© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1719320

Modelling the Prevalence and Predictors of Childhood Obesity in Children Aged 5–10 Years Using Regularized Regression: A Cross-Sectional Study in Kogi State, Nigeria

SUNDAY JOHN EKELE[1], G. I. ONWUKA[2], TOLULOPE O. JAMES[3]
[1, 2, 3] *Department of Mathematics (Statistics), Abdullahi Fodiyo University of Science and Technology, Aliero, Kebbi State*

Abstract*— Background: Childhood obesity is rising fastest in low- and middle-income countries, yet sub-national evidence from Nigeria remains scarce. We examined the distribution and predictors of body mass index (BMI) among primary-school children aged 5–10 years in Kogi State and compared the performance of ordinary least squares (OLS) and three regularized regression estimators. Methods: In a descriptive cross-sectional design, anthropometric measurements (weight, height, BMI) and questionnaire data (age, sex, physical activity, diet type) were obtained from primary-school pupils selected by multi-stage stratified, cluster and simple random sampling. BMI was modelled with OLS, Lasso (L_1), Ridge (L_2) and Elastic Net (combined L_1–L_2) regression. Penalty parameters were tuned by cross-validation; models were compared using R^2, AIC, BIC, MSE and RMSE, with residual diagnostics for normality, homoscedasticity and autocorrelation. Results: Weight, height and BMI category were strong and statistically robust predictors of BMI across all four estimators ($p < 0.001$). Age and sex were not significant; the significance of physical activity and diet type was not supported once test statistics were recomputed from the reported coefficients and standard errors (see Note to Table 5). Elastic Net returned the most favourable fit metrics ($R^2 = 0.987$, AIC = 131.0, BIC = 150.0, MSE = 2.25, RMSE = 1.50) and the lowest variance-inflation factors among the models. All models satisfied residual assumptions (Shapiro–Wilk $p > 0.05$; Breusch–Pagan $p > 0.05$; Durbin–Watson ≈ 1.98). Conclusions: Regularized regression, particularly Elastic Net, controls multicollinearity more effectively than OLS for this anthropometric data structure. Because BMI is a deterministic function of weight and height, the very high R^2 should be interpreted with caution. The findings nonetheless support continued investment in school-based physical-activity and nutrition programmes, consistent with the global evidence base.

***Index Terms*— Childhood Obesity, Body Mass Index, Elastic Net, Regularized Regression, Multicollinearity, Nigeria**

I. INTRODUCTION

Overweight and obesity, defined by abnormal or excessive fat accumulation, are among the fastest-growing public-health threats of the twenty-first century. The most recent pooled analysis of 3,663 population-representative studies by the NCD Risk Factor Collaboration estimated that 159 million school-aged children and adolescents (aged 5–19 years) were living with obesity in 2022, with the prevalence rising in the large majority of countries between 1990 and 2022 (NCD-RisC, 2024). Forecasts from the Global Burden of Disease programme project that, without decisive intervention, the burden among children and adolescents will continue to climb through 2050 (GBD 2021 Adolescent Obesity Collaborators, 2025).

Body mass index (BMI), computed as weight in kilograms divided by height in metres squared, remains the standard population-level metric for classifying weight status, although it does not capture body-fat distribution or composition and should be interpreted alongside its known limitations (Nuttall, 2015). Childhood obesity tracks into adulthood and elevates the long-term risk of type 2 diabetes, cardiovascular disease, certain cancers, and psychosocial harm, making early identification of modifiable determinants a priority for health systems.

The burden is shifting toward low- and middle-income countries. In sub-Saharan Africa, a systematic scoping

review of 81 studies across 20 countries documented a rising and under-monitored childhood obesity burden, with most evidence concentrated in South Africa and Nigeria and dominated by cross-sectional designs (Danquah et al., 2020). A more recent meta-analysis of preschool children in the region estimated a pooled overweight/obesity prevalence of 14.8% (Tolasa et al., 2025). For Nigeria specifically, a systematic review and meta-analysis confirmed an upward national trajectory but emphasised the limited epidemiological understanding that constrains targeted public-health responses (Adeloye et al., 2021). Localized Nigerian and West African studies report rising prevalence linked to dietary transition, sedentary behaviour and rapid urbanization (Adeniyi et al., 2019; Ganle et al., 2019; Diallo et al., 2023).

Statistically, anthropometric and behavioural predictors of obesity are frequently collinear, which destabilises ordinary least squares (OLS) estimates and inflates their standard errors. Regularized estimators address this directly: Ridge regression shrinks coefficients via an L_2 penalty (Hoerl & Kennard, 1970), Lasso performs simultaneous shrinkage and variable selection via an L_1 penalty (Tibshirani, 1996), and the Elastic Net combines both penalties to retain groups of correlated predictors while preserving sparsity (Zou & Hastie, 2005). Despite their suitability for correlated public-health data, these methods are rarely applied to childhood-obesity modelling in the Nigerian context.

Aim and objectives. This study analyses the distribution and predictors of BMI among children aged 5–10 years in Kogi State, Nigeria, and compares OLS, Lasso, Ridge and Elastic Net regression in order to (i) quantify associations between BMI and demographic/anthropometric/behavioural factors, (ii) identify the predictors most robustly associated with BMI, and (iii) determine which estimator best balances fit and parsimony for this data structure.

II. MATERIALS AND METHODS

2.1 Study design and population

A descriptive cross-sectional design was used to capture weight status at a single time point among primary-school pupils aged 5–10 years in Kogi State, Nigeria. The sampling frame comprised both public and private primary schools across urban and rural communities, ensuring socio-economic and geographic diversity.

2.2 Sample size and sampling technique

A target sample of 1,000 pupils was specified to support reliable population inference. A multi-stage procedure was applied: (i) *stratified sampling* of schools into public and private strata; (ii) *cluster sampling* by geographic location (urban/rural); and (iii) *simple random sampling* of schools within strata and of eligible pupils within selected schools. This procedure was designed to reflect the demographic diversity of the population while limiting selection bias.

2.3 Data collection

Weight was measured with a calibrated digital scale (light clothing, no shoes) and height with a stadiometer, recorded in metres. BMI was computed as BMI = weight (kg) / [height (m)]² and classified against age-specific WHO/CDC percentile charts. Demographic and lifestyle variables (age, sex, diet type, physical activity) were collected through a structured questionnaire administered with the assistance of trained class teachers using digital forms. Written informed consent was obtained from parents/guardians, instruments were calibrated throughout fieldwork, and research assistants were trained to standardise measurement.

2.4 Statistical models

For a sample of n observations and p predictors, the response y (BMI) is modelled as a linear function of the design matrix X:

$$y = X\beta + \varepsilon, \quad y \in \mathbb{R}^n, \; X \in \mathbb{R}^{n \times p}, \; \beta \in \mathbb{R}^p, \; \varepsilon \sim N(0, \sigma^2 I_n) \qquad (1)$$

The OLS estimator minimises the residual sum of squares:

$$\hat{\beta}OLS = \arg\min_\beta \|y - X\beta\|_2^2 \qquad (2)$$

but is unstable when predictors are strongly correlated (as below). Two penalties address this. Ridge regression adds an L2 penalty, shrinking coefficients without setting any to zero:

$$\hat{\beta}ridge = \arg\min_\beta \{ \|y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 \} \qquad (3)$$

Lasso replaces the L2 penalty with an L1 penalty, inducing sparsity by driving weak coefficients to exactly zero:

$$\hat{\beta}\text{lasso} = \arg\min_{\beta} \{ \|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 \}, \quad \|\beta\|_1 = \Sigma_{j=1}^{p} |\beta_j| \qquad (4)$$

The Elastic Net combines both penalties, retaining correlated predictors as a group while preserving variable selection:

$$\hat{\beta}\text{EN} = \arg\min_{\beta} \{ \|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 \} \qquad (5)$$

Introducing a mixing parameter α, the combined penalty can be written equivalently as

$$\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 = \lambda [\alpha\|\beta\|_1 + \tfrac{1}{2}(1 - \alpha)\|\beta\|_2^2] \qquad (6)$$

with overall penalty strength $\lambda = \lambda_1 + \lambda_2$ and mixing parameter $\alpha = \lambda_1 / (\lambda_1 + \lambda_2)$, $0 \le \alpha \le 1$. At $\alpha = 1$ the estimator reduces to Lasso, and at $\alpha = 0$ to Ridge.

Because the L1 term is non-differentiable at zero, the solution is obtained by cyclic coordinate descent. For each coefficient the update is

$$\beta_j \leftarrow S((1/n) \Sigma_i x_{ij}(y_i - \hat{y}_i^{(j)}), \lambda\alpha) / (1 + \lambda(1 - \alpha)) \qquad (7)$$

where $\hat{y}_i^{(j)} = \Sigma_{k \neq j} x_{ik}\beta_k$ and $S(\cdot)$ is the soft-thresholding operator,

$$S(z, \gamma) = \text{sign}(z) \cdot \max(|z| - \gamma, 0) \qquad (8)$$

The fitted (prediction) model for BMI is therefore the linear predictor evaluated at the penalized estimates; the penalty terms govern estimation only and do not enter the prediction equation:

$$\hat{Y} = \hat{\beta}0 + \Sigma_{j=1}^{p} \hat{\beta}_j X_j \qquad (9)$$

2.5 Model evaluation

Penalty parameters (λ_1, λ_2) were tuned by cross-validation. Predictive performance and parsimony were assessed with the coefficient of determination (R^2), Akaike and Bayesian information criteria (AIC, BIC), mean squared error (MSE), root mean squared error (RMSE) and mean absolute error (MAE):

$$\text{MSE} = (1/n) \Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (10)$$

$$\text{RMSE} = \sqrt{(\text{MSE})}, \quad \text{MAE} = (1/n) \Sigma_{i=1}^{n} | y_i - \hat{y}_i | \qquad (11)$$

$$R^2 = 1 - [\Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2] / [\Sigma_{i=1}^{n} (y_i - \bar{y})^2] \qquad (12)$$

Multicollinearity was quantified with the variance-inflation factor (VIF). Residual assumptions were tested with the Shapiro–Wilk (normality), Breusch–Pagan (homoscedasticity; Breusch & Pagan, 1979) and Durbin–Watson (autocorrelation; Durbin & Watson, 1950) statistics.

III. RESULTS

3.1 Descriptive statistics

Children averaged 7.5 years of age (range 5–10). Mean weight was 31.92 kg (SD 10.11) and mean height 1.25 m (SD 0.15). Mean BMI was 21.22 (SD 8.69; range 7.1–47.0) with a slight positive skew (0.64), consistent with an emerging upper tail of overweight/obese children. Physical activity averaged 4.94 h/week (SD 2.58). Table 1 summarises the distribution of all study variables.

Table 1. *Descriptive statistics for the childhood-obesity dataset, Kogi State.*

Variable	Mean	Median	Min	Max	SD
Age (years)	7.500	8.000	5.000	10.000	1.723
Sex (1=M, 2=F)	1.566	2.000	1.000	2.000	0.496
Weight (kg)	31.918	31.500	15.200	49.900	10.113
Height (m)	1.254	1.260	1.000	1.500	0.145
BMI (kg/m²)	21.216	19.800	7.100	47.000	8.693
BMI category	1.762	2.000	1.000	3.000	0.792
Physical activity (h/wk)	4.936	5.000	1.000	9.000	2.584
Diet type (1=unhealthy)	1.522	2.000	1.000	2.000	0.500

Note. Sex coded 1 = male, 2 = female; a mean of 1.57 indicates a modestly higher proportion of girls. BMI category coded 1 = normal, 2 = overweight, 3 = obese.

3.2 Multicollinearity

Under OLS, height (VIF = 24.04) and age (VIF = 16.93) exhibited severe multicollinearity, with weight in the moderate-to-high range (VIF = 10.15). All three regularized estimators reduced these values, with Lasso achieving the largest reductions and Elastic Net providing a balanced reduction while retaining all predictors (Table 2). The behavioural and demographic indicators (physical activity, sex, diet type) remained in the acceptable range throughout.

Table 2. *Variance-inflation factors by estimator.*

Feature	OLS	Lasso	Ridge	Elastic Net	Risk
Age (years)	16.93	10.82	12.53	11.42	High

Weight (kg)	10.15	6.86	7.94	7.14	Moderate–high
Height (m)	24.04	15.73	18.31	16.55	High
Physical activity	4.46	2.92	3.42	3.11	Acceptable
Sex	2.29	1.83	1.98	1.89	Acceptable
Diet type	2.07	1.65	1.77	1.70	Acceptable

3.3 Residual diagnostics

All four models satisfied the key regression assumptions. Shapiro–Wilk tests indicated normally distributed residuals ($p > 0.05$), Breusch–Pagan tests indicated homoscedasticity ($p > 0.05$), and Durbin–Watson statistics were close to 2 (≈ 1.98), indicating no meaningful autocorrelation (Table 3).

Table 3. *Residual diagnostics by estimator.*

Test	Statistic	Threshold	Interpretation
Shapiro–Wilk (normality)	$W \approx 0.93$ (all models)	$p > 0.05$	Residuals normal
Breusch–Pagan (homoscedasticity)	$p \approx 0.96$ (all models)	$p > 0.05$	Constant variance
Durbin–Watson (autocorrelation)	≈ 1.85–1.99 (all models)	≈ 2	Independent errors

3.4 Model comparison

Elastic Net produced the most favourable values on every comparison metric: the highest R^2 (0.987), the lowest information criteria (AIC = 131.0, BIC = 150.0) and the smallest error (MSE = 2.25, RMSE = 1.50). The three remaining estimators were closely matched and somewhat poorer (Table 4).

Table 4. *Model comparison criteria.*

Model	R^2	AIC	BIC	RMSE
OLS	0.9569	131.84	150.08	1.8013
Lasso	0.9565	132.42	150.63	1.8106
Ridge	0.9567	132.09	150.35	1.8061
Elastic Net	0.9868	131.02	150.00	1.5014

Note. Reported information criteria (AIC ≈ 131, BIC ≈ 150) imply an analytic sample of roughly 100 observations rather than the stated 1,000; this discrepancy should be reconciled against the source output (see Statistical-validation note, §3.6).

3.5 Elastic Net coefficient estimates

In the Elastic Net model, weight, height and BMI category were strongly associated with BMI (all $p < 0.001$) with internally consistent test statistics. Age and sex were not significant. The reported test statistics for physical activity and diet type were not reproducible from their coefficients and standard errors; recomputing t = coefficient / standard error yields $|t| < 1.3$ for both, indicating non-significance at conventional thresholds (Table 5). Identical patterns hold in the Lasso, Ridge and OLS estimates.

Table 5. *Elastic Net regression estimates for BMI.*

Predictor	Coef.	Std. error	T	p	Sig.
Intercept	43.021	0.872	49.32	< 0.001	***
Age (years)	0.054	0.042	1.29	0.198	n.s.
Sex	0.036	0.147	0.25	0.805	n.s.
Weight (kg)	0.672	0.0085	79.09	< 0.001	***
Height (m)	−34.793	0.522	−66.70	< 0.001	***
BMI category	−0.779	0.114	−6.82	< 0.001	***
Physical activity	0.020	0.028	0.70	0.486	n.s.
Diet type	−0.173	0.146	−1.18	0.237	n.s.

Note. t-statistics are recomputed as coef / std. error for internal consistency. In the source manuscript, physical activity and diet type were reported as highly significant, but those t-values are arithmetically inconsistent with the tabulated coefficients and

standard errors and are corroborated as non-significant by the OLS estimates; the discrepancy should be re-derived from the original software output. *** $p < 0.001$; n.s. = not significant.

3.6 Statistical-validation note

Internal consistency. Three reporting issues were identified and should be reconciled before publication: (i) the t-statistics for physical activity and diet type are not reproducible from the reported coefficients and standard errors (Table 5); (ii) the information criteria imply $n \approx 100$ rather than the stated $n = 1{,}000$ (Table 4); and (iii) the model-specification intercept reported elsewhere in the source ($\beta_0 \approx 21.5$) differs from the estimated intercept (≈ 43.0).

Structural caution. Because BMI is defined as weight / height², regressing BMI on weight and height (and on BMI category, which is itself derived from BMI) is close to an algebraic identity. This largely explains the exceptionally high R^2 and means the dominant "predictors" are partly mechanical rather than epidemiological. For a behaviourally interpretable model we recommend either modelling BMI as a function of behavioural and demographic variables only (excluding weight, height and BMI category) or modelling weight status with logistic/ordinal regression.

IV. DISCUSSION

This study examined the predictors of BMI among children aged 5–10 years in Kogi State and compared four regression estimators. The most robust and internally consistent associations were with the anthropometric variables — weight, height and BMI category — which were highly significant across all models. Age and sex were not significant, consistent with several previous reports (Kaioglou & Venetsanou, 2017; da Silva et al., 2019), and plausibly reflecting the narrow 5–10-year age band in which anthropometric variation already captures most developmental differences.

Once test statistics were recomputed for internal consistency, physical activity and diet type were not statistically significant in the adjusted models — a result that agreed with the OLS estimates and contrasts with the source manuscript's original significance claims. This null finding does not contradict the wider evidence base: physical inactivity and energy-dense diets are well-established drivers of childhood obesity in global (NCD-RisC, 2024; GBD 2021 Adolescent Obesity Collaborators, 2025) and regional analyses (Adeniyi et al., 2019; Ganle et al., 2019; Diallo et al., 2023). Their non-significance here is most plausibly attributable to the narrow age range, measurement error in self-reported behaviour, and the dominance of the anthropometric predictors that are definitionally linked to BMI.

Methodologically, the results illustrate why regularization is valuable for correlated anthropometric data: the OLS VIFs for height and age signalled severe multicollinearity, and all three penalized estimators attenuated it. Elastic Net achieved the best balance of fit and parsimony, consistent with its theoretical grouping property for correlated predictors (Zou & Hastie, 2005). However, the exceptionally high R^2 should not be read as evidence of strong epidemiological prediction, because BMI is an algebraic function of two of its predictors (see §3.6).

Substantively, the slight positive skew of BMI and the spread of physical-activity and dietary habits are consistent with an early-stage obesity transition in this population, mirroring the upward national trajectory reported for Nigeria (Adeloye et al., 2021) and the broader sub-Saharan pattern (Danquah et al., 2020; Tolasa et al., 2025). The findings reinforce the case for school-based physical-activity and nutrition programmes and parental education, in line with global policy guidance.

V. LIMITATIONS

First, the design is cross-sectional, precluding causal inference. Second, the analysis is geographically restricted to Kogi State, limiting generalizability. Third, behavioural variables were self-reported and may be subject to recall and social-desirability bias.

VI. CONCLUSION

Among children aged 5–10 years in Kogi State, weight, height and BMI category were the most robust correlates of BMI, while age and sex were not significant; the apparent significance of physical

activity and diet type was not supported once test statistics were verified. Elastic Net regression best controlled multicollinearity and balanced fit and parsimony, making it a suitable estimator for correlated anthropometric data. Given the definitional dependence of BMI on weight and height, the headline R^2 should be interpreted cautiously, and future work should adopt longitudinal designs, expand geographic coverage, and model weight status using behaviourally interpretable specifications to better inform school- and policy-level interventions.

DECLARATIONS

Ethics: Written informed consent was obtained from parents/guardians; school authorities were briefed on the study's objectives and procedures. *Funding:* None declared. *Conflicts of interest:* The authors declare no competing interests. *Data availability:* Data are available from the corresponding author on reasonable request.

REFERENCES

[1] Adeloye, D., Ige-Elegbede, J. O., Ezejimofor, M., Owolabi, E. O., Ezeigwe, N., Omoyele, C., … Harhay, M. O. (2021). Estimating the prevalence of overweight and obesity in Nigeria in 2020: A systematic review and meta-analysis. Annals of Medicine, 53(1), 495–507. https://doi.org/10.1080/07853890.2021.1897665

[2] Adeniyi, O. F., Fagbenro, G. T., & Olatona, F. A. (2019). Overweight and obesity among school-aged children and maternal preventive practices against childhood obesity in Lagos, Southwest Nigeria. International Journal of MCH and AIDS, 8(1), 70–83. https://doi.org/10.21106/ijma.273

[3] Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica, 47(5), 1287–1294.

[4] da Silva, D. D., de Lima, M. V. M., Muniz, P. T., Holanda, M. N., Câmara, O. F., Monteiro, A., & Wajnsztejn, R. (2019). Prevalence and factors associated with obesity in children under five years old in Rio Branco, Acre. Journal of Human Growth and Development, 29(2), 263–273.

[5] Danquah, F. I., Ansu-Mensah, M., Bawontuo, V., Yeboah, M., & Kuupiel, D. (2020). Prevalence, incidence, and trends of childhood overweight/obesity in sub-Saharan Africa: A systematic scoping review. Archives of Public Health, 78, 109. https://doi.org/10.1186/s13690-020-00491-2

[6] Diallo, R., Baguiya, A., Baldé, M. D., Camara, S., Diallo, A., Camara, B. S., … Compaoré, E. (2023). Prevalence and factors associated with overweight in children under five years in West African countries. Journal of Public Health Research, 12(3). https://doi.org/10.1177/22799036231181845

[7] Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression. Biometrika, 37(3–4), 409–428.

[8] Ganle, J. K., Boakye, P. P., & Baatiema, L. (2019). Childhood obesity in urban Ghana: Evidence from a cross-sectional survey of in-school children aged 5–16 years. BMC Public Health, 19, 1561. https://doi.org/10.1186/s12889-019-7898-3

[9] GBD 2021 Adolescent Obesity Collaborators. (2025). Global, regional, and national prevalence of child and adolescent overweight and obesity, 1990–2021, with forecasts to 2050. The Lancet, 405. https://doi.org/10.1016/S0140-6736(25)00397-6

[10] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.

[11] Kaioglou, V., & Venetsanou, F. (2017). Overweight and obesity prevalence in young children living in Athens. Public Health Open Journal, 2(1), 26–32.

[12] NCD Risk Factor Collaboration (NCD-RisC). (2024). Worldwide trends in underweight and obesity from 1990 to 2022: A pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. The Lancet, 403(10431), 1027–1050. https://doi.org/10.1016/S0140-6736(23)02750-2

[13] Nuttall, F. Q. (2015). Body mass index: Obesity, BMI, and health — A critical review. Nutrition Today, 50(3), 117–128.

[14] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288.

[15] Tolasa, B., et al. (2025). Prevalence of overweight and obesity and its associated factors

among preschool children in sub-Saharan Africa: A systematic review and meta-analysis. (PMC12908065).

[16] World Health Organization. (2018). Report of the Commission on Ending Childhood Obesity. Geneva: WHO.

[17] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x