Hindawi
Journal of Probability and Statistics
Volume 2019, Article ID 9750538, 5 pages
https://doi.org/10.1155/2019/9750538

Review Article

A Comparison of Mean-Based and Quantile Regression Methods for Analyzing Self-Report Dietary Intake Data

Michelle L. Vidoni,[1] Belinda M. Reininger,[2] and MinJae Lee [1,3]

[1]*Biostatistics, Epidemiology, and Research Design (BERD) Core, Center for Clinical and Translational Sciences (CCTS), The University of Texas Health Science Center at Houston, 6410 Fannin, Houston, TX 77030, USA*
[2]*Health Promotion & Behavioral Sciences, Hispanic Health Research Center, The University of Texas School of Public Health Brownsville Regional Campus, One West University Blvd., Brownsville, TX 78520, USA*
[3]*Division of Clinical and Translational Sciences, Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, 6410 Fannin, Houston, TX 77030, USA*

Correspondence should be addressed to MinJae Lee; minjae.lee@uth.tmc.edu

Received 17 July 2018; Accepted 12 February 2019; Published 3 March 2019

Guest Editor: Min Zhang

In mean-based approaches to dietary data analysis, it is possible for potentially important associations at the tails of the intake distribution, where inadequacy or excess is greatest, to be obscured due to unobserved heterogeneity. Participants in the upper or lower tails of dietary intake data will potentially have the greatest change in their behavior when presented with a health behavior intervention; thus, alternative statistical methods to modeling these relationships are needed to fully describe the impact of the intervention. Using data from *Tu Salud ¡Si Cuenta! (Your Health Matters!) at Home Intervention*, we aimed to compare traditional mean-based regression to quantile regression for describing the impact of a health behavior intervention on healthy and unhealthy eating indices. The mean-based regression model identified no differences in dietary intake between intervention and standard care groups. In contrast, the quantile regression indicated a nonconstant relationship between the unhealthy eating index and study groups at the upper tail of the unhealthy eating index distribution. The traditional mean-based linear regression was unable to fully describe the intervention effect on healthy and unhealthy eating, resulting in a limited understanding of the association.

1. Introduction

Many health behavior interventions focus on positive lifestyle changes in the areas of increasing physical activity and healthy diets. Adopting these behavior changes can prevent or reduce the negative health consequences of obesity in minority US populations. Mexican Americans are particularly prone to physical inactivity and poor diets because of lack of fruit and vegetable consumption compared to Non-Hispanic Whites [1, 2]. Despite research showing poorer dietary intake than other ethnic groups, within the Mexican American population there is heterogeneity in healthy and unhealthy food intake [3].

Dietary intake data is typically measured using self-report tools and individual food intake is aggregated into compositional data or patterns to describe overall diets. When the dietary data are analyzed using mean-based approaches, such as ordinary least squares (OLS) regression, potentially important relationships with disease risk at the lower and upper levels of the distribution could be obscured due to unobserved heterogeneity. Participants in the upper or lower tails of dietary intake data, where inadequacy or excess is greatest, will theoretically have the greatest change in their behavior when presented with a health behavior intervention; thus, alternative statistical methods to modeling these relationships are needed to fully describe the impact of the intervention. This is particularly notable in certain populations, such as Mexican Americans, where variation in factors such as acculturation and language influence food choices and adherence to traditional and western diet patterns [3–6].

As an alternative to mean-based regression, quantile regression (QR) was developed by Koenker and Bassett and has primarily been used in the fields of risk management and business [7]. Quantile regression has been extended for handling longitudinal data based on different approaches that account for serial correlations within a subject and has been used as an important alternative to mean-based regression approaches because of its flexibility for modeling nonnormal data, or heterogeneous conditional distributions [8]. QR can model the conditional distribution of the response, not only on the conditional mean, giving the research critical insights when valuable information lies in the tails. Despite QR being computationally intensive and not equipped to handle small data sets, it is more robust to outliers than mean-based regression, where estimates of the conditional mean can be strongly influenced by outliers.

Application of QR to health and behavioral sciences is increasing and could be a valuable statistical tool for health researchers. QR has been used to evaluate the effects of physical activity or dietary intake on varying quantile levels of certain variables, such as BMI [9–12], waist circumference [13], socioeconomic status [14], and risk factors of disease outcomes including health-related scores and biomarker data [15–19]. A limited number of studies have introduced a QR-based approach specifically applied to behavioral data [20–22]. Yet, there is limited research focusing on how to use and apply QR results to improve behavioral interventions and maintenance of behavior change over time by possibly addressing the upper and lower tails of the population distribution differently.

The goal of this review was to compare traditional mean-based linear regression with QR through the illustration of their applications to real data from the behavioral intervention study aimed at improving healthy eating and to demonstrate the usefulness of QR in fully describing the relationships.

2. Linear Quantile Mixed Effect Regression

Let y_{it} be the measurement for the i-th subject ($i = 1, \ldots, n$) at time t ($t = 1, \ldots, n_i$), then we define a linear mixed effect regression model as

$$y_{it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \boldsymbol{z}_{it}^T \boldsymbol{\gamma}_i + \varepsilon_{it}, \tag{1}$$

where $\boldsymbol{x}_{it}$ is a vector of p covariates at t, $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression parameters, and the correlation among the observations within the i-th subject is induced by the subject-level residuals, i.e., $q \times 1$ vector $\boldsymbol{\gamma}_i$ and an associated vector $\boldsymbol{z}_{it}$ for q random effect variables. The error term can be defined as $\boldsymbol{z}_{it}^T \boldsymbol{\gamma}_i + \varepsilon_{it}$, where random errors for individual records, ε_{it}, are independent of each other. We assume that linear quantile mixed models are determined based on the asymmetric Laplace distribution (ALD) [23], which has a good performance on data generated from many error distributions, and a relationship with the L_1-norm objective function [7]. Let a response variable y be an ALD, denoted ALD(μ, σ, τ), then we can define a probability density function,

$$f(y \mid \mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\}, \tag{2}$$

where $0 < \tau < 1$ is the skewness parameter, μ is the location parameter, σ is the scale parameter, and a loss function $\rho_\tau(v) = (\tau - I(v \le 0))$ represents the contribution by residuals v. Assuming the location parameter is $\mu_{\tau,it} = \boldsymbol{x}_{it}^T \boldsymbol{\beta}_\tau + \boldsymbol{z}_{it}^T \boldsymbol{\gamma}_i$, a quantile regression model related to the τ-th quantile of a response variable y_{it}, conditional on $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$, has the form:

$$q_\tau(y_{it}) = \boldsymbol{x}_{it}^T \boldsymbol{\beta}_\tau + \boldsymbol{z}_{it}^T \boldsymbol{\gamma}_i + \varepsilon_{\tau,it}, \quad 0 < \tau < 1, \tag{3}$$

where $\boldsymbol{\beta}_\tau$ is a vector of quantile-specific regression parameters corresponding to the coefficient $\boldsymbol{\beta}$ in a linear regression model (1) and $\varepsilon_{\tau,it} \sim$ ALD$(0, \sigma, \tau)$, which are also dependent on τ. The objective function for y_{it} for fixed τ is expressed as

$$Q_n(\boldsymbol{\beta}_\tau) = \sum_{i=1}^{n} \sum_{t=1}^{n_i} \rho_\tau\left(y_{it} - \boldsymbol{x}_{it}^T \boldsymbol{\beta} - \boldsymbol{z}_{it}^T \boldsymbol{\gamma}_i\right). \tag{4}$$

We can estimate quantile-specific regression parameters that minimize the objective function above. As we assume $y_{it} \sim$ ALD(μ_{it}, σ, τ), ALD is determined as a scale mixture of normal distribution based on Laplace distribution with the skewness parameter τ that is treated here as a quantile level. Then a likelihood for y_{it} at τ-th quantile can be expressed as

$$L(\boldsymbol{\beta}, \sigma \mid y_{it}, \tau) = \frac{\tau(1-\tau)}{\sigma^N} \exp\left\{-\sum_{i=1}^{n} \sum_{t=1}^{n_i} \rho_\tau\left(\frac{y_{it} - \boldsymbol{x}_{it}^T \boldsymbol{\beta} - \boldsymbol{z}_{it}^T \boldsymbol{\gamma}_i}{\sigma}\right)\right\}. \tag{5}$$

If σ is considered a nuisance parameter, then the maximization of this likelihood above is equivalent to the minimization of the objective function of quantile regression (4) defined above. More details regarding estimation process are available elsewhere [8].

3. Example

3.1. Tu Salud ¡Si Cuenta! (Your Health Matters!) at Home Intervention. The behavioral data used in the current study were from the *Tu Salud ¡Si Cuenta! (Your Health Matters!) at Home Intervention*. One of the main objectives of this randomized control trial was to increase participant intake of healthy foods and decrease unhealthy food intake through exposure to community health workers delivering a behavioral modification intervention. The study was conducted in the Texas Rio Grande Valley area and included participants who were Mexican American adults, aged 18-75 years, and enrolled in the ongoing Cameron County Hispanic Cohort [1, 24] . Participants were randomly selected and randomized into either the intervention or standard care group from June 2010 to April 2013. The intervention group received up to six monthly community health work home visits in the first 6 months of the intervention, which included lifestyle change

education, motivation, and support. No other intervention elements, other than that equivalent to the standard care group, were offered during the last 6 months of the trial. The standard care group participants were potentially exposed to a community-wide physical activity and healthy diet campaign across the 12 months. Data were collected at baseline, 6- and 12-month follow-ups.

Participants completed a dietary intake questionnaire that asked if yesterday they had eaten 20 commonly and culturally appropriate foods and how many times with the following responses available: no, once, twice, three times, four times, and five or more times [25, 26]. Responses were summed into Healthy and Unhealthy Eating Indices (HEI and UNHEI, respectively). The HEI score was comprised responses to the 10 healthy food items (baked or grilled fish, turkey or chicken; eggs; beans; fruit; fruit juice; orange vegetables; other vegetables; salad; whole grain breads; and whole grain cereals) with a possible response range from 0 to 50. The UNHEI was composed of the responses to the 9 unhealthy food items (baked goods; french fries or chips; fried meat; frozen desserts; red and processed meats; nonchocolate candy; regular sodas; sweetened or sports drinks; and white bread) with a possible range from 0 to 45 [27]. Both HEI and UNHEI scores appeared to be well-approximated by a normal distribution.

3.2. Quantile Regression and Mean-Based Regression. To assess intervention effect on healthy and unhealthy eating, a multivariable longitudinal QR and mean-based model were conducted based on the linear mixed effect model equation below.

$$\begin{aligned} y_{it} &= \alpha + \beta_1 x_{1i} \\ &\quad + \beta_{21} v_{1it} + \beta_{22} v_{2it} + \beta_{31} x_{1i} v_{1it} + \beta_{32} x_{1i} v_{2it} + \boldsymbol{u}_{it}^T \boldsymbol{\delta} + \gamma_i + \varepsilon_{it}, \end{aligned} \tag{6}$$

where the index score, y_{it}, can be either the HEI or UNHEI measurement for the i-th participant ($i = 1, \ldots, n$) at visit t ($t = 1, 2, 3$) and a binary variable x_1 for study group (x_1=1 if intervention) and v_{1i} and v_{2i} are dummy variables for two follow-up visits, i.e., month 6 (visit 2) and month 12 (visit 3), respectively. Interaction terms between study group and follow-up visits were included in the model to obtain estimates of the intervention effect at each time point. $\boldsymbol{u}_{it}$ is a vector of a set of potential confounders that were adjusted for in the model (i.e., gender, age, diabetes, marital status, years in school, employment status, type of insurance, generation, and preferred language) and $\boldsymbol{\delta}$ is an associated parameter vector. We also considered a random intercept by including an error term γ_i for the i-th subject. We used *lqmm* R package [8] for QR models and SAS *proc mixed* for mean-based models.

3.3. Results. There were 500 participants randomized to either the standard care or intervention groups, n=250 respectively. At baseline, the mean HEI score was 6.6 (standard deviation (SD)=3.3) for the standard care group and 6.9 (SD=3.5) for the intervention group. The mean UNHEI score for the standard care group was 5.4 (SD=3.4) and for the intervention group was 5.6 (SD=3.6).

Results from QR and mean-based regression are presented in Figure 1. The red line indicates estimated beta coefficients based on mean-based model for the effect of the study group at each time point, showing slight differences (i.e., beta coefficient <0.4) in mean HEI and mean UNHEI between intervention and standard care groups at baseline and follow-ups.

With regard to HEI, the results for QR and mean-based regression do not substantially differ. In contrast, the QR results indicate a nonconstant relationship between unhealthy eating and study groups at the upper tail of the distribution of the UNHEI. At baseline, the association between the distribution of UNHEI scores and study groups is not constant, as the intervention group is more likely to be in the upper tail of the UNHEI distribution at the start of the study. At month 6, the effect of the intervention is inconsistent across the UNHEI distribution. For example, at the upper tail of the UNHEI distribution the intervention group had higher UNHEI scores, yet around the quantile level τ=0.05 and 0.75 the intervention group reported lower UNHEI scores than the control group. The strength of the association in the upper tail of the distribution is attenuated at 6 months compared to baseline. More strikingly, at the 12-month follow-up QR suggests that there is an increase in unhealthy food intake in intervention group compared to control for the participants in the upper tail of the UNHEI data distribution.

4. Discussion

Mean-based regression results showed minimal differences in the healthy eating index at any visit between intervention and standard care groups, likewise for the unhealthy eating index. These results would lead a researcher to incorrectly assume that the intervention failed to increase intake of healthy foods or decrease unhealthy food intake or possibly conclude that the reasons for the lack of change might not be due to the intervention itself but to information bias or environmental changes in the community based intervention.

In contrast, the results of the QR highlight a different relationship between the study groups and outcomes. The estimated coefficients were not constant across the distribution of the UNHEI outcome at baseline and follow-ups. These results may indicate a baseline imbalance in the UNHEI outcome, which under mean-based regression would have not been identified, and approaches to adjust for the imbalance should be considered. Likewise at the 6-month follow-up, the protective effect of the intervention would have also been ignored using mean-based methods. The QR results for the unhealthy index at the 12-month follow-up identified an inconsistent relationship between study group and UNHEI. At the lower tail of UNHEI, the intervention was protective, then this relationship reversed at the upper tail of UNHEI. Overall, there was little difference in the UNHEI between intervention and standard care groups, except at the upper tail of the UNHEI distribution. This indicates purely mean-based approach may not be appropriate for evaluating the effect of the intervention on dietary uptake behaviors in populations with unobserved heterogeneity.

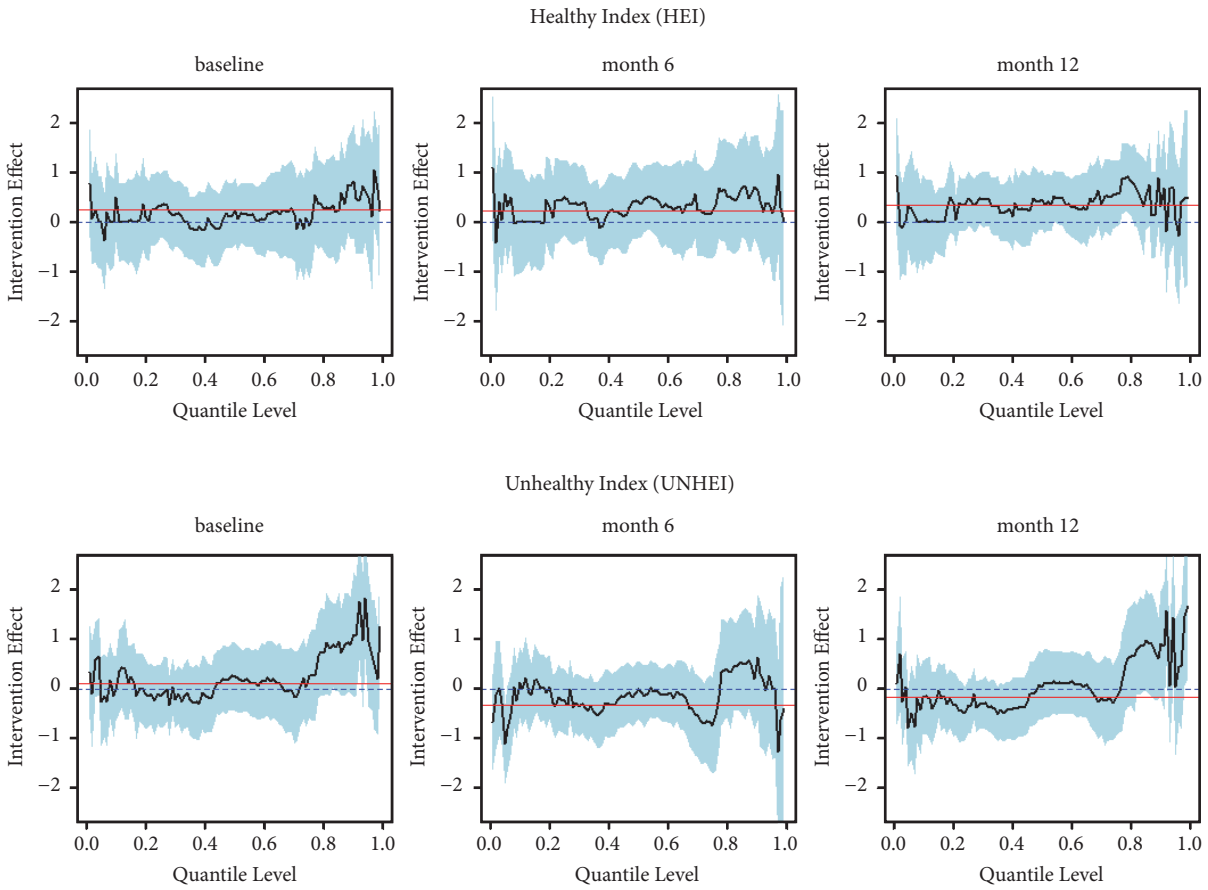

FIGURE 1: Estimated adjusted∗ parameter of x_1 for study group (intervention vs. standard care) at each visit based on mean-based regression (red line) and quantile regression (black line with 95% confidence limits) by quantile levels of healthy (HEI) and unhealthy index (UNHEI). ∗Adjusted for gender, age, diabetes, marital status, years in school, employment status, type of insurance, generation, and preferred language.

5. Conclusions

The traditional mean-based linear regression was unable to fully describe the relationship between healthy and unhealthy eating and the intervention, resulting in a limited understanding of the intervention effect. Use of quantile regression identified a different relationship by modeling the coefficients across the distribution of the outcome resulting in a more complete picture of the association. These findings from the quantile regression results could be applied towards developing more effective behavioral intervention trials in heterogeneous populations.

Disclosure

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH CTSA or NIMHD or UTCO.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to acknowledge and thank Dr. Belinda Reininger for her guidance and support on demonstrating the statistical approaches through the behavioral data observed from the intervention program, the *Tu Salud ¡Si Cuenta! (Your Health Matters!) at Home Intervention,* which was supported by the UT Health Clinical and Translational Science Award (UL1 TR000371), NIH/National Institute on Minority Health and Health Disparities (MD000170 P20), and the Texas Department of State Health Services funding for University of Texas Community Outreach (UTCO). The authors would like to recognize the support provided by the Biostatistics/Epidemiology/Research Design (BERD) component of the Center for Clinical and Translational Sciences (CCTS) at the UT Health Science Center at Houston, which is mainly funded by the NIH Centers for Translational Science Award (NIH CTSA), grant UL1 RR024148.

References

[1] B. M. Reininger, L. Mitchell-Bennett, M. Lee et al., "Tu Salud,¡ Si Cuenta!: exposure to a community-wide campaign and

its associations with physical activity and fruit and vegetable consumption among individuals of mexican descent," *Social Science and Medicine*, vol. 143, pp. 98–106, 2015.

[2] B. M. Reininger, J. Wang, S. P. Fisher-Hoch, A. Boutte, K. Vatcheva, and J. B. McCormick, "Non-communicable diseases and preventive health behaviors: a comparison of Hispanics nationally and those living along the US-Mexico border Health behavior, health promotion and society," *BMC Public Health*, vol. 15, no. 1, article 564, 2015.

[3] B. M. Reininger, M. Lee, R. Jennings, A. Evans, and M. Vidoni, "Healthy eating patterns associated with acculturation, sex and BMI among Mexican Americans," *Public Health Nutrition*, vol. 20, no. 7, pp. 1267–1278, 2017.

[4] K. J. Duffey, P. Gordon-Larsen, G. X. Ayala, and B. M. Popkin, "Birthplace is associated with more adverse dietary profiles for US-born than for foreign-born Latino adults," *Journal of Nutrition*, vol. 138, no. 12, pp. 2428–2435, 2008.

[5] I. B. Ahluwalia, E. S. Ford, M. Link, and J. C. Bolen, "Acculturation, weight, and weight-related behaviors among Mexican Americans in the United States," *Ethnicity and Disease*, vol. 17, no. 4, pp. 643–649, 2007.

[6] J. K. Montez and K. Eschbach, "Country of birth and language are uniquely associated with intakes of fat, fiber, and fruits and vegetables among Mexican-American women in the United States," *Journal of the Academy of Nutrition and Dietetics*, vol. 108, no. 3, pp. 473–480, 2008.

[7] R. Koenker and G. Bassett Jr., "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[8] M. Geraci, "Linear quantile mixed models: the lqmm package for laplace quantile regression," *Journal of Statistical Software*, vol. 57, no. 13, pp. 1–29, 2014.

[9] D. A. Amugsi, Z. T. Dimbuene, P. Bakibinga, E. W. Kimani-Murage, T. N. Haregu, and B. Mberu, "Dietary diversity, socioeconomic status and maternal body mass index (BMI): quantile regression analysis of nationally representative data from Ghana, Namibia and Sao Tome and Principe," *BMJ Open*, vol. 6, no. 9, Article ID e012615, 2016.

[10] M. Bottai, E. A. Frongillo, X. Sui et al., "Use of quantile regression to investigate the longitudinal association between physical activity and body mass index," *Obesity*, vol. 22, no. 5, pp. E149–E156, 2014.

[11] S. Azagba and M. F. Sharaf, "Fruit and vegetable consumption and body mass index: a quantile regression approach," *Journal of Primary Care & Community Health*, vol. 3, no. 3, pp. 210–220, 2012.

[12] J. A. Mitchell, R. R. Pate, V. España-Romero, J. R. O'Neill, M. Dowda, and P. R. Nader, "Moderate-to-vigorous physical activity is associated with decreases in body mass index from ages 9 to 15 years," *Obesity*, vol. 21, no. 3, pp. E280–E286, 2013.

[13] J. A. Mitchell, M. Dowda, R. R. Pate et al., "Physical activity and pediatric obesity: a quantile regression analysis," *Medicine and Science in Sports and Exercise*, vol. 49, no. 3, pp. 466–473, 2017.

[14] Ø. Seippel, "Physical exercise and social inequality in Norway – A comparison of OLS and quantile regression analysis," *European Journal for Sport and Society*, vol. 12, no. 4, pp. 355–376, 2016.

[15] A. D'Silva, P. A. Gardiner, T. Boyle, D. G. Bebb, S. T. Johnson, and J. K. Vallance, "Associations of objectively assessed physical activity and sedentary time with health-related quality of life among lung cancer survivors: A quantile regression approach," *Lung Cancer*, vol. 119, pp. 78–84, 2018.

[16] Z. Wang, P. Gordon-Larsen, A. M. Siega-Riz et al., "Sociodemographic disparity in the diet quality transition among Chinese adults from 1991 to 2011," *European Journal of Clinical Nutrition*, vol. 71, no. 4, pp. 486–493, 2017.

[17] L. Liu, "Using multivariate quantile regression analysis to explore cardiovascular risk differences in subjects with chronic kidney disease by race and ethnicity: findings from the US chronic renal insufficiency cohort study," *International Cardiovascular Forum Journal*, 2015.

[18] A. K. Monroe, T. T. Brown, C. Cox, S. M. Reynolds, D. J. Wiley, F. J. Palella et al., "Physical activity and its association with insulin resistance in multicenter AIDS cohort study men," *AIDS Research and Human Retroviruses*, vol. 31, no. 12, pp. 1250–1256, 2015.

[19] E. Verly, J. Steluti, R. M. Fisberg, and D. M. L. Marchioni, "A quantile regression approach can reveal the effect of fruit and vegetable consumption on plasma homocysteine levels," *PLoS ONE*, vol. 9, no. 11, Article ID e111619, 2014.

[20] J. N. Variyam, J. Blaylock, and D. Smallwood, "Characterizing the distribution of macronutrient intake among U.S. Adults: A quantile regression approach," *American Journal of Agricultural Economics*, vol. 84, no. 2, pp. 454–466, 2002.

[21] Y. Wei, Y. Ma, and R. J. Carroll, "Multiple imputation in quantile regression," *Biometrika*, vol. 99, no. 2, pp. 423–438, 2012.

[22] Y. Wei and R. J. Carroll, "Quantile regression with measurement error," *Journal of the American Statistical Association*, vol. 104, no. 487, pp. 1129–1143, 2009.

[23] D. V. Hinkley and N. S. Revankar, "Estimation of the Pareto law from underreported data: a further analysis," *Journal of Econometrics*, vol. 5, no. 1, pp. 1–11, 1977.

[24] S. P. Fisher-Hoch, A. R. Rentfro, J. J. Salinas et al., "Socioeconomic status and prevalence of obesity and diabetes in a Mexican American community, Cameron County, Texas, 2004-2007," *Preventing Chronic Disease*, vol. 7, no. 3, article A53, 2010.

[25] D. M. Hoelscher, R. S. Day, E. S. Lee et al., "Measuring the prevalence of overweight in Texas schoolchildren," *American Journal of Public Health*, vol. 94, no. 6, pp. 1002–1008, 2004.

[26] A. Pérez, D. M. Hoelscher, H. S. Brown, and S. H. Kelder, "Peer reviewed: differences in food consumption and meal patterns in texas school children by grade," *Preventing Chronic Disease*, vol. 4, no. 2, 2007.

[27] C. E. Velazquez, K. E. Pasch, N. Ranjit, G. Mirchandani, and D. M. Hoelscher, "Are adolescents' perceptions of dietary practices associated with their dietary behaviors?" *Journal of the Academy of Nutrition and Dietetics*, vol. 111, no. 11, pp. 1735–1740, 2011.