Journal of Biometrics & Biostatistics

ISSN: 2155-6180

Ajayi and Bello, J Biom Biostat 2018, 9:3
DOI: 10.4172/2155-6180.1000402

Research Article Open Access

Predictors of Body Mass Index among Pregnant Women in Nigeria: A Comparison of Ordinary Least Squares Regression and Quantile Regression Models Using Machine Learning Approach

David Taiwo Ajayi* and Segun Bello

Department of Epidemiology and Medical Statistics, Faculty of Public Health, College of Medicine, University of Ibadan, Nigeria

Abstract

Poor nutrition during pregnancy is a major public health problem. Maternal under nutrition is a significant risk factor for maternal morbidity, mortality, poor birth outcomes (e.g. low birth weight), and infant mortality. Maternal under nutrition is defined as having a body mass index (BMI) <18.5 kg/m^2. Previous studies on maternal BMI utilized classical statistical approach, whose criteria for model assessment are goodness-of-fit test and residual examination. The aim of this study was to identify predictors of BMI among pregnant women in Nigeria, and to compare the performance of ordinary least squares (OLS) regression and quantile regression using machine learning approach.

This study utilized data from the 2013 Nigeria Demographic and Health Survey. A total of 3,049 pregnant women were included in the study. Data were summarized using descriptive statistics. The assumption of normality of the outcome variable (BMI) was tested using one-sample Kolmogorov-Smirnov test. Bivariate associations of BMI with independent variables were assessed using robust (nonparametric) statistical techniques: Kendall's tau correlation for continuous predictors, Wilcoxon rank sum test for binary predictors and Kruskal-Wallis test for multinomial predictors. Predictors of maternal BMI were investigated using OLS and quantile regression analyses. Model assessment was made using 10-fold cross-validation. A two-tailed p-value <0.05 was considered statistically significant.

The respondents had a mean age of 28.22 ± 6.30 years, and a mean BMI of 23.81 ± 4.18 kg/m^2. Multivariate analyses identified respondent's age, duration of pregnancy, wealth class, and residence as predictors of maternal BMI. The cross-validated mean squared error for the OLS regression model was lower than that for the quantile regression model.

Respondent's age, duration of pregnancy, wealth class, and residence were significantly associated with maternal BMI. OLS regression model fit the data more than the quantile regression model.

Keywords: Body mass index; Pregnant women; Machine learning; Ordinary least squares regression; Quantile regression; Cross-validation

Introduction

Poor nutrition during pregnancy is a major public health problem. Maternal undernutrition is a significant risk factor for maternal morbidity, mortality, poor birth outcomes (e.g. low birth weight), and infant mortality [1]. Undernutrition is responsible for more than 3.5 million deaths of mothers and children under the age of 5 each year in developing countries [2]. The prevalence of maternal undernutrition in Nigeria is 6.7% [3].

Maternal nutrition refers to the nutritional needs of women during antenatal and postnatal periods and sometimes also to the period prior to conception (i.e. during adolescence). Maternal undernutrition (or chronic energy deficit) is defined as having a body mass index (BMI) <18.5 kg/m^2 [1]. Maternal nutrition plays a critical role in fetal growth and development. Intrauterine environment is a major determinant of fetal growth. Studies have shown that maternal undernutrition during pregnancy reduces placental and fetal growth, a condition known as intrauterine growth restriction (IUGR) [4,5]. Maternal undernutrition causes clinical complications in fetuses and infants. For example, about 50% of nonmalformed stillbirths in humans are attributed to IUGR [6]. Moreover, perinatal mortality rates are 5-30 times greater in infants who weigh <2.5 kg at birth than those who have average birth weights [6]. Infants with IUGR are more likely to develop neurological, respiratory, intestinal, and circulatory disorders than those without IUGR [5].

Predictors of maternal nutrition have been investigated. In a review of studies on maternal and child health indicators in the South Asian region, Bhutta et al. [7] concluded that maternal illiteracy, poverty, and lack of empowerment of women were factors associated with maternal undernutrition. Begum and Sen [8] reported that education, exposure to media, and domestic decision-making were associated with maternal nutritional status in Bangladesh. Senbanjo and colleagues [3] identified age at first birth, maternal education level, and number of births as determinants of maternal nutritional status in Nigeria. However, previous studies employed classical statistical methods, such as ordinary least squares (OLS) regression and logistic regression, where models are validated using goodness-of-fit tests and residual examination. Breiman et al. [9] demonstrated that predictive accuracy of a statistical technique on a test data set is the appropriate criterion for how good the model is, and this is the hallmark of machine learning approach.

Machine learning entails estimating the systematic relationship (f) between an outcome variable and input variable(s) using a subset of the data set (training set), and assessing the model performance

***Corresponding author:** David Taiwo Ajayi, Department of Epidemiology and Medical Statistics, Faculty of Public Health, College of Medicine, University of Ibadan, Nigeria, Tel: +2347036421475; E-mail: dtb.ajayi@gmail.com

Received June 08, 2018; **Accepted** June 25, 2018; **Published** June 30, 2018

Citation: Ajayi DT, Bello S (2018) Predictors of Body Mass Index among Pregnant Women in Nigeria: A Comparison of Ordinary Least Squares Regression and Quantile Regression Models Using Machine Learning Approach. J Biom Biostat 9: 402. doi: 10.4172/2155-6180.1000402

on the hold-out set (observations not used in fitting or training the statistical model) [10]. In the regression setting, the most commonly used measure of model performance or accuracy is the mean squared error (MSE), given by:

$$MSE = \frac{1}{n}\sum(y_i - \hat{f}(x_i))^2,$$

Where y_i represents the response variable for the *i*th observation, $\hat{f}$ is the estimate of *f*, and $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the *i*th observation [10]. Cross-validation (CV) is a widely used technique for model assessment [10]. *k*-fold CV involves randomly dividing the set of observations into *k* folds, of approximately equal size. The first fold is treated as a validation set (test set), and the method is fit on the remaining *k*–1 folds. The MSE is then computed on the observations in the held-out fold. This procedure is repeated *k* times; each time, a different group of observations is treated as a validation set. This process results in *k* estimates of the test error, and the *k*-fold CV estimate is computed by averaging these values [10].

Studies have shown that BMI distribution is skewed; thus, researchers employed quantile regression to investigate factors associated with BMI, in addition to OLS regression [11-13]. Quantile regression is a semi-parametric technique that models quantiles of the response variable conditional on the covariates [14]. Quantile regression allows a comprehensive evaluation of the associations between predictor(s) and the outcome at various quantiles (or percentiles). Unlike the OLS regression, quantile regression makes no assumptions about the distribution of the errors; thus, it is more robust to non-normal errors and outliers [15].

Studies comparing predictive performance of quantile regression models with different quantiles are lacking. Moreover, studies on comparison of OLS regression and quantile regression models using CV are not available in the literature. Therefore, the aim of this study was to identify predictors of BMI among pregnant women in Nigeria, and to compare the performance of OLS regression and quantile regression using machine learning approach.

Methods

This study utilized data from the 2013 Nigeria Demographic and Health Survey (NDHS), implemented by the National Population Commission (NPC) in conjunction with the ICF International. The survey, a population-based cross-sectional study, employed a stratified three-stage cluster sampling to select the respondents. Respondents were selected from 904 clusters, comprising 372 urban areas and 532 rural areas in 36 states and the Federal Capital Territory, Abuja, Nigeria. A total of 38,868 women aged 15-49 years and 17,317 men aged 15-49 years were interviewed. Detailed description of the sample design and implementation is available in the 2013 NDHS report [16].

The Individual Recode data set was used. A total of 3,049 pregnant women were included in the analysis, after listwise deletion of missing values. Data were summarized using descriptive statistics. The assumption of normality of the outcome variable (BMI) was tested using one-sample Kolmogorov-Smirnov test. Bivariate associations of BMI with independent variables were assessed using robust (nonparametric) statistical techniques: Kendall's tau correlation for continuous predictors, Wilcoxon rank sum test for binary predictors and Kruskal-Wallis test for multinomial predictors. Predictors having p-value <0.2 in bivariate analysis were included in multivariate model [17]. Multivariate outliers were identified using the robust Mahalanobis distance. Predictors of maternal BMI were investigated using OLS and quantile regression analyses. Multiple Quantile regression models were fitted using quantiles from 0.10 to 0.95 with increment of 0.05. Ten-fold CV was performed the select the best quantile regression model using MSE as a performance criterion. The performance of the OLS regression model and the selected quantile model was compared using 10-fold cross-validation. A two-tailed p-value <0.05 was considered statistically significant. All analyses were done in R, version 3.4.4 (R Core Team, Vienna, Austria).

Results

Sample characteristics

The mean age of the respondents was 28.22 ± 6.30 years (Table 1). Most (99.1%) of the respondents were married (Table 2). About one-

Variable	Mean	SD	Median	Range
Respondent's age (years)	28.22	6.30	28.00	34.00
Number of living children	2.99	1.96	3.00	12.00
Preceding birth interval (years)	2.29	1.33	2.00	20.00
BMI (Kg/m^2)	23.81	4.18	23.05	36.46

Table 1: Descriptive statistics for continuous variables (N=3049).

Characteristic	Frequency	Percentage
Marital status		
Married	3021	99.1
Divorced/Widowed	28	0.9
Respondent's education level		
No education	1542	50.6
Primary	592	19.4
Secondary	751	24.6
Higher	164	5.4
Respondent's employment status		
Not employed	986	32.3
Employed	2063	67.7
Duration of pregnancy (trimester)		
First	737	24.2
Second	1228	40.3
Third	1084	35.6
Antenatal care		
No	1190	39.0
Yes	1859	61.0
Partner's education level		
No education	1213	39.8
Primary	588	19.3
Secondary	897	29.4
Higher	351	11.5
Partner's employment status		
Not employed	21	0.7
Employed	3028	99.3
Wealth class		
Poor	2089	68.5
Rich	960	31.5
Residence		
Rural	2111	69.2
Urban	938	30.8
Region		
North Central	409	13.4
North East	678	22.2
North West	1075	35.3
South East	253	8.3
South South	302	9.9
South West	332	10.9

Table 2: Frequency distribution of categorical variables (N=3049).

half (50.6%) of the respondents had no formal education. About 61% of respondents were on antenatal care and 40.3% were in the second trimester. About a third (31%) of the respondents resided in the urban area.

The mean BMI of the respondents was 23.81 ± 4.18 kg/m^2. The BMI was right skewed (Skewness=1.52), and the Kolmogorov-Smirnov test showed a significant deviation from normality (p-value <0.001). Figure 1 shows the distribution of maternal BMI.

Bivariate analyses

Table 3 shows the results of bivariate analyses of factors associated with BMI among pregnant women. Respondent's age, respondent's education level, respondent's employment status, number of living children, duration of pregnancy, antenatal care, partner's employment status, wealth class, residence, and region were significantly associated with maternal BMI.

Multivariate analyses

Figure 2 shows the plot of the robust Mahalanobis distance against the observation index. The plot implied the presence of several outliers in the predictor space. The 10-fold CV approach demonstrated that 0.55 quantile (55th percentile) had the lowest MSE among the 18 quantile models tested (Table 4). Table 5 shows the results of multivariate analyses of factors associated with maternal BMI. Similar to the results of OLS regression, quantile regression revealed that respondent's age, duration of pregnancy, wealth class, and residence were significant predictors of maternal BMI. Maternal BMI increased with age. Respondents in the second and third trimesters had higher BMI than those in the first trimester. BMI was higher among respondents in the rich wealth class than those in the poor wealth class. The cross-validated MSE for the OLS regression model was lower than that for the 0.55 quantile regression model (14.396 vs. 14.431).

Discussion

This study demonstrated that the model with 0.55 quantile had the highest predictive performance. At this quantile, respondent's age, duration of pregnancy, wealth class, and residence were significant predictors of maternal BMI. OLS regression also identified respondent's

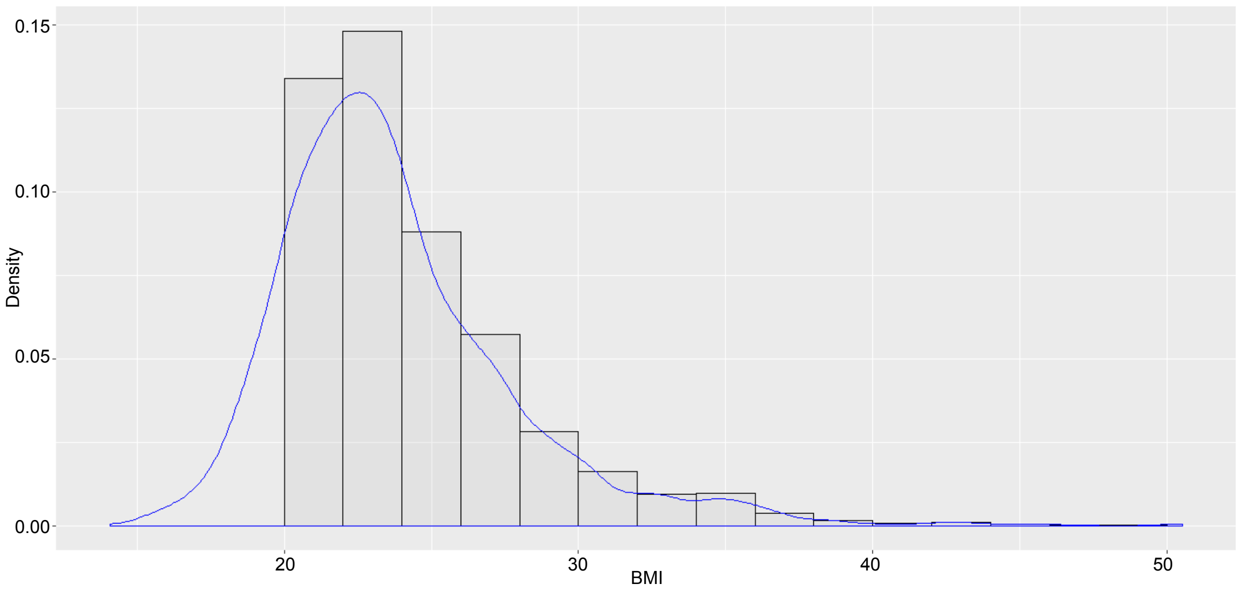

Figure 1: Histogram of maternal BMI.

Variable	Test statistic	P-value
Respondent's age (years)	12.3[†]	<0.001
Marital status	39014.0[‡]	0.480
Respondent's education level	284.8[*]	<0.001
Respondent's employment status	923670.0[‡]	<0.001
Number of living children	3.7[†]	<0.001
Preceding birth interval (years)	1.7[†]	0.097
Duration of pregnancy (trimester)	148.7[*]	<0.001
Antenatal care	880020.0[‡]	<0.001
Partner's education level	203.9[*]	<0.001
Partner's employment status	32924.0[‡]	0.779
Wealth class	646460.0[‡]	<0.001
Residence	717100.0[‡]	<0.001
Region	187.2[*]	<0.001

[†]Kendall's tau correlation, [‡]Wilcoxon rank sum test, [*]Kruskal-Wallis test.

Table 3: Bivariate analyses of factors associated with maternal BMI.

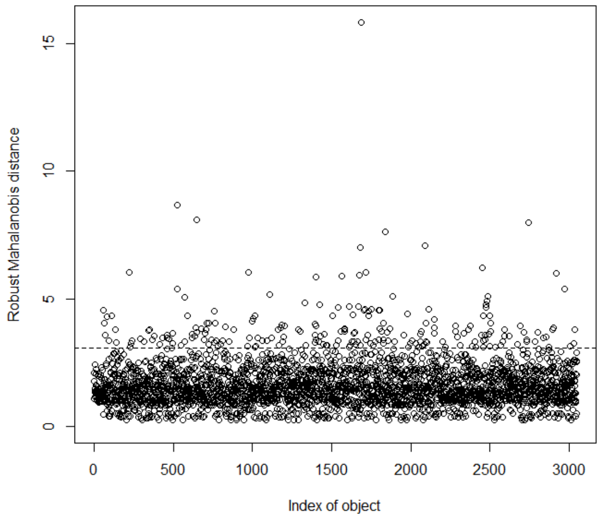

Figure 2: A Plot of the robust Mahalanobis distance against the observation index.

Quantile	MSE
0.10	29.735
0.15	25.790
0.20	22.559
0.25	20.239
0.30	18.482
0.35	17.118
0.40	16.069
0.45	15.247
0.50	14.702
0.55	14.448
0.60	14.488
0.65	14.910
0.70	15.822
0.75	17.389
0.80	20.066
0.85	25.171
0.90	36.353
0.95	61.892

Table 4: Results of 10-fold cross-validation of quantile regression models.

age, duration of pregnancy, wealth class, and residence as significant predictors of maternal BMI.

The finding that wealth class was associated with maternal BMI is consistent with Bhutta et al. result [7]. The significant association of residence with maternal BMI observed in this study agrees with the report of Senbajo and colleagues from Nigeria [3]. In contrast to the findings by Bhutta et al. [7], and Senjobi et al. [3], maternal education was not significantly associated with maternal BMI. Similar to the findings by Kusin et al. [18], number of living children and birth interval were not associated with maternal BMI in this study.

This study also found that OLS regression model had better predictive accuracy than quantile regression model. It has been recommended that OLS regression should not be used when the assumption of normality is violated or when the data contain outliers [19]. In this case, Madadizadeh et al. [20] suggested the use of quantile regression, instead. Although maternal BMI violated the assumption of normality and the data had many outliers, the OLS regression model demonstrated better fit than the quantile regression model, using machine learning approach. This finding upholds Breiman's proposition, that model assessment and model selection should be based on predictive accuracy [9].

This study had some limitations. For example, the study utilized NDHS data, the study design of which was cross-sectional; thus, the temporal sequence of the observed associations of predictors with maternal BMI could not be established. Analytic epidemiological studies (e.g. cohort design) would be more suitable in establishing the temporal sequence. Moreover, the determinants of maternal BMI investigated were not exhaustive because the data set was secondary. Other factors that might influence maternal BMI, such as diseased state, health education, and access to health care, could not be investigated. In spite of these limitations, this study utilized a population-based data; thus, the findings have considerable external validity. Also, this study provides evidence on model assessment using machine learning approach.

Variable	Quantile Regression (τ=0.55)			OLS Regression		
	Beta	SE	P-value	Beta	SE	P-value
Respondent's age (years)	0.121	0.016	<0.001	0.112	0.017	<0.001
Respondent's education level						
No education						
Primary	0.408	0.190	0.032	0.324	0.223	0.146
Secondary	0.087	0.238	0.713	0.342	0.258	0.185
Higher	2.573	0.441	<0.001	2.400	0.416	<0.001
Respondent's employment status						
Not employed						
Employed	-0.227	0.129	0.078	-0.085	0.155	0.584
Number of living children	-0.065	0.050	0.198	0.001	0.052	0.986
Preceding birth interval (years)	-0.020	0.053	0.700	0.050	0.055	0.356
Duration of pregnancy (trimester)						
First						
Second	0.955	0.157	<0.001	0.820	0.177	<0.001
Third	1.943	0.155	<0.001	1.865	0.182	<0.001
Antenatal care						
No						
Yes	0.212	0.144	0.142	0.283	0.166	0.088
Partner's education level						
No education						
Primary	0.235	0.184	0.201	0.409	0.216	0.058
Secondary	0.395	0.182	0.031	0.469	0.225	0.037
Higher	0.730	0.297	0.014	0.932	0.301	0.002
Wealth class						
Poor						
Rich	1.056	0.211	<0.001	1.091	0.211	<0.001
Residence						
Rural						
Urban	0.612	0.201	0.002	0.546	0.190	0.004
Region						
North Central						
North East	-0.600	0.207	0.004	-0.574	0.248	0.021
North West	-0.991	0.193	<0.001	-0.710	0.238	0.003
South East	-0.811	0.369	0.028	-0.154	0.323	0.634
South South	0.499	0.372	0.180	0.582	0.302	0.054
South West	-1.066	0.293	<0.001	-0.947	0.293	0.001
Cross-validated MSE	14.431			14.396		

Table 5: Multivariate analyses of factors associated with BMI – Quantile regression vs. OLS regression.

In conclusion, respondent's age, duration of pregnancy, wealth class, and residence were significant associated with maternal BMI. OLS regression model fit the data more than the quantile regression model. Predictive accuracy is a more suitable criterion for model assessment than the classical goodness-of-fit test or residual examination.

Acknowledgements

The authors would like to thank the Demographic and Health Surveys Program for allowing them to use the 2013 NDHS data.

References

1. Ahmed T, Hossain M, Sanin KI (2012) Global burden of maternal and child undernutrition and micronutrient deficiencies. Ann Nutr Metab 61: 8-17.

2. World Health Organization (2005) Severe Malnutrition: Report of a Consultation to Review Current Literature. World Health Organization, Geneva, p: 46.

3. Senbanjo IO, Olayiwola IO, Afolabi WA, Senbanjo OC (2013) Maternal and child under-nutrition in rural and urban communities of Lagos state, Nigeria: the relationship and risk factors. BMC Research Notes 6: 286.

4. Barker DJ, Clark PM (1997) Fetal undernutrition and disease in later life. Rev Reprod 2: 105-112.

5. Wu G, Bazer FW, Cudd TA, Meininger CJ, Spencer TE (2004) Maternal nutrition and fetal development. J Nutr 134: 2169-2172.

6. Marsal K (2002) Intrauterine growth restriction: Curr Opin Obstet Gynecol 14: 127-135.

7. Bhutta ZA, Gupta I, de'Silva H, Manandhar D, Awasthi S, et al. (2004) Maternal and child health: Is South Asia ready for change? BMJ 328: 816-819.

8. Begum S, Sen B (2005) Maternal Health, Child Well-being and Intergenerationally Transmitted Chronic Poverty: Does Women's Agency Matter?. SSRN Electronic J 32: 69-93.

9. Breiman L (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statist Sci 16: 199-231.

10. Tibshirani R, James G, Witten D, Hastie T (2013) An introduction to statistical learning-with applications in R. Springer, New York.

11. Mitchell JA, Rodriguez D, Schmitz KH, Audrain-McGovern J (2013) Greater screen time is associated with adolescent obesity: a longitudinal study of the BMI distribution from ages 14 to 18. Obesity 21: 572-575.

12. Mitchell JA, Pate RR, Espana-Romero V, O'Neill JR, Dowda M, et al. (2013) Moderate-to-vigorous physical activity is associated with decreases in body mass index from ages 9 to 15 years. Obesity 21: E280-293

13. Bottai M, Frongillo EA, Sui X, O'Neill JR, McKeown RE, et al. (2014) Use of quantile regression to investigate the longitudinal association between physical activity and body mass index. Obesity 22: E149–E156.

14. Koenker R, Bassett Jr G (1978) Regression quantiles. Econometrica: Journal of the Econometric Society 46: 33-50.

15. Petscher Y, Logan JA (2014) Quantile regression in the study of developmental sciences. Child Dev 85: 861-881.

16. National Population Commission (NPC) [Nigeria] and ICF International (2014) Nigeria Demographic and Health Survey 2013. NPC and ICF International, Abuja, Nigeria, and Rockville, Maryland, USA.

17. Hosmer Jr DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression. (3rd edn.) John Wiley & Sons, New Jersey.

18. Kusin JA, Kardjati S, Renqvist UH (1993) Chronic undernutrition in pregnancy and lactation. Proceedings of the Nutrition Society 52: 19-28.

19. Seber GA, Lee AJ (2012) Linear regression analysis. Vol. 936. John Wiley & Sons.

20. Madadizadeh F, Asar ME, Bahrampour A (2016) Quantile Regression and its Key Role in Promoting Medical Research. Iran J Public Health 45: 116.