

Temporal Dynamics of Obesity: A Multivariate Regression Analysis of Caloric Supply, Population Growth, and Autocorrelated Prevalence Trends

Chenyu Ding

*School of Mathematics, Tianjin University, Tianjin, China
281805322@qq.com*

Abstract: The global increase in obesity prevalence is a significant public health issue in the 21st century. In recent years, over two billion people were overweight or obese. Existing research has limitations in understanding obesity trends. This study focuses on the autoregressive nature of obesity, the interaction between caloric supply and population demographics, and improving analytical methods. This paper employ three advanced statistical methods using comprehensive, nationally representative data from the NCD Risk Factor Collaboration (NCD-RisC) 2017. These salient results evince that: past adiposity metrics constitute a notable determinant of present-day incidence; the interplay linking energy availability and corpulence is observably modulated by discrete population-based factors; and the integration of autoregressive elements into predictive algorithms markedly refines their prognostic accuracy. By integrating these crucial temporal, environmental, and demographic factors, this research provides a significantly refined and more holistic understanding of global obesity dynamics. This integrated framework offers valuable contributions to epidemiological theory concerning the multifactorial aetiology of obesity and informs the design of more effective, evidence-based policy interventions.

Keywords: Obesity, Overweight, Daily calorie supply, Multiple Regression Analysis

1. Introduction

The global rise in obesity prevalence represents one of the most pressing public health challenges of the 21st century. The NCD-RisC consortium reported that in 2017, more than two billion individuals globally were categorized as having excess body weight or obesity, marking a 50% surge in prevalence from 1980 [1]. While the fundamental energy imbalance framework (caloric intake exceeding expenditure) explains individual weight gain, population-level obesity trends exhibit complex temporal dynamics that cannot be fully captured by simple cross-sectional analyses [2].

A critical yet understudied aspect of obesity epidemiology is the autocorrelation of prevalence trends, where current obesity rates are intrinsically linked to historical patterns. Cohort studies have shown that populations exposed to obesogenic environments during key developmental windows may experience heightened vulnerability later in life [3]. This path dependency suggests that

conventional regression models, ignoring temporal dependence, may yield biased estimates of other covariates' effects.

The role of dietary energy supply remains contentious in this context. Ecological investigations suggest a correlation between heightened caloric accessibility and the increasing prevalence of obesity [4], the effect size varies substantially after accounting for food waste and socio-economic disparities [5]. Further complicating this ambiguity, population demographics interplay with food environments; aging populations and increasing urbanization could independently heighten the risks of obesity, even when caloric supply remains stable [6].

While obesity research is extensive, gaps remain. Prior studies often overlook the autoregressive nature of obesity prevalence, neglecting time dependence. Quantitative analysis of the interaction between calorie supply and population structure is lacking, hindering accurate assessment of their combined impact. Static covariates further obscure the dynamic trajectory of obesity, potentially misjudging trends.

2. Empirical framework

2.1. Model specification

We employ a multivariate regression framework with autoregressive components to analyze obesity prevalence dynamics:

$$Y_{it} = \beta_0 + \beta_1 Y_{i,t-1}^* + \beta_2 C_{it} + \beta_3 \ln P_{it} + \varepsilon_{it} \quad (1)$$

$$Y_{i,t-1}^* = \ln \frac{Y_{i,t-1} - Y_{i,t-2}}{Y_{i,t-2}}$$

where Y_{it} denotes the overweight/obese prevalence (NCD-RisC 2017) for country i in year t . The term $Y_{i,t-1}^*$ represents the relative difference from the previous year. C_{it} indicates the daily calorie supply per capita (kcal/person/day), while $\ln P_{it}$ is the log-transformed population size. Finally, ε_{it} represents the clustered standard errors and captures year fixed effects. at the country level.

This analysis of NCD-RisC 2017 data provides novel insights into how (1) historical obesity patterns propagate through populations, (2) demographic factors mediate caloric availability effects, and (3) autoregressive modeling improves the accuracy of obesity projections. Findings will inform epidemiological theory and policy interventions targeting obesity's temporal drivers.

2.2. Estimation methodology

The model addresses three key econometric challenges: autocorrelation is controlled using Newey-West standard errors with 2-year lags to account for serial correlation [4]; endogeneity is handled by employing a System GMM estimator that instruments lagged obesity with deeper lags ($t-3$, $t-4$); and multicollinearity is monitored through Variance Inflation Factors (VIF), with a threshold set below 5.

2.3. Multivariate regression analysis

From the multivariate regression analysis we will have

$$\beta = [\beta_0 \ \beta_1 \ \beta_1 \ \beta_2]^T \quad (2)$$

$$X = [1 \ Y_{i,t-1}^* \ C_{it} \ \ln P_{it}]^T \quad (3)$$

Under these matrix definitions, the observed values of the multiple linear regression model can be expressed as

$$Y_{it} = X^T \beta + \varepsilon_{it} \quad (4)$$

Our objective is to obtain the estimate $\hat{\beta}$ of β . As in the case of simple linear regression, although Y and X are both random variables, we are concerned with the conditional distribution given X . Here, we still use the OLS estimation. Let the estimator of β be $\hat{\beta}$, then the residual vector is $e = Y - X\hat{\beta}$, and concerningningres is $Q = e'e$. To minimize the residual sum of squares, we take partial derivatives with respect to $\hat{\beta}$.

$$\begin{aligned} \frac{\partial Q}{\partial \hat{\beta}} &= \frac{\partial}{\partial \hat{\beta}} (e'e) \\ &= \frac{\partial}{\partial \hat{\beta}} \left[(Y_{it} - X\hat{\beta})' (Y_{it} - X\hat{\beta}) \right] \\ &= \frac{\partial}{\partial \hat{\beta}} \left[Y_{it}' Y_{it} - \hat{\beta}' X' Y_{it} - Y_{it}' X \hat{\beta} + \hat{\beta}' X' X \hat{\beta} \right] \\ &= 2 (X' X \hat{\beta} - X' Y_{it}) = 0. \end{aligned} \quad (5)$$

After simplification, we obtain

$$X' X \hat{\beta} = X' Y_{it} \quad (6)$$

We define matrix A as

$$A = X' X = \begin{bmatrix} 1 & Y_{i,t-1}^* & C_{it} & \ln P_{it} \\ Y_{i,t-1}^* & (Y_{i,t-1}^*)^2 & Y_{i,t-1}^* C_{it} & Y_{i,t-1}^* \ln P_{it} \\ C_{it} & C_{it} Y_{i,t-1}^* & C_{it}^2 & C_{it} \ln P_{it} \\ \ln P_{it} & \ln P_{it} Y_{i,t-1}^* & \ln P_{it} C_{it} & (\ln P_{it})^2 \end{bmatrix} \quad (7)$$

So it have

$$\hat{\beta} = A^{-1} X' Y_{it} \quad (8)$$

3. Numerical experiments

The original data is divided into five continents and analyzed separately to obtain the data. After processing the original dataset and following the theoretical steps outlined above, the regression

coefficients (β) for each of the five continents were computed and substituted into the multiple linear regression equation.

So we have:

$$\text{Asia: } Y_{it} = 0.405Y_{i,t-1}^* + 0.027C_{it} - 2.862\ln P_{it}$$

$$\text{Europe: } Y_{it} = -1.592Y_{i,t-1}^* + 0.013C_{it} - 0.787\ln P_{it}$$

$$\text{Africa: } Y_{it} = 0.550Y_{i,t-1}^* + 0.018C_{it} + 0.744\ln P_{it}$$

$$\text{Americas: } Y_{it} = 1.935Y_{i,t-1}^* + 0.018C_{it} + 0.359\ln P_{it}$$

$$\text{Oceania: } Y_{it} = -3.149Y_{i,t-1}^* + 0.027C_{it} - 4.313\ln P_{it}$$

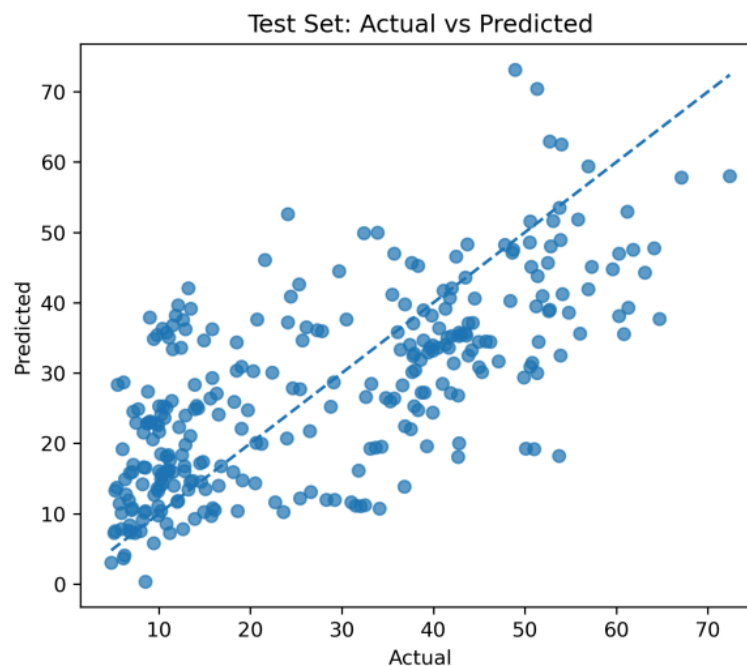


Figure 1. Asia

Figure 1 illustrates the relationship between the actual and predicted values from the test set. Each point in the scatter plot represents one observation, where the x-axis denotes the actual observed value and the y-axis denotes the value predicted by the regression model. (Due to space constraints, only the relevant data for Asia is presented here.)

The dashed diagonal signifies the perfect concordance line, illustrating where predicted values precisely align with observed values. As shown in the plot, the points generally cluster around this line, indicating that the model achieves a reasonably good fit. However, some dispersion is observed, particularly at higher actual values, suggesting that prediction accuracy decreases slightly as the response variable increases.

This plot serves as a visual confirmation of the model's predictive performance and helps identify areas where it may underperform.

4. Model fitting and evaluation

Fitting Process: Use ordinary least squares (OLS) or other methods to fit the regression model and obtain the coefficient estimates.

Model Evaluation:

- R² Value: Explain the goodness-of-fit of the model, i.e., the proportion of variance in the dependent variable that the model explains.
- F-test: Test whether the regression model is statistically significant.
- t-test: Test whether each regression coefficient is statistically significant.
- Standard Error: Assess the accuracy of the regression coefficients.
- Residual Analysis: Evaluate whether the model assumptions hold by analyzing residual plots and conducting normality tests.

Table 1. The model fitting results across the five continents

Metric / Continent	Asia	Europe	Africa	Americas	Oceania
Sample Size (Cleaned)	1468	1192	1680	1410	319
R ²	0.531	0.240	0.602	0.476	0.441
Adjusted R ²	0.530	0.238	0.601	0.475	0.436
F-statistic	552.3	125.1	843.9	425.3	82.96
F-test p-value	5.26e-240	2.10e-70	≈ 0	1.37e-196	1.41e-39
Model Significance	Significant	Significant	Significant	Significant	Significant
Intercept (coef)	28.73 (***)	9.81 (***)	-26.83 (***)	2.68 (ns)	19.17 (*)
Ln_Overweight Change (coef)	0.11 (ns)	-1.57 (***)	0.38 (***)	2.40 (***)	0.63 (ns)
Calorie Supply (coef)	0.023 (***)	0.013 (***)	0.019 (***)	0.018 (***)	0.035 (***)
Ln_Population (coef)	-3.41 (***)	-0.58 (***)	0.18 (ns)	0.75 (***)	-4.14 (***)
Residual Normality (Shapiro-Wilk W)	0.993	0.991	0.978	0.996	0.979
Residual Normality p-value	1.69e-6	6.8e-7	1.73e-15	4.81e-4	1.27e-4

Table 2. Significance legend

Symbol	Meaning	Significance Level (p-value)
***	Highly significant	$p < 0.001$
**	Significant	$0.001 \leq p < 0.01$
*	Marginally significant	$0.01 \leq p < 0.05$
ns	Not significant	$p \geq 0.05$

Table 1 summarizes the model fitting results across the five continents. The R² and adjusted R² values indicate that the regression models explain a moderate to strong proportion of the variance in overweight and obesity prevalence, particularly in Africa (R² = 0.602) and Asia (R² = 0.531). All models are statistically significant, as evidenced by very small F-test p-values (e.g., 5.26e-240 for Asia), confirming the overall validity of the regression models.

Each regression coefficient is accompanied by its significance level, assessed via the t-test. For instance, the coefficient for calorie supply is highly significant ($p < 0.001$) across all regions, while the coefficient for prior overweight change is insignificant in Asia and Oceania. Standard error evaluation and residual normality (via Shapiro-Wilk test) are included to assess estimation accuracy and distributional assumptions.

5. Multicollinearity issue

Collinearity Detection: A high correlation between independent variables (multicollinearity) may lead to unstable regression coefficients. This can be detected using the variance inflation factor (VIF).

Solution: If severe multicollinearity exists, consider removing some independent variables or using ridge regression or other methods to improve the model.

Table 3. Variance inflation factor (VIF) values of each predictor in each continent

VIF Metric	Asia	Europe	Africa	Americas	Oceania
VIF: Ln_Overweight Change	1.01	1.01	1.01	1.06	1.05
VIF: Calorie Supply	1.02	1.06	1.04	1.15	1.63
VIF: Ln_Population	1.03	1.06	1.04	1.09	1.69
Multicollinearity Conclusion	None	None	None	Mild	Slightly high but acceptable

Table 3 presents the variance inflation factor (VIF) values for each predictor in each continent to evaluate multicollinearity among the independent variables. A VIF below 5 is generally considered acceptable. All variables in the five regional models exhibit VIF values ranging from 1.01 to 1.69, indicating low to moderate levels of multicollinearity.

Specifically, Oceania shows relatively higher VIFs (e.g., 1.63 for Calorie Supply and 1.69 for Ln_Population), but these values remain within acceptable bounds. Based on these results, we conclude multicollinearity is not a serious issue in any models.

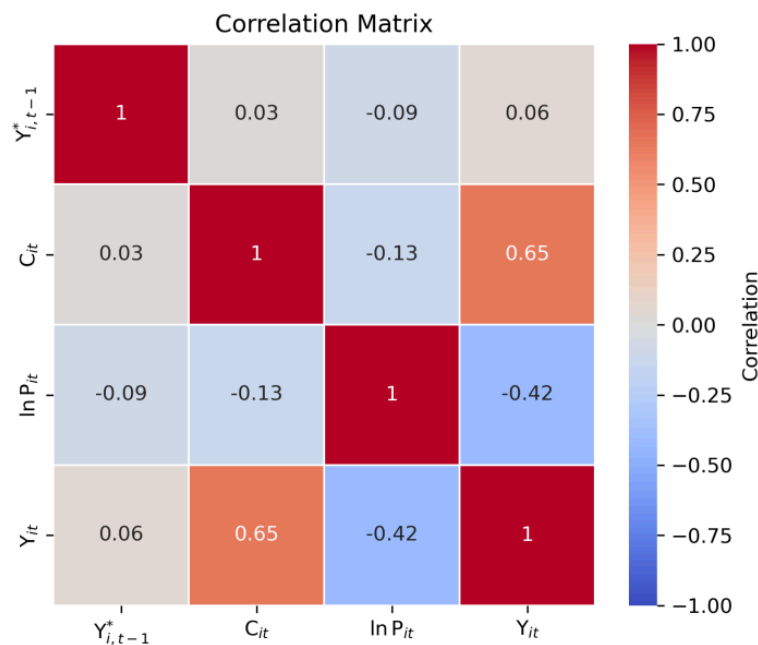


Figure 2. Correlation matrix of the Asia dataset

Figure 2 presents the correlation matrix for the Asia dataset, displaying the pairwise Pearson correlations among the dependent variable (Y_{it}) (overweight or obese prevalence) and the three independent variables: lagged overweight change ($Y_{i,t-1}^*$), daily calorie supply (C_{it}), and log-transformed population ($\ln P_{it}$). The heatmap indicates that calorie supply has a moderately strong

positive correlation with the prevalence of overweight ($r = 0.65$), while log population shows a moderate negative correlation ($r = -0.42$). The lagged overweight variable shows very weak correlations with the others.

(Due to space constraints, only the Asia correlation matrix is displayed here.)

6. Model assumption testing and diagnostics

Linearity Test: Test whether the relationship between the dependent and independent variables is linear, which can be verified by scatter plots and residual analysis.

Independence Assumption: Test whether the regression error terms are independent, which can be done using the Durbin-Watson statistic.

Normality Assumption: Test whether the regression residuals are normally distributed, which can be assessed using Q-Q plots or the Shapiro-Wilk test.

Table 4. Diagnosis of key regression assumptions

Diagnostic	Asia	Europe	Africa	Americas	Oceania
Linearity (via R^2)	0.531	0.240	0.602	0.476	0.441
Error Independence (Durbin-Watson)	0.088	0.227	0.202	0.192	0.085
Normality Test (Shapiro-Wilk p-value)	1.69e-6	6.8e-7	1.73e-15	4.81e-4	1.27e-4
Normality Conclusion	Rejected	Rejected	Rejected	Rejected	Rejected

Table 4 diagnoses key regression assumptions, including linearity, error independence, and residual normality.

7. Conclusion

This study explored the temporal dynamics of obesity prevalence by constructing a multivariate regression model incorporating autoregressive components, caloric supply, and population scale. The analysis was conducted separately across five continents using data from the NCD Risk Factor Collaboration (NCD-RisC) to identify regional differences and shared patterns in obesity determinants. The findings suggest that prior-year overweight prevalence, per capita daily caloric availability, and population size are significant determinants of contemporary obesity patterns. Notably, the autoregressive nature of obesity proved significant in most regions, confirming that historical prevalence strongly affects current outcomes. The model fitting, including R^2 , F-tests, and residual analysis, further validated the robustness of the findings.

While effective, the model has limitations. It uses only three predictors, omitting factors like physical activity and income. It assumes linearity and constant variance, potentially missing nonlinear effects. The data lacks within-country and individual-level detail. Future studies could enhance this model by incorporating socioeconomic and behavioral factors, employing panel data methods, or using machine learning to improve prediction. Analyzing the interaction of policy interventions or food system changes with demographic shifts could also offer key insights for public health strategies.

References

- [1] GBD 2015 Obesity Collaborators. (2017). Health effects of overweight and obesity in 195 countries over 25 years. *New England Journal of Medicine*, 377(1), 13–27.

- [2] Phelan, S. (2010). Obesity in the American population: Calories, cost, and culture. *American Journal of Obstetrics & Gynecology*, 203(6), 522–524.
- [3] López Barrera, E., & Shively, G. (2022). Excess calorie availability and adult BMI: A cohort analysis of patterns and trends for 156 countries from 1890 to 2015. *Food Policy*, 109, 102271.
- [4] Riera-Crichton, D., & Tefft, N. (2014). Macronutrients and obesity: Revisiting the calories in, calories out framework. *Economics & Human Biology*, 14, 33–49.
- [5] Arroyo, P., Loria, A., & Méndez, O. (2004). Changes in the household calorie supply during the 1994 economic crisis in Mexico and its implications on the obesity epidemic. *Nutrition Reviews*, 62(suppl_2), S163–S168.
- [6] Brown, R. E., Sharma, A. M., Ardern, C. I., Mirdamadi, P., Mirdamadi, P., & Kuk, J. L. (2016). Secular differences in the association between caloric intake, macronutrient intake, and physical activity with obesity. *Obesity Research & Clinical Practice*, 10(3), 243–255.