the data set, which should be pre-formatted to a table or a matrix. The next part is tau, which is obvious that we can type in the different quantile value of interest here. The nk controls the quadrature knots, and we need to try different nk values for best results. There are two types of quadrature process. In Fig.3, we used type = “robust”, which means Gauss-Laguerre process, and the other possible type is “normal”, which uses Gauss-Hermite process. You can also control the numerical part of the lqmm process using “lqmmcontrol” command. In this command, two optimization methods can be chosen. One is gradient search called as “gs” and the other one is derivation free called as “df”. Also, the loop tolerance and max iteration can also be pre-set.

5. Results and Discussion

This study demonstrates a case when a quantile regression is advantageous over traditional or classical regression models. LQMM is effective in examining the effect of baseline SDEE on upper percentiles of BMI after 12 months. Quantile regression methods provide a more robust approach to analyzing the BMI data compared to linear regression methods.

In Table 2 below, we show one of the best result we obtained using robust mode and gradient search optimization method. The asterisk mark on the value means from the result, its is significant to the quantile regression result. In another word, for 75th quantile regression, we found strongest linear relationship between BMI difference and energy expenditure. In Fig.7, we plot points of data set with both genders, the three quantile regression result lines, and linear mean regression result. It is worth to mention that from our model, we considered both mean energy expenditure as well as the gender but in order to make the 2D graph looks straight forward, we included both genders. From Fig.7, we can see that it is hard to interpret the relationship between BMI change and mean energy expenditure. The linear mean regression (red line) does not provide useful result because of the outliers. However, we can see that for 75th quantile, higher energy expenditure results in larger positive BMI change. We need to mention that the positive BMI difference in this plot means the BMI of this individual changes to a normal and healthy range. Vice versa, the negative values means the BMI of this individual changes to a worth range. Thus, we can see that the individuals who spend more energy in a day will have a higher possibility to move toward good BMI range with larger BMI change. If we only perform linear mean regression, the information we can intercept from the data set will be limited, we might draw the conclusion that mean energy expenditure is irrelevant to BMI change.

	Intercept	Mean EE	Gender	Interaction of EE and Gender
25th quantile	-0.18684	0.09027	0.36184	-0.11339
50th quantile	-0.002989	0.129981	0.201494	-0.146919
75th quantile	-0.0417498*	0.2592706*	0.7389187*	-0.2894821*

Table 2 One of the best results from robust type

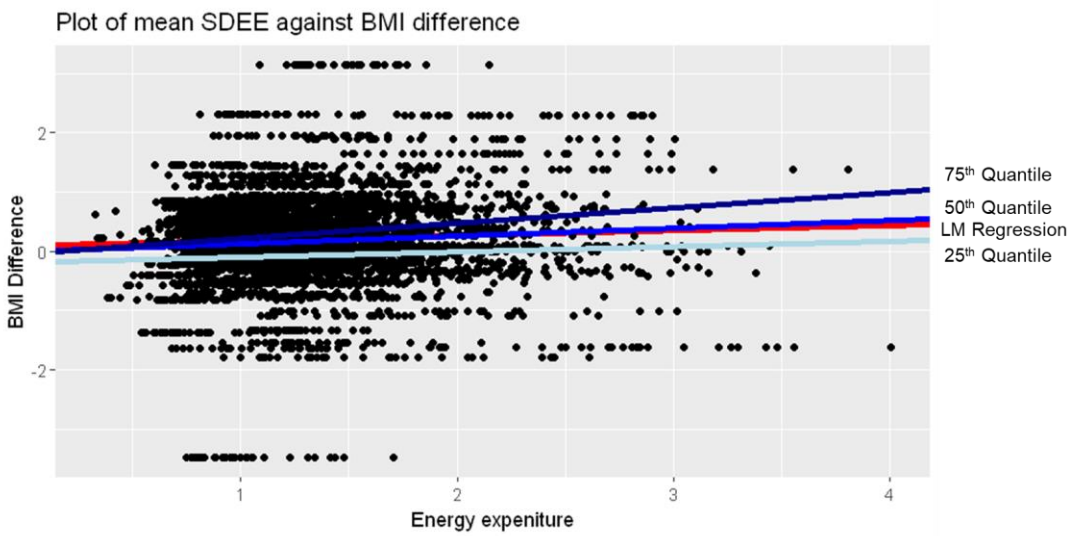


Figure 7 Plot of mean energy expenditure against BMI difference

Another point we want to discussion is the different computational method in the lqmm package. Our final result is based on the “robust” mode, which uses Gauss-Laguerre method. Compared to “normal” mode, which uses Gauss-Hermite method, robust mode yields more stable results. We met less convergence warning with robust mode without costing extremely more computational cost. Also, in the optimization control process, we tried both derivation free and gradient search method. We found that derivation free is fast and stable, but the result is meaningless. However, the gradient search method costs more computational cost but the result is accurate. In some of our cases, there are waning messages like the one below.

```
"Lower loop did not converge in: lqmm. Try increasing max number of iteration
s (1000) or tolerance (0.001)
"
```

This warning message happens because when the algorithm requires a certain number of data points to estimate the specified regression model but one or more bootstrap samples do not provide adequate information because of a particular configuration of their units (Geraci 2014). We noticed that the robust mode will yield considerably less warning messages compared to normal mode. Increasing iteration numbers of lower loop may reduce these warnings but it does not help to improve the final results.

6. Limitation

Despite its strengths of the statistical technology used, there are several limitations to this study that might affect the results. First of all, A larger sample size would have improved our results of the study, especially for the analysis of effects of predictors on the extreme quantiles of BMI. Besides that, there might be an issue of measurement error, which is the difference between a measured value of a quantity and its true value, as the both of the step counts and SDEE were measured by the armband device. Even though such error is not exactly a “mistake”, we would still like it to be as small as possible in order to minimize the variability of the data. Another limitation to this study is the problem of possible cofounders including diet and socioeconomic status, which were not taken into account in our analysis.

Reference

Geraci M (2014). lqmm: Linear Quantile Mixed Models. R package version 1.5.1, URL http://CRAN.R-project.org/package=lqmm.

Geraci M, Bottai M (2007).”Quantile Regression for Longitudinal Data Using the Asymmetric Laplace Distribution." *Biostatistics*, **8**(1), 140-154.

Geraci M, Bottai M (2014). “Linear Quantile Mixed Models." *Statistics and Computing*, **24**(3), 461-479.

Geraci, M. (2014). Linear Quantile Mixed Models: The lqmm Package for Laplace Quantile Regression. *Journal of Statistical Software*, **57**(13), 1 - 29.

Liu, L. L., Lawrence, J. M., Davis, C., Liese, A. D., Pettitt, D. J., Pihoker, C., Dabelea, D., Hamman, R., Waitzfelder, B., and Kahn, H. S. (2010). Prevalence of overweight and obesity in youth with diabetes in usa: the search for diabetes in youth study. Pediatric Diabetes, 11(1):4–11.

Bandini, L. G., Must, A., Phillips, S. M., Naumova, E. N., and Dietz, W. H. (2004). Re- lation of body mass index and body fatness to energy expenditure: longitudinal changes from preadolescence through adolescence. The American Journal of Clinical Nutrition, 80(5):1262–1269.

Wechsler, H., Devereaux, R. S., Davis, M., and Collins, J. (2000). Using the school environ- ment to promote physical activity and healthy eating. Preventive Medicine, 31(2):S121–S137.

Benden, M. E., Zhao, H., Jeffrey, C. E., Wendel, M. L., and Blake, J. J. (2014). The evaluation of the impact of a stand-biased desk on energy expenditure and physical activity for elementary school students. International Journal of Environmental Research and Public Health, 11(9):9361–9375.

Lanningham-Foster, L., Foster, R. C., McCrady, S. K., Manohar, C. U., Jensen, T. B., Mitre, N. G., Hill, J. O., and Levine, J. A. (2008). Changing the school environment to increase physical activity in children. Obesity, 16(8):1849–1853.