Unveiling Obesity Dynamics: Interactions Between Lifestyle and Genetic Factors

Interplay of Genetic Predispositions, Physical Activity, Alcohol Consumption, and Dietary Habits in Obesity Outcomes

Ilseop Lee, Ramil Mammadov, Tursunai Turumbekova, Yirang Liu

Abstract

Understanding the complex interactions between genetic predispositions and lifestyle factors is essential for developing effective and personalized obesity prevention strategies. This study examines (1) the interaction between family history of overweight and physical activity frequency affecting Body Mass Index (BMI), and (2) the combined effects of meal frequency and alcohol consumption on obesity levels. Using a dataset from Mexico, Peru, and Colombia, multiple linear regression (MLR) and ordinal logistic regression models were applied. The results reveal that physical activity significantly reduces BMI for individuals with a family history of overweight, highlighting the importance of personalized weight management strategies. Additionally, frequent meals combined with alcohol consumption were strongly associated with higher obesity levels, particularly severe obesity, offering actionable insights for targeted interventions.

Introduction

Obesity, a global health crisis, has nearly tripled since 1975, with over 650 million adults classified as obese in 2016 (WHO). This epidemic imposes significant economic burdens, driven by rising healthcare costs, and underscores the importance of lifestyle factors, such as diet and physical activity, as primary contributors to weight gain. Genetic predispositions, like a family history of overweight, often interact with these behaviors, while factors such as alcohol consumption may further exacerbate their effects.

Given the complex interplay of genetic and lifestyle factors, this study investigates two key aspects of obesity: (1) the interaction between family history of overweight and physical activity frequency on Body Mass Index (BMI), and (2) the interaction between meal frequency and alcohol consumption on obesity levels. By addressing these questions, the study aims to uncover insights that can inform personalized obesity prevention and intervention strategies tailored to individual genetic and lifestyle factors.

Methods

The Obeity Levels dataset (Kaggle) consists of 2111 records with no missing values from individuals in Mexico, Peru, and Colombia, representing demographic, dietary, and lifestyle factors (Original research source). Of these, 23% were collected via a web platform, while 77% were synthetically generated using the SMOTE algorithm.

Exploratory Data Analysis (EDA) was performed for each research questions to summarize variable distributions and relationships. Continuous variables were visualized with histograms and boxplots, while categorical variables (e.g., family history of overweight, alcohol consumption frequency, obesity levels) were summarized using bar charts. Interaction effects were analyzed using scatterplots and interaction plots.

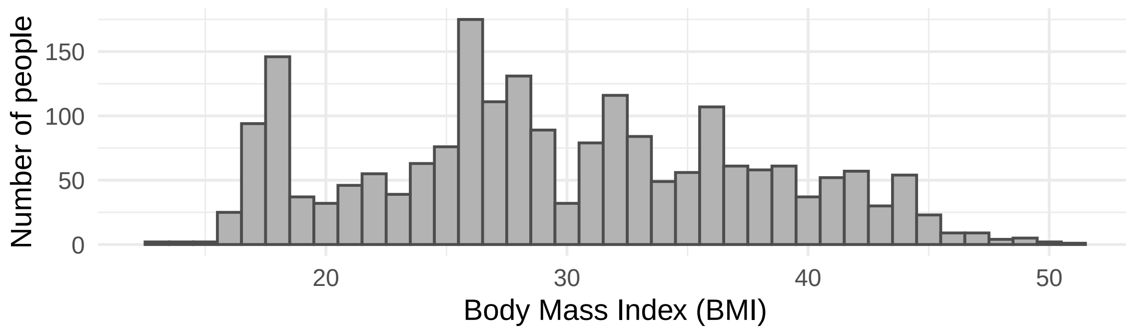
To address the research objectives, two models were developed. For question 1, a multiple linear regression (MLR) model was applied to examine the relationship between Body Mass Index (BMI) and three key predictors: family history of overweight, physical activity frequency, and their interaction term. Model diagnostics included residual analysis to assess linearity, normality, and homoscedasticity, as well as Variance Inflation Factor (VIF) calculations to evaluate multicollinearity. For question 2, an ordinal logistic regression model was applied to analyze obesity levels, focusing on the interaction between meal frequency and alcohol consumption. Variables controlling for confounding effects were included. Diagnostics included an assessment of the proportional odds assumption, a confusion matrix, and an analysis of predicted probabilities. During model development, variables with a Variance Inflation Factor (VIF) below 3 were retained to address multicollinearity. To improve analytical clarity, the target variable obesity levels for research question 2, initially categorized into seven groups, was consolidated into three—Obese (972 records), Overweight (580 records), and Underweight/Normal (559 records). Similarly, alcohol consumption levels were re-categorized into “Yes” (1472 records) and “No” (639 records) to address small sample sizes and ensure robust analysis.

Results

Question 1. How does the interaction between family history of overweight and the frequency of physical activity influence an individual’s BMI?

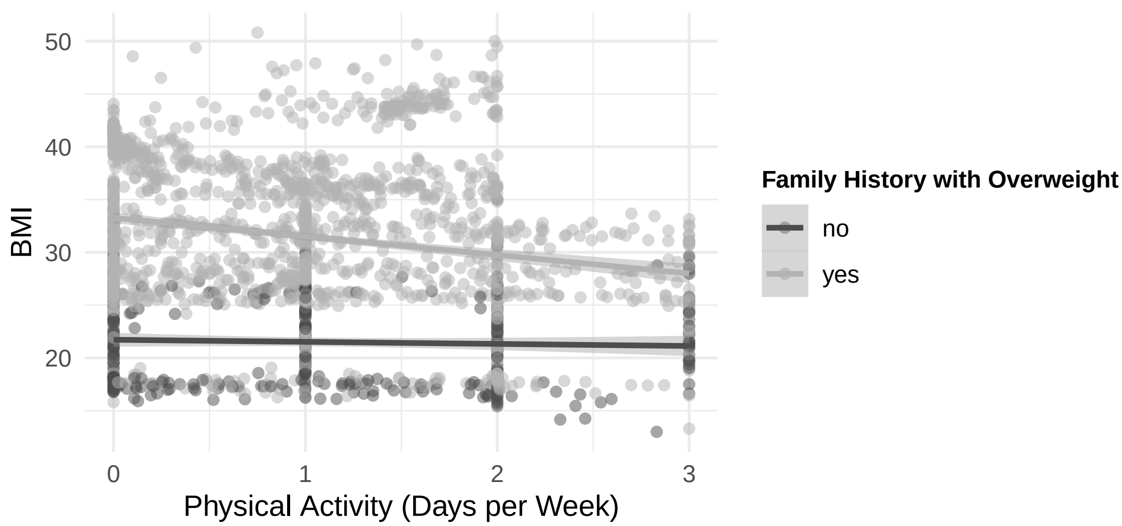
The distribution of BMI is typically right-skewed as shown on figure 1, as most individuals in the dataset are overweight or obese (73%). The median BMI is around 29.7, which is consistent with a population that has a higher proportion of overweight and obese individuals. The distribution of BMI shows a peak around 25–30, which according to National Heart, Lung, and Blood Institute is categorized as overweight (National Heart, Lung, and Blood Institute, n.d.). This supports the general notion that the population has a significant proportion of individuals at risk for obesity-related health conditions. The histogram illustrates the distribution of BMI, highlighting the concentration of individuals with BMI values above 25.

Figure 1. BMI Distribution



On the figure 2, the scatter plot examines the interaction between physical activity and family history on BMI. The results indicate that individuals with a family history of overweight (light gray) show a stronger decrease in BMI as physical activity increases, evidenced by the steeper negative slope of the light gray line. In contrast, individuals without a family history of overweight (dark gray line) show a relatively flat decrease in BMI with increased physical activity. This suggests that while genetic factors influence BMI, individuals with a family history of overweight might respond more significantly to increased physical activity in terms of BMI reduction.

Figure 2. Physical Activity and Family History on BMI



Model : Multiple Linear Regression - BMI

The multiple linear regression (MLR) model was fitted to mean BMI, including the main effects of family history and physical activity frequency, as well as their interaction term.

Table 1. Results of the Multiple Liner Regression - BMI

Predictor	Estimate	Std. Error	t-value	p-value
Intercept	21.71	0.55	39.57	<0.001
Family History of Overweight (Yes)	11.55	0.61	19.04	<0.001
Physical Activity Frequency	-0.20	0.38	-0.52	0.604
Interaction (Family History & Physical Activity Frequency)	-1.56	0.43	-3.64	<0.001

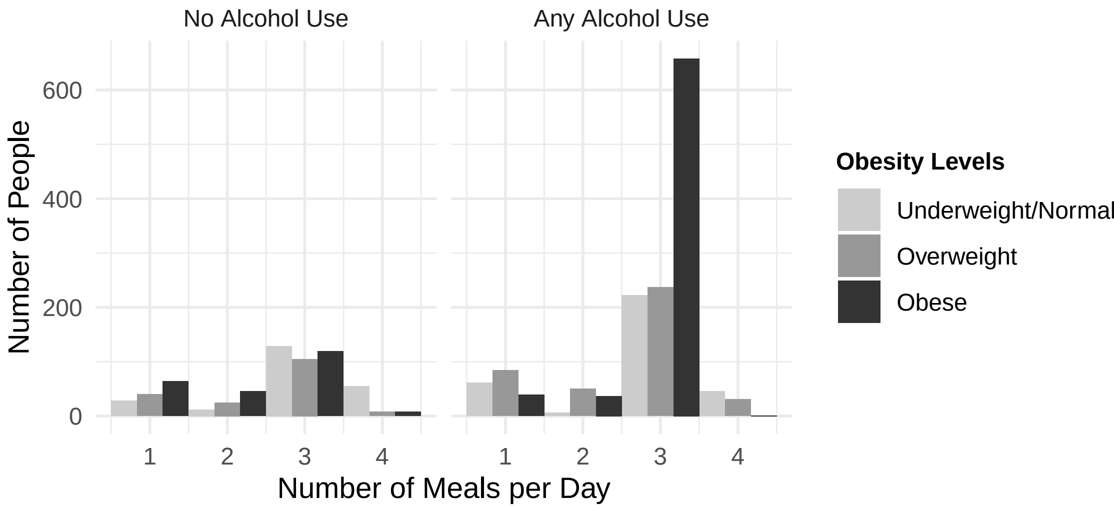
The mean BMI for individuals with no family history of overweight and no physical activity was 21.72. The interaction between family history of overweight and physical activity frequency indicates that the effect of physical activity on BMI varies across groups. Physical activity is significantly more effective in reducing BMI for individuals with a family history of overweight. For this group, physical activity leads to a substantial and statistically significant reduction in BMI ($p < 0.001$), reducing it by an additional 1.56 units compared to those without a family history. Conversely, for individuals without a family history of overweight, the estimated effect of physical activity frequency on BMI is a decrease of 0.20 units; however, this effect is not statistically significant ($p = 0.604$) . This suggests that targeting physical activity interventions could be especially impactful for individuals with genetic or family predispositions to overweight.

The model explained 26.1% of the variability in BMI (Multiple R-squared = 0.261), indicating that the predictors included in the model account for a moderate proportion of the variability in BMI. The statistically significant F-statistic ($F = 248.1$, $p < 2.2e-16$) indicates that at least one predictor or the interaction term is significantly associated with BMI.

Question 2: How do meal frequency and alcohol consumption interaction influence obesity levels?

Figure 3 illustrates the distribution of obesity levels and meal frequency across two groups within the alcohol consumption category: "No Alcohol Use" and "Any Alcohol Use". The analysis reveals that meal frequency significantly influences obesity levels, with consuming three meals per day being the most common pattern in both group (262 records for No Alcohol Use, 942 records for Any Alcohol Use). This suggests a relationship between meal frequency and obesity. Furthermore, "Obese" category has the highest number of individuals in the "Any Alcohol Use" group (735 records), indicating a potential association between alcohol consumption and higher obesity levels.

Figure 3. Impact of Alcohol Consumption and Meal Frequency on Obesity levels



Model : Ordinal Logistic Regression model - Obesity Levels

Table 2. Impact of Meal Frequency and Alcohol Consumption on Obesity Outcomes

Variable	Odds Ratio	Estimate	Std Error	T-Value	P-Value
Number of Meals per Day	0.46	-0.77	0.09	-8.30	<0.001
Alcohol Consumption (Yes)	0.10	-2.34	0.32	-7.39	<0.001
Interaction between Number of Meals per Day & Alcohol Consumption	2.86	1.05	0.12	9.12	<0.001
Underweight/Normal vs. Overweight	15.24	2.72	0.39	6.91	<0.001
Overweight vs. Obese	66.96	4.20	0.40	10.47	<0.001

According to the ordinal logistic regression model results, the interaction between meals frequency and alcohol consumption has a significant impact on increasing obesity levels (T-Value: 9.12, P-Value: < 0.001). This suggests that individuals who consume alcohol and have higher meals per day meal are 2.9 (Odds Ratio) times more likely to fall into higher obesity categories. However, individually, larger number of meals per day is associated with a lower likelihood of falling into higher obesity categories (Odds Ratio of 0.46, P-Value: < 0.001). Similarly, alcohol consumption independently decreases the likelihood of higher obesity levels (Odds Ratio of 0.10). However, the interaction effect indicates that this independent impact can be reversed or overridden when meal frequency and alcohol consumption are combined.

Table 3. Ordinal Logistic Regression Confusion Matrix: Actual vs. Predicted Obesity Categories

Predicted/Actual	**Underweight/Normal**	**Overweight**	**Obese**
Underweight/Normal	320	152	91
Overweight	49	56	39
Obese	190	372	842

Overall, the model excels at detecting obesity but faces challenges in accurately classifying individuals as Underweight/Normal or Overweight. As shown on Table 4, the model performed well in identifying individuals with obesity, correctly detecting 86.6% of cases. However, it struggled with classifying others as obese, with a specificity of only 50.7%. For the Underweight/Normal category, the model identified just over half of these cases correctly (achieved a moderate sensitivity of 57.3%), while showing fewer misclassifications (specificity was higher at 84.3%). In contrast, the model had significant difficulty with the Overweight category, correctly identifying only 9.7% of cases, despite of rare misclassification of others as overweight (specificity of 94.3%).

The ordinal logistic regression model showed moderate overall performance with 58% accuracy, however, it was chosen over the multinomial model for its ability to preserve the ordered nature of obesity weight status categories. While the multinomial model achieved higher accuracy (0.61) and kappa (0.37), it does not account for the inherent order, reducing its interpretability. The ordinal model performed competitively, with an accuracy of 0.58 and a kappa of 0.30. It showed strong sensitivity for the "Obese" category (0.85) and reasonable specificity across categories, though it struggled with the "Overweight" category due to lower sensitivity and overlapping probabilities.

Conclusion

This study underscores the multifaceted nature of obesity, highlighting the critical interactions between genetic predisposition and lifestyle behaviors. The findings reveal that family history of overweight is a significant predictor of BMI, with physical activity playing a vital role in mitigating its impact. Specifically, individuals with a family history of overweight benefit significantly from increased physical activity, demonstrating the value of tailored interventions that account for genetic factors. However, for those without such a family history, the relationship between physical activity and BMI is less pronounced, suggesting that other factors may drive weight outcomes in this group.

Additionally, the interaction between meal frequency and alcohol consumption was found to significantly influence obesity levels. While consuming fewer meals per day is associated with maintaining a normal weight, frequent meals combined with alcohol consumption substantially increase the risk of obesity. This highlights the compounded effects of certain lifestyle behaviors and the need for integrated strategies that address multiple contributing factors simultaneously.

The findings emphasize the importance of personalized approaches to obesity prevention and management. Public health strategies should focus on promoting physical activity, especially among those with a family history of overweight, and encouraging moderation in meal frequency and alcohol consumption to mitigate the risk of obesity. However, the limitations of using BMI as the sole measure of obesity, along with the synthetic nature of the dataset, underscore the need for further research. Future studies should include additional measures of body composition, such as body fat percentage or waist-to-hip ratio, to offer a more comprehensive understanding of obesity. Furthermore, access to less synthetically generated datasets could facilitate the development of granular classification criteria for obesity levels and alcohol consumption categories. This approach would enable more tailored and effective solutions for obesity prevention and management.

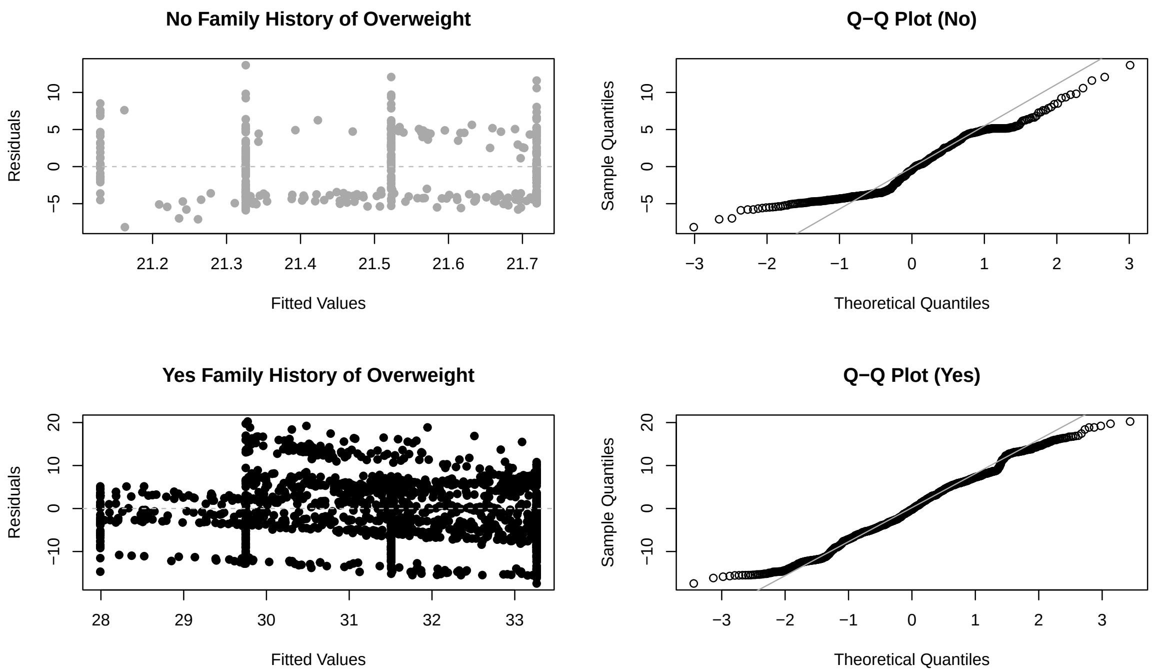
References

1. Johnson, M. (2022). Behavioral influences on obesity levels. *Obesity Research Review, 45*(4), 290–307.
2. Kähler, F., et al. (2021). Effects of alcohol consumption on obesity. *Journal of Nutrition, 38*(2), 151–159.
3. National Heart, Lung, and Blood Institute. (n.d.). Assessing your weight and health risk. Retrieved from [https://www.nhlbi.nih.gov/health/educational/lose_wt/bmitools.htm%5D(https://www.nhlbi.nih.gov/health/educational/lose_wt/bmitools.htm#:~:text=People%20who%20are%20overweight%20(BMI,in%20relation%20to%20their%20height.)
4. Obesity Medicine Association. (2022, August 2). Alcohol and obesity. Retrieved from https://obesitymedicine.org/blog/alcohol-and-obesity
5. Rauner, A., Mess, F., & Woll, A. (2013). The relationship between physical activity, physical fitness, and overweight in adolescents: A systematic review of studies published in or after 2000. *BMC Pediatrics, 13*(19). https://doi.org/10.1186/1471-2431-13-19
6. Smith, J., et al. (2021). The role of genetic and lifestyle factors in obesity. *Journal of Public Health, 12*(3), 112–119.
7. World Health Organization (WHO). (2020). Obesity and overweight. Retrieved from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

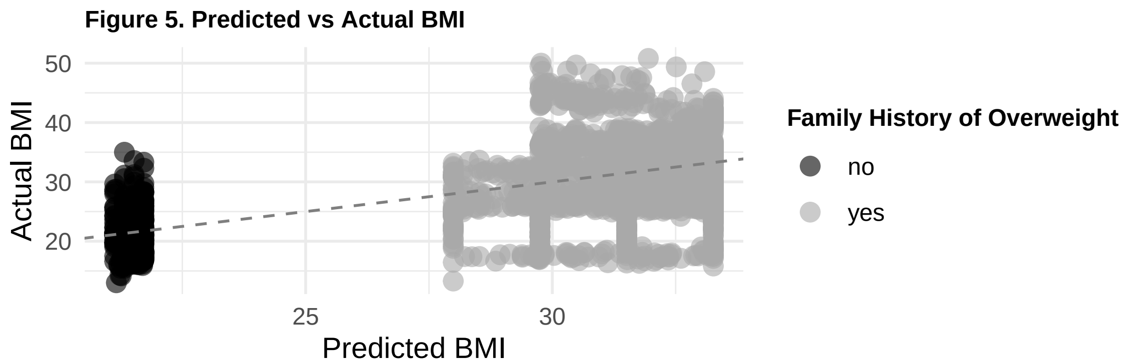
Appendices:

Model assessment for MLR in research question 1:

Figure 4. MLR model assessment plots



By separating Family History of Overweight into two groups, we observe that the residual plots do not display clear clustering, suggesting the absence of systematic patterns that would indicate model misspecification. The residuals are evenly distributed around the zero line for both groups, demonstrating reasonable homoscedasticity. However, in the Q-Q plots, the "No" group shows significant deviations from the theoretical quantile line in the tails, indicating a lack of normality in the residuals for this subgroup. In contrast, the "Yes" group exhibits better alignment with the theoretical quantile line, although slight deviations in the tails persist. The "Yes" group also shows slightly more variability and deviation in the residual plot compared to the "No" group, potentially pointing to heteroscedasticity or the need for additional predictors to improve the model fit for this subgroup. Overall, while the residual plots for both groups suggest that the model fits reasonably well without major biases, the poor fit in the Q-Q plot for the "No" group highlights concerns with normality assumptions for this subgroup. Refining the model, addressing potential outliers, or applying transformations may improve the overall model performance, particularly for the "No" group.



The graph indicates that the interaction between family history of overweight and frequency of physical activity does not effectively explain variations in BMI. The predicted BMI values are concentrated within a narrow range (20–30), while the actual BMI values exhibit much greater variability.

Table 4. Multicollinearity Assessment Summary 1

Predictor	**Family History of Overweight**	**Physical Activity Frequency**	**Interaction: Family History of Overweight & Physical Activity Frequency**
VIF	2.44	4.62	5.79

The Variance Inflation Factor assessment for the predictors in the model indicates acceptable levels of multicollinearity.

Model Assessment for Question 2

Table 5: Predicted Probabilities and Classes from Ordinal Logistic Regression

Predicted.Class	Fit.Prob.Underweight.Normal	Fit.Prob.Overweight	Fit.Prob.Obese
Underweight/Normal	0.758	0.174	0.068
Underweight/Normal	0.641	0.246	0.113
Underweight/Normal	0.534	0.300	0.166
Overweight	0.326	0.354	0.320
Underweight/Normal	0.688	0.218	0.094
Obese	0.130	0.266	0.604

The proportional odds assumption, a cornerstone of the ordinal logistic regression model, posits that the odds ratios between adjacent outcome categories remain consistent. This assumption aligns with the natural hierarchy of obesity levels (Underweight/Normal, Overweight, Obese), and the results in Table 5 support this assumption. Table 5 shows the predicted probabilities and classifications generated by the ordinal logistic regression model. For cases classified as Underweight/Normal, the model assigns the highest probabilities to this category (e.g., 0.758, 0.641, 0.534, and 0.688), with progressively lower probabilities for Overweight and Obese. Similarly, for the case classified as Obese, the highest probability is assigned to the Obese category (0.604), while the probabilities for Underweight/Normal and Overweight are lower. The case classified as Overweight (with probabilities of 0.326, 0.354, and 0.320) demonstrates more evenly distributed probabilities, reflecting the overlap often observed in adjacent categories. This pattern of predictions supports the validity of the proportional odds assumption. By aligning predicted probabilities with the ordinal structure of the outcome, the model demonstrates improved interpretability and consistency, strengthening its applicability to hierarchical data like obesity levels.

The ordinal logistic regression model is preferred for analyzing obesity levels due to its better performance, interpretability, and alignment with the ordinal nature of the outcome. It achieves higher overall accuracy (61.39% vs. 57.7%) and kappa (0.3694 vs. 0.3001) compared to the multinomial model. Sensitivity for the Overweight category improves significantly (25.86% vs. 9.66%), while maintaining high specificity for Underweight/Normal (87.50%) and Overweight (91.44%). The Akaike Information Criterion (AIC) also favors the ordinal model (3675.93 vs. 3955.49), reflecting its efficiency.

These results underscore the ordinal model's ability to respect the hierarchical structure of obesity levels and maintain proportional odds across categories. Multicollinearity diagnostics show that while the interaction term (Meal Frequency × Alcohol Consumption) has VIF values exceeding 10, all other variables remain well below the threshold of 5, indicating no significant multicollinearity issues.

Table 7: Multicollinearity Assessment Summary 2

Predictor	Meal Frequency	Alcohol Consumption	Age	High-Fat Food Consumption	Water Consumption	Vegetable Intake	Calorie Monitoring	Interaction: Meal Frequency × Alcohol Consumption
VIF	2.93	10.94	1.03	1.04	1.04	1.02	1.04	13.32

* A VIF value above 10 suggests multicollinearity, interaction terms and their related variables, values above 10 are often acceptable