MIKAILALSYS
Journal of Multidisciplinary Sciences

e-ISSN : 2987-2286
p-ISSN : 2987-3924

Index : Dimensions, Scilit, Crossref, Garuda, Semantic, Lens,Google Scholar,Base, etc

https://doi.org/10.58578/mikailalsys.v3i2.5322

Application of Quantile Regression and Ordinary Least Squares Regression in Modeling Body Mass Index in Federal Medical Centre Jalingo, Nigeria

Adeniyi Oyewole Ogunmola & Benjamin Ekene Okoye
Federal University Wukari, Taraba State, Nigeria
ogunmola@fuwukari.edu.ng

Article Info:

Submitted:	Revised:	Accepted:	Published:
Mar 7, 2025	Mar 22, 2025	Apr 3, 2025	Apr 8, 2025

Abstract

Body mass index is a measure of nutritional status of an individual. Malnutrition is a leading public health problem in developing countries like Nigeria, it is also a major cause of morbidity and mortality. In this study, Body mass index is modeled using ordinary least squares method and quantile regression method. Data is collected from Antiretroviral therapy Clinic in Federal Medical Centre, Jalingo. Variables in the data collected are the Body mass index, age, weight, height, sex and occupation of the patients. Results showed that the ordinary least square regression and quantile regression at 25^{th} percentile, median percentile, 75^{th} percentile and 95^{th} percentile fit the data. Weight, age, sex and height of patients are significant in determining the BMI of the patients when OLS method is applied. While weight, sex and height of patients are significant in determining the BMI of the patients. It is also discovered that OLS method fits the data more than quantile regression method using AIC and MSE

Keywords: Body mass index (BMI), Ordinary least squares (OLS), Quantile regression, modeling, Akaike Information Criterion (AIC)

Mikailalsys is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

INTRODUCTION

Body Mass Index (BMI) is a widely used anthropometric measurement that assesses an individual's body weight in relation to their height. It is commonly utilized to categorize individuals into different weight groups, such as underweight, normal weight, overweight, and obesity, based on potential health risks (Adedia et al. 2020, Mohajan and Mohajan, 2023). These classifications are employed in various health assessments and public health guidelines to gauge the risk of weight-related health conditions, including cardiovascular disease, diabetes, and other metabolic disorders (NIH 2024, Safaei et al. 2021, WHO 2024a, WHO 2024b).

Modeling BMI requires understanding the relationships between various demographic, genetic, environmental, and behavioral factors that influence body weight.

METHODOLOGY

The data source in this study is secondary data obtained from Anti-retroviral therapy clinic in Federal Medical Centre Jalingo, Taraba State. R statistical software is used to analyse the data. Since the study is on applying OLS and quantile regression methods to estimate BMI from some regressors based on the data collected, the dependent variable is BMI while the independent variables are age, sex , weight, height and occupation of patients in the clinic. Generally the regression BMI model is expressed as:

$$BMI = \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 weight + \beta_4 height + \beta_5 occupation + \varepsilon.$$

Quantile regression method and OLS method are statistical methods used for modeling relationship between a dependent variable (outcome) and a set of independent variables (predictors or regressors). However, they serve different purposes and make different assumptions.

OLS Regression

OLS estimates the conditional mean of the dependent variable given the independent variables. It minimizes the sum of the squared differences (errors) between the observed values and the predicted values. That is, it minimizes the distance between the predicted regression line and the actual data points.

$$\hat{\beta} = arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\boldsymbol{\beta})^2$$

The assumptions of the model include:

Linearity: the relationship between the independent and the dependent variables is linear.

Homoscedasticity: the variance of errors is constant across all levels of the independent variables

Normality: The residuals (errors) are normally distributed and they are independent to each other.

The OLS is best used when interested in estimating the average relationship between the predictors and outcome.

Quantile Regression

Quantile regression estimates the conditional quantile of the dependent variable given the independent variables. Instead of modeling the mean, quantile regression focuses on the quantile on a specific quantile.

Instead of minimizing the squared errors, quantile regression minimizes the sum of absolute deviations weighted according to the desired quantile. The loss function is minimized differently than for the mean.

$$\hat{\beta}_r = arg\min_{\beta} \sum_{i=1}^{n} \rho_r (y_i - X_i\boldsymbol{\beta})^2$$

where ρ_r is the loss function that gives different weights depending on whether the residual is above or below the quantile.

No Assumptions about Error Distribution: Unlike OLS, quantile regression does not assume normality or homoscedasticity, making it more robust to outliers and heteroscedasticity.

Quantile regression is useful when it is desired to understand how the relationship between the independent variables and the outcome differs across different points of the distribution. Here 25th. 50th, 75th and 95th percentiles are compared.

Comparing the OLS method with the different quantiles, Akaike information criterion (AIC) will be used to select the best model. Mean squared error (MSE) will also be used to measure prediction accuracy by quantifying how close predicted values are to actual values.

AIC

AIC is a measure of a model's goodness of fit while penalizing for the number of parameters used. It helps in comparing different models and selecting the best one by balancing model complexity and goodness of fit (Akaike 1974, Hurvich and Tsai 1989, Burnham and Anderson 2004). AIC is given as:

$$AIC = 2k - 2log_e L$$

where, k is the number of parameters in the model, and L is the likelihood of the model (a measure of how well the model fits the data). Lower AIC values indicate a better model, as it balances goodness of fit with simplicity (parsimony). AIC penalizes models with more parameters to avoid over fitting. Even if a model fits the data well, if its parameters are too many, its AIC will be higher.

MSE

MSE is a measure of the average squared difference between the actual (observed) values and the predicted values from the model (Hastie et al 2009, Montgomery et al. 2012, and James et al. 2013). It is given as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where, y_i is the observed value, $\hat{y}_i$ is the predicted value and n is the number of observations. Lower MSE values indicate a better model, as the predictions are closer to the actual values. MSE heavily penalizes large errors because of the squaring of the residuals (differences between observed and predicted values).

RESULTS

Table 1 shows the parameter estimates from the conditional mean (OLS method), 25th quantile, 50th quantile, 75th quantile and 95th quantile of the patients BMI given the age, sex, weight, height and occupation of the patients. The significance of the parameter at 0.05

level of significance is denoted by the asterisk symbol (*) which indicates the parameter is significant.

Table 1: Parameter estimates for OLS and Quantiles regressions

	Estimated Parameters values				
Parameters	**OLS**	**q = 0.25**	**q = 0.50**	**q = 0.75**	**q = 0.95**
Intercept	2.220994*	1.48841	1.81151*	2.83004*	4.14132*
Weight	0.298043*	0.29616*	0.29712*	0.30199*	0.30016*
Height	-0.345810*	-0.30904*	-0.32688	-0.34383*	-0.37700
Age	-0.014380*	-0.01871	-0.001007	-0.01606	-0.00683
Sex	2.039214*	1.82389*	2.06017*	2.28828*	2.08832*
Occupation	0.012203	0.07873	0.05022	-0.05207	0.26450
R-Square	0.85				
MSE	2.814698	4.117106	2.846878	3.937304	11.25643
AIC	2631.967	2697.733	2669.001	2829.183	3243.804

From the table, when OLS method is fitted to the data, about 85 percent of the variations in the BMI are explained by the set of independent variables, and intercept, weight, height, age and sex of the patients are significant in determining the patients BMI at 5 percent level of significant. When 25th quantile regression is fitted to the data, weight, height and sex of patients are significant in determining the patients BMI. When 50th quantile regression is applied, intercept, weight, and sex are significant in determining the patients BMI at 5 percent level of significance. When 75th quantile regression is applied, intercept, weight, height and sex are significant in determining the BMI of the patients. Lastly, when 95th quantile regression is applied to the data, intercept, weight and sex are significant in determining the patients BMI at 5 percent level of significant. In all the five models, sex and weight of patients are significant in determining the patients BMI. In the OLS and three out of the four quantiles regression, intercept is significant. In the OLS and two out of the four quantiles regression, height is significant. In only OLS model is age of patients significant in determining the patients BMI.

Based on AIC, OLS model, 25th quantile and 50th quantile regressions are close and lower than 75th quantile and 95th quantile regressions. The 95th quantile regression has the largest AIC value farthest away from those that are close (OLS, 25th quantile and 50th quantile regressions). The 95th quantile regression and the 75th quantile regression model are far

away from being a better model suitable for modeling the patients BMI based on the data. While OLS model, 25th quantile regression and 50th quantile regression, are much better in modeling the patients BMI. The best among these three models is the OLS with the lowest AIC value of 2631.967.

Based on MSE, OLS model and 50th quantile regression are the closely best models for predicting accurately the patients BMI based on the data. The worst model among the five models for predicting the patients BMI accurately is 95th quantile regression model. The best model for prediction is the OLS model.

CONCLUSION

Generally, OLS model and 50th quantile regression give both a better model and better prediction accuracy compared to the remaining three models based on AIC and MSE. Then intercept, weight and sex are best variables that determined BMI of the patients, while addition of age and sex of patients will come as the next model that fits the and predicts BMI of the patients. The overall best model is that of the OLS.

Recommendation

The OLS model is the best and it should be used when it is desired to estimate the average relationship between the predictors and the outcome. The 50th quantile regression is the best model among the quantile regression, and it should be used to understand how the relationship between the independent variables and the outcome differs across the median of the distribution.

REFERENCES

Adedia D., Boakye A.A., Mensah D., Lokpo S.Y., Afeke I., and Duedu K.O. (2020). *Comparative assessment of anthropometric and bioimpedence methods for determining adiposity*. Heliyon, Volume 6, Issue 12, e05740, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2020.e05740. (https://www.sciencedirect.com/science/article/pii/S2405844020325834)

Akaike, H. (1974). *A new look at the statistical model identification*. IEEE Transactions on Automatic Control, 19(6), 716-723.

Burnham, K. P., & Anderson, D. R. (2004). *Multimodel inference: Understanding AIC and BIC in model selection*. Sociological Methods & Research, 33(2), 261-304.

Hurvich, C. M., & Tsai, C. L. (1989). *Regression and time series model selection in small samples*. Biometrika, 76(2), 297-307.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Mohajan D. and Mohajan H.K. (2023). *Body Mass Index (BMI) is a Popular Anthropometric Tool to Measure Obesity Among Adults*. Paradigm Academic Press Journal of Innovations in Medical Research ISSN 2788-7022 APR. 2023 VOL.2, NO.4. doi:10.56397/JIMR/2023.04.06

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). John Wiley & Sons.

National Heart, Lung, and Blood Issue (NIH) (2024). Assessing Your Weight and Health Risk. Assessing Your Weight and Health Risk (nih.gov) September, 15th 2024.

Safaei M., Sundararajan E.A., Driss M., Boulila W., Shapi'I A. (2021). *Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity*. Computers in Biology and Medicine, Volume 136, 2021, 104754, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104754. (https://www.sciencedirect.com/science/article/pii/S0010482521005485)

World Health Organsation (WHOa) (2024). *Health Topics Obesity*. https://www.who.int/health-topics/obesity September, 15th 2024.

World Health Organsation (WHOb) (2024). *Health Topics Diabetics*. https://www.who.int/health-topics/diabetics September, 15th 2024.