Aaron Guan
Felix Peng
Sarah Wu

August 15th, 2025

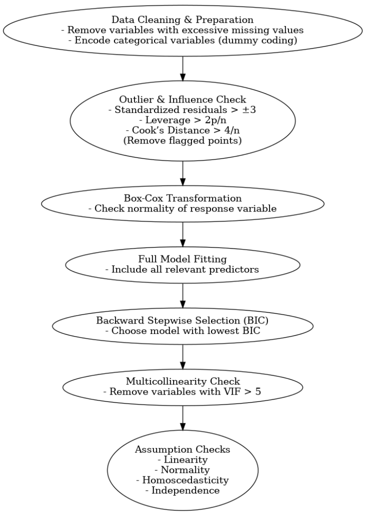
The Influence of Lifestyle and Mental Health on BMI

Introduction:

Body Mass Index (BMI) is a widely used metric for assessing obesity in adults, serving as a practical indicator of body fatness in population health research. Elevated BMI is strongly associated with chronic health conditions, including cardiovascular disease, Type 2 diabetes, and hypertension (NHLBI, accessed July 17 2025). Since the 1960s, obesity rates in the United States have increased from ~13% to ~ 43%, generating their most pressing public health concern today. Prior studies have demonstrated that lifestyle factors such as shorter sleep durations are strongly associated with concurrent and future obesity (Patel, S.R., Hu, F.B., 2012), and smokers with lower socioeconomic status have been associated with increased weight gain and obesity (Chiolero, A., et al., 2007). Increased sedentary time was also associated with higher body-fat percentage and mental health issues (Feng, Yue., et al., 2024). While prior research has emphasized the role of lifestyle choices such as physical activity, diet, alcohol consumption, and smoking, it remains unclear how they interact with mental health to predict BMI outcomes. Understanding these complex relationships is essential for developing effective and equitable public health strategies. Our preliminary analysis of the NHANES survey data has identified multiple predictors with correlational relationships to BMI. Thus, our research question aims to investigate **whether lifestyle and mental health factors can be used to predict BMI in adults in the United States.**

Methods

To investigate how mental health and lifestyle factors are related to Body Mass Index (BMI), we began by using data from the National Health and Nutrition Examination Survey (NHANES). From the full dataset, exploratory data analysis (EDA) is performed by reviewing the summary statistics for each variable. Then, we will identify all variables relevant to lifestyle habits (e.g., physical activity, smoking, alcohol use, sleep patterns) and mental health (e.g., depression scores, emotional well-being). Variables unrelated to these topics are excluded from further consideration.


Data Cleaning & Preparation
- Remove variables with excessive missing values
- Encode categorical variables (dummy coding)
Outlier & Influence Check
- Standardized residuals > ±3
- Leverage > 2p/n
- Cook's Distance > 4/n
(Remove flagged points)
Box-Cox Transformation
- Check normality of response variable
Full Model Fitting
- Include all relevant predictors
Backward Stepwise Selection (BIC)
- Choose model with lowest BIC
Multicollinearity Check
- Remove variables with VIF > 5
Assumption Checks
- Linearity
- Normality
- Homoscedasticity
- Independence

Before modeling, the data will be assessed for completeness. Variables with excessive missing values will be removed to prevent large reductions in sample size. This ensures that we retain a dataset that contains the most relevant predictors while maintaining adequate observations for reliable model estimation.

Next, we coded categorical variables into binary or dummy variables where appropriate, to ensure compatibility with linear regression. All "Yes/No" survey questions were re-coded as 1 for "Yes" and 0 for "No." Multi-category variables are converted into dummy variables, with one category set as the reference. For example, the Depressed variable, which originally had three categories ("None," "Several," and "Most"), are created into two dummy variables: depressedSeveral and depressedMost. The first dummy variable was coded as 1 if the respondent reported "Several" days of depression in the past two weeks, and 0 otherwise. The second was coded as 1 if the respondent reported "Most" days, and 0 otherwise. This coding allows the regression coefficients for depressedSeveral and depressedMost to be interpreted relative to individuals reporting no depression. For marijuana use, we merged the Marijuana and RegularMarij variables into a single three-level variable named MarijFrequency: which included "Never", "Sometimes", and "Frequently". Continuous predictors such as hours of sleep per night and age were kept in their numeric form.

Any outliers, leverage points, or influential points will then be removed from the data. Outliers will be identified using standardized residuals greater than ±4, high-leverage points will be flagged using hat values exceeding 2(p)/n, where p is the number of predictors and n is the sample size. Influential points will be detected using Cook's distance values greater than 4/n.

Before building the main model, we assessed whether BMI needed transformation to meet the normality assumption and whether we needed to apply a box-cox transformation. The Box-Cox method identifies an optimal transformation parameter to improve normality and stabilize variance, which supports the regression assumptions.

We will then fit the full model and perform backward stepwise selection using the Bayesian Information Criterion (BIC) as the selection metric. BIC balances model fit and complexity by penalizing models with more predictors more strongly than AIC, helping to prevent overfitting.

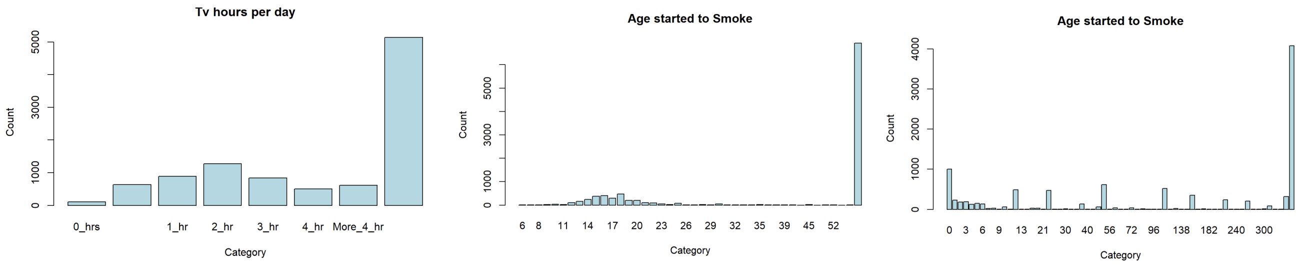
Finally, we will calculate Variance Inflation Factors (VIF) to check for multicollinearity. Variables with VIF values above the threshold of 5 will be removed to ensure stable and interpretable regression estimates.

Results:

We first decided to use every lifestyle variable and some health variables (Depressed, LittleInterest, SleepHrsNight, SleepTrouble) which we think are relevant to our study. After performing exploratory data analysis (EDA). We removed some variables with a significant amount of missing values. Specifically, PhysActiveDays, TVHrsDay, CompHrsDay, AlcoholDay, AlcoholYear, SmokeAge, AgeRegMarij, AgeRegMarij. Examples are shown in **Figure 1,** the last bar for each of the plots represents the number of occurrences of missing values. We also removed SexAge, to avoid excluding individuals who have never had sex. With the remaining response variables, we clean the data by dropping rows with missing entries and duplicate data. After the data cleaning process we are left with 2873 valid survey responses and 20 predictor variables.

Figure 1

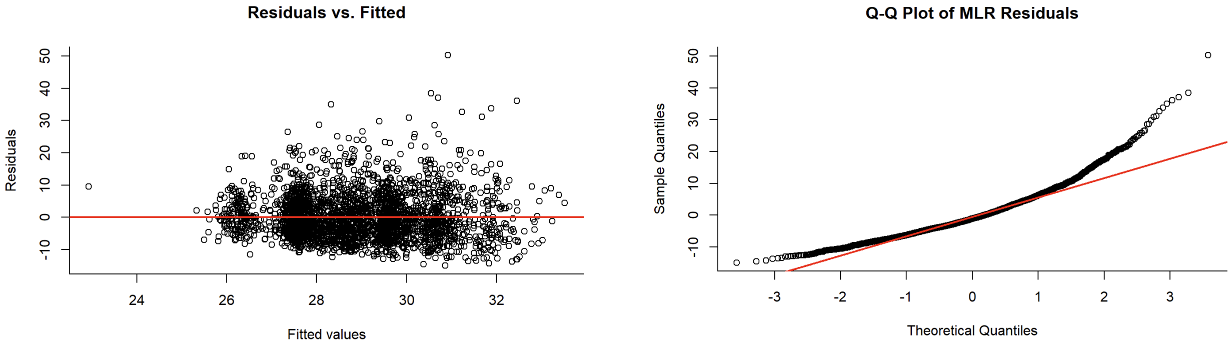
Barplot for some explanatory variables



We start with building an initial model which contains all the 20 predictor variables, the resulting residual vs. fitted value plot and qq-plot is shown in **Figure 2**. We noticed some of the problems for this initial model. The Q-Q plot of BMI in its original scale showed clear deviations from the reference line, indicating that BMI did not follow a normal distribution **(Figure 2)**. Moreover, on the plot of residuals vs. fitted value plot, the occurrence of residuals is slightly higher above the horizontal line at 0 across the range of predicted value. Therefore the initial model is not applicable.

Figure 2

Residuals plot and qq-plot for the initial model



To improve model accuracy and meet regression assumptions, we identified and removed influential observations based on standardized residuals (>|3|), high leverage points (hat values > 2p/n), and influential points (Cook's distance > 4/n). After applying these thresholds, it removed a total of 330 outliers/leverage/influential points from the dataset. Furthermore, after removing outliers and leverage points, we dropped the predictors SexOrientationBi and SexOrientationHomo due to zero or near-zero variance.

We selected the final model using backward stepwise selection with BIC, which was preferred over AIC because our large sample size warranted a stronger penalty for additional predictors, and BIC penalised more complex models under large data size. This approach reduced the risk of overfitting and ensured that we only kept the most essential variables. The final model included three significant predictors, SleepTroubleDummy, PhysActiveDummy, and SmokeRegular. It showed no significant loss of fit compared with the larger candidate model (partial-F p = 0.14).

Once the final model was selected, we evaluated the four key assumptions of linear regression:

1. **Linearity** – We examined the Residuals vs. Fitted Value plot from **Figure 3** and Residual vs predictors (**Figure 4**). These plots show that residuals are randomly distributed with no curve nor systematic pattern, and scattered around the horizontal line at 0 across the range of predicted values. Suggesting the assumption of linearity is met.
2. **Normality of residuals** – After the Box-Cox Transformation, the qq-plot in **Figure 3** shows that all data sticks closely to the reference line and no appearance of skewness. Suggesting the assumption is met.
3. **Homoscedasticity** – By examining the Residual vs. Fitted value plot in **Figure 3,** the distribution of residuals is not affected by the change in BMI, therefore the assumption of Homoscedasticity is met.
4. **Independence of errors** – assumed reasonable given the cross-sectional NHANES design; also supported by the absence of patterns in residual plots.

Since all assumptions are met, this model is applicable. The summary of the model is shown in **Table 1**.

Figure 3

Residual plot and qq-plot for the final model

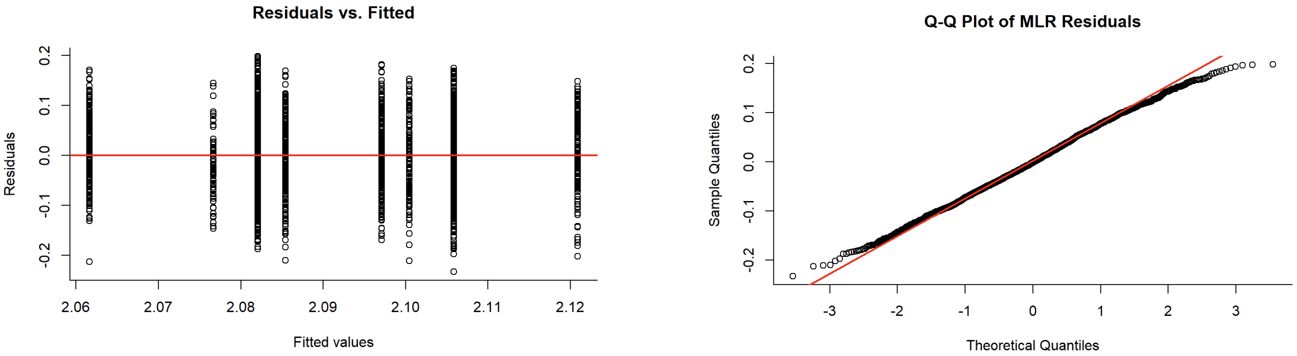
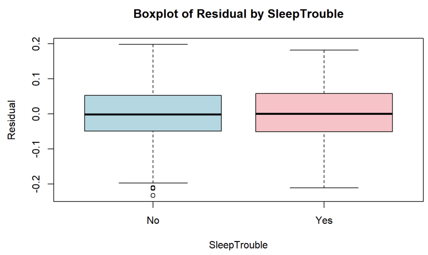
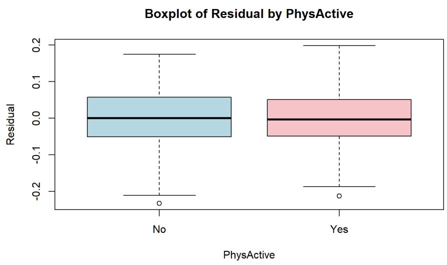


Figure 4

Residuals vs. Predictor Variables





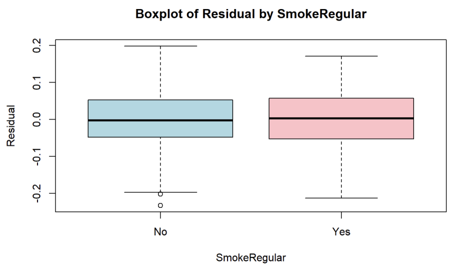


Table 1

Regression Model Estimates

	Coefficients Estimates	P-value	VIF
Intercept	2.105841	< 2e-16	NA
SleepTroubleDummy	0.015038	1.57e-5	1.008970
PhysActiveDummy	-0.023792	1.43e-15	1.024063
SmokeRegular	-0.020423	6.85e-9	1.030598

The final model:

$\hat{y}_{bc} = 2.105841 + 0.015038x_1 - 0.023792x_2 - 0.020423x_3$

Where,

$\hat{y}_{bc}$ is the BMI under Box-Cox transformation

x_1 is the dummy variable indicates whether the participants have sleep trouble

x_2 is the dummy variable indicates whether the participants have is physically active

x_3 is the dummy variable indicates whether the participants currently a smoker

To interpret these numbers we use the Box-Cox back transformation BMI = $(\lambda \cdot y_{bc}+1)^{\wedge}(1/\lambda)$, then our model indicates that people having SleepTrouble will be estimated to have 0.4 kg/m^2 higher BMI, people who are physically active will be estimated to have 0.64 kg/m^2 lower BMI, and people who Smoke regularly have 1.56 kg/m^2 lower BMI. All the predictors have p-value less than 0.05, which means they are all statistically significant. Furthermore no predictor had a VIF above 5, which indicated low redundancy between variables.

Conclusion and limitations:

The final regression model revealed a statistically significant relationship between variables SleepTroubleDummy, PhysActiveDummy, and SmokeRegular to Body Mass Index (BMI). Specifically, the analysis showed that higher physical activity and regular smoking were associated with lower BMI, both having negative coefficient estimates**.** Contrarily, individuals with sleep trouble were associated with higher BMI, indicated by the positive coefficient estimate. All three results are statistically significant, indicated by the low P-values (Table 1). These findings are consistent with previous literature cited regarding sleep duration and physical activity; however, results regarding smoking are surprising compared to the literature cited. Our results address the research question by quantifying the influence of lifestyle factors on BMI in the sample.

The results are relevant because BMI is a widely used indicator of health risk, associated with many chronic health conditions. By identifying sleep, physical activity, and smoking as meaningful contributors to BMI variation, our results on how specific factors predict BMI can inform both public health strategies and personal lifestyle choices.

However, several limitations must be acknowledged. The data was collected through surveys, which could cause bias in the data through misreporting, limiting the ability to establish causal relationships, and may reduce generalizability to other populations. Potential unmeasured confounding variables, such as genetic predisposition, underlying medical conditions, or socioeconomic factors, were not included in the model and could partly explain the observed relationships as well.

Assumptions made: Data points were independent of errors.

Despite these limitations, the model offers a statistically sound and interpretable explanation for BMI variation in this dataset. Future research could strengthen these findings by using larger, more diverse samples, incorporating longitudinal data to track BMI changes over time, and including additional predictors such as dietary patterns, sleep quality, and genetic factors. Such extensions would help create a more comprehensive and accurate model for predicting BMI and guiding health interventions.

Contributions: Felix focused on the results section. Sarah focused on the methods section. Aaron focused on the introduction and conclusion sections.

Bibliography

Chiolero, A., Faeh, D., Paccaud, F., & Cornuz, J. (2008, April 1). *Consequences of smoking for body weight, body fat distribution, and insulin resistance*. The American Journal of Clinical Nutrition. https://www.sciencedirect.com/science/article/pii/S0002916523235479?via%3Dihub#ab0005

Feng Y, Jia Y, Jiang J, Wang R, Liu C, Liu W, Wang R. (2024, August 6). *Association between lifestyle factors and mental health in apparently healthy young men*. BMC Public Health. https://pmc.ncbi.nlm.nih.gov/articles/PMC11301853/#Abs1

Patel, S.R., Hu, F.B. (2012, September 6). *Short Sleep Duration and Weight Gain: A systematic Review*. Wiley Online Library: Obesity. https://onlinelibrary.wiley.com/doi/10.1038/oby.2007.118

Pruim, R. (2015). *NHANES: Data from the US National Health and Nutrition Examination Study* (Version 2.1.0) [R package]. https://cran.r-project.org/package=NHANES

U.S. Department of Health and Human Services. (accessed July 17 2025.). *Assessing your weight and health risk*. National Heart Lung and Blood Institute. https://www.nhlbi.nih.gov/health/educational/lose_wt/risk.htm#:~:text=BMI%20is%20a%20useful%20measure,breathing%20problems%2C%20and%20certain%20cancer