How do lifestyle and poverty influence BMI across different age groups and genders*

Negative Correlations with Poverty and Physical Activity, Positive Correlations with Age, Sleep Duration, and Gender

Sakura Hu

December 14, 2024

This paper investigates the relationship between BMI and factors such as lifestyle, poverty, age, and gender, aiming to identify patterns that influence maintaining a healthy BMI. The analysis uses the NHANES dataset from the US National Health and Nutrition Examination Survey and applies a multilinear regression model. The findings suggest weak overall correlations between log(BMI) and the predictors, though some significant patterns are observed: BMI is negatively associated with poverty and physical activity levels and positively associated with age, sleep duration, and being male. These results underscore the complexity of factors influencing BMI and highlight potential areas for targeted public health interventions to promote healthier lifestyles.

1 Introduction

Body mass index (BMI) is a widely used measure for assessing whether an individual's weight is within a healthy range, with significant implications for health outcomes such as heart disease, diabetes, and mortality. Given its importance, understanding the factors that influence BMI, such as socioeconomic and lifestyle variables, has become an essential area of public health research. This paper aims to address the relationship about how variables like poverty, physical activity, sleep duration, age and gender interact to influence BMI outcomes.

Using resampled version of the data from the US National Health and Nutrition Examination Survey (NHANES), this study models log-transformed BMI as a function of five predictors: poverty level, physical activity frequency (measured in days), age, sleep duration, and gender. A multiple linear regression approach was employed to quantify these relationships and

[*] Code and data are available at: https://github.com/xycww/BMI.

identify patterns within the dataset. The estimand in this study is the expected change in log-transformed BMI resulting from a one-unit change in each predictor variable, holding all other variables constant.

The results indicate that while the correlations between log(BMI) and these predictors are generally weak, several significant relationships are observed. Poverty is associated with a 0.01 decrease in log(BMI), suggesting that higher income corresponds to lower BMI levels. Physical activity frequency also demonstrates a negative relationship with BMI; each additional day of physical activity per week is associated with a 0.004 decrease in log(BMI), indicating a modest benefit of regular exercise for maintaining lower BMI. Sleep duration shows a negative relationship, where one additional hour of sleep per night is linked to a 0.01 decrease in log(BMI). In contrast, age is positively associated with BMI, with each additional year corresponding to a 0.001 increase in log(BMI), reflecting the gradual weight gain commonly seen with aging. Lastly, being male is associated with a 0.02 increase in log(BMI) compared to females, suggesting possible physiological or behavioral differences between genders.

These results are essential for informing public health initiatives aimed at addressing weight-related health challenges. By identifying specific socioeconomic and lifestyle factors that influence BMI, this research provides a foundation for developing targeted interventions and strategies to promote healthier weight maintenance.

The remainder of this paper is structured as follows. Section 2 provides an overview and measurement of the dataset, as well as an introduction to the variables. Section 3 presents the model setup and justification. Section 4 discusses the results of the model. Section 5 provides a detailed discussion of the results. Section B offers a more in-depth justification of the data and discusses the surveys and sampling methods. Finally, Section D examines the diagnostics of the model by analyzing the assumptions of linear regression.

2 Data

2.1 Overview

The dataset used in this analysis is derived from the 2009-2010 cycle of the US National Health and Nutrition Examination Survey (NHANES). This version of the dataset includes survey results collected between October 2009 and December 2010. NHANES is a long-running study conducted by the US National Center for Health Statistics (NCHS) that has been gathering health and nutrition data since the early 1960s. Since 1999, approximately 5,000 individuals from various age groups have been interviewed annually in their homes and undergone health examinations at mobile examination centers (MEC).

Two datasets in this package were considered for this analysis: NHANESraw, which is the original raw data, and NHANES, a resampled version of the NHANES data. NHANESraw contains the original survey data with 20,293 observations and additional variables describing the sample weighting scheme, while NHANES is a simplified version with 10,000 resampled observations to account for oversampling effects. NHANES is used in this analysis due to its ability to reduce the potential biases from the complex survey design in NHANESraw. Additional details about this choice are provided in Section B.1. The data used here was originally compiled by Michelle Dalrymple from Cashmere High School and Chris Wild from the University of Auckland for educational purposes.

For the current study, the data was cleaned to focus on variables pertinent to the analysis of BMI. Specifically, variables such as BMI, poverty index, physical activity days, sleep hours, gender, and age were retained. After cleaning the missing values and the duplicate rows in the dataset, 2841 observations remained.

The dataset was prepared, cleaned, and analyzed using R (R Core Team 2023) with the following libraries: tidyverse (Wickham et al. 2019) and dplyr (Wickham et al. 2023) for data manipulation, ggplot2 (Wickham 2016), kableExtra (Zhu 2024) and patchwork (Pedersen 2024) for visualizations, and broom (Wickham 2024) for model summaries. The arrow (Richardson et al. 2024) library was used for efficient data storage and retrieval, while knitr (Xie 2024) facilitated report generation. The NHANES package (Pruim 2015) provided access to the dataset, and styler (Müller, Walthert, and Patil 2024) was employed to ensure well-structured R code. Additionally, testthat (Wickham 2011) and pointblank (Iannone, Vargas, and Choe 2024) were utilized for data validation and testing.

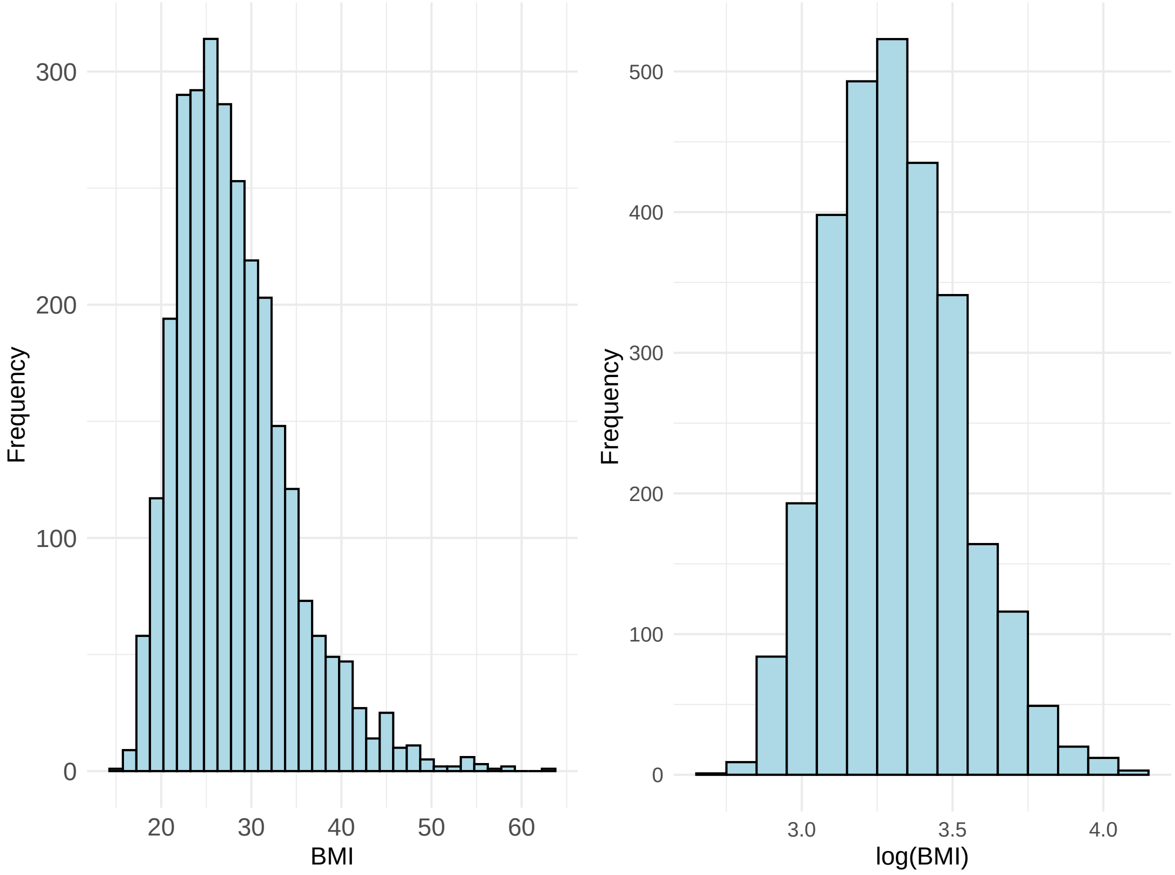
A summary table of cleaned data is shown in Table 1.

Table 1: Summary statistics for `BMI`, `log(BMI)`, `Poverty Index`, `Physical Activity Days`, `Sleep Hours`, `Age`, and `Gender`.

Variable	Mean	Median	Min	Max	1st Quantile	3rd Quantile
BMI	28.059	27.000	15.02	63.3	23.500	31.49
log(BMI)	3.311	3.296	2.709	4	3.157	3.45
Poverty Index	2.940	2.910	0	5	1.340	5.00
Physical Activity Days	3.723	3.000	1	7	2.000	5.00
Sleep Hours	43.850	43.000	16	80	29.000	57.00
Age	6.926	7.000	2	12	6.000	8.00
Gender	NA	NA	1399 (females)	1442 (males)	NA	NA

2.2 Measurement

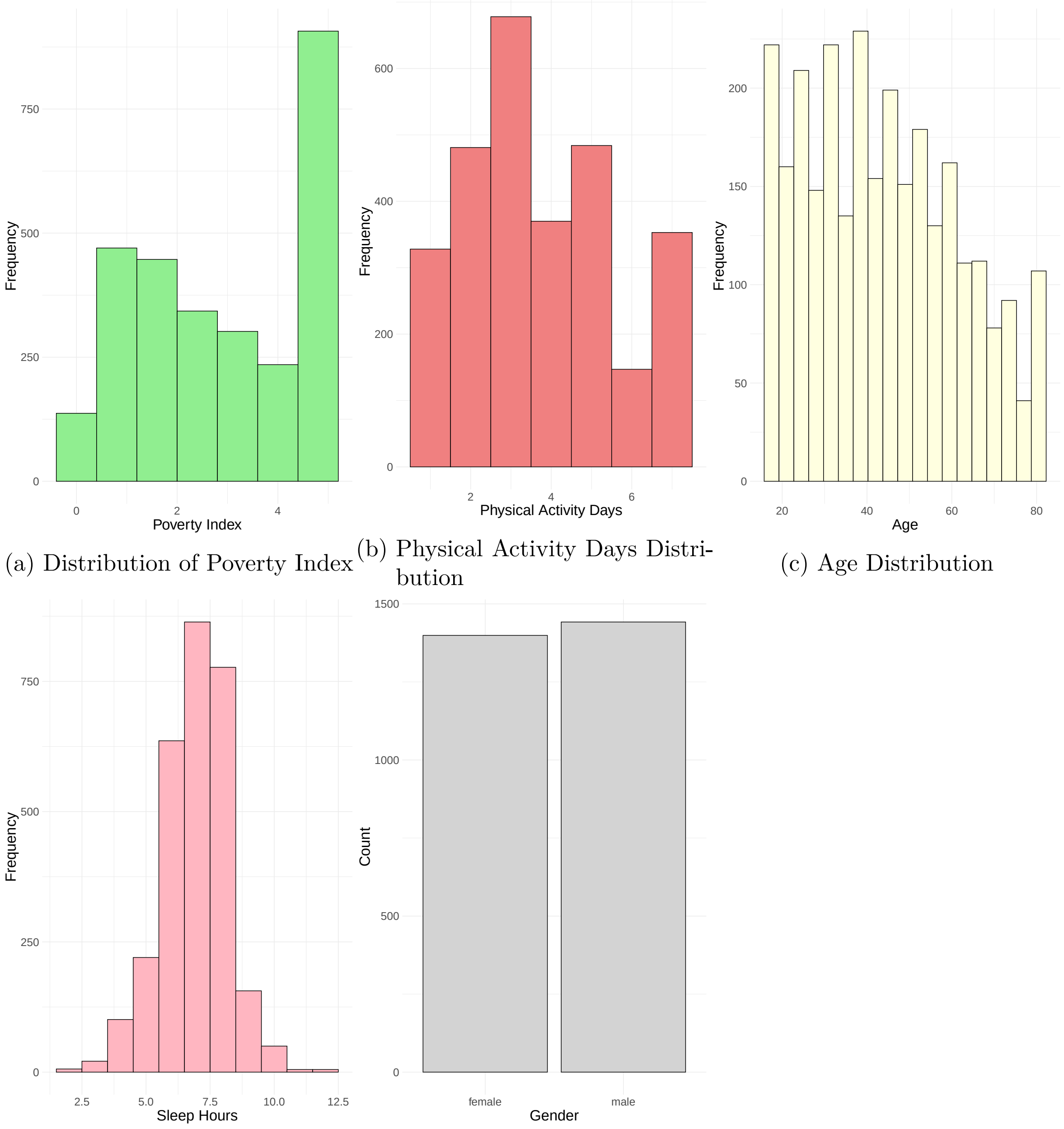
The NHANES dataset provides a comprehensive view of the health and nutritional status of the U.S. population, transforming real-world phenomena into structured data entries. The data collection process begins with the selection of a sample from the U.S. population using a complex, multistage probability sampling design. This ensures that the sample is representative of



(a) Distribution of BMI

(b) Distribution of log(BMI)

Figure 1: Distributions of BMI and log(BMI) in the NHANES dataset.



(a) Distribution of Poverty Index

(b) Physical Activity Days Distribution

(c) Age Distribution

(d) Sleep Hours Distribution

(e) Gender Distribution

Figure 2: Distributions of predictor variables in the NHANES dataset, including Poverty, Physical Activity Days, Age, Sleep Hours, and Gender.

the civilian, non-institutionalized population, while also oversampling certain subgroups, such as racial minorities, to ensure sufficient data for subgroup analysis (Curtin et al. 2013). Once selected, individuals are surveyed on various health and lifestyle factors, such as their eating habits, physical activity, medical conditions, and demographics. The participants are then examined in a Mobile Examination Center (MEC) where they undergo physical assessments, laboratory tests, and interviews.

The process of measurement involves both self-reported data and direct observations. For example, data on an individual's lifestyle habits, such as hours of physical activity, alcohol consumption, or smoking, is obtained through self-reports in interviews (Zipf et al. 2013). These responses are recorded as variables in the dataset, and any missing values or refusals are appropriately coded. On the other hand, physical measurements like height, weight, blood pressure, and cholesterol levels are directly taken in the MEC using medical equipment (Zipf et al. 2013). These physical measurements are then processed and entered into the dataset. For each individual, these variables represent their health status at a specific time, providing a snapshot that is used for statistical analysis.

Importantly, NHANES also accounts for missing data, a common occurrence in surveys. Missing values may arise if a participant refuses to answer a particular question or does not undergo a specific examination. These missing values are systematically handled by assigning special codes such as a period (.) for numeric variables or blanks for character variables, ensuring that they are not treated as valid data (Johnson et al. 2013). Furthermore, to mitigate potential bias due to missing data, adjustments are made using sample weights to produce national estimates that are representative of the U.S. population.

In the case of variables related to health measurements, the data undergo further validation through the use of sampling weights and variance estimation to ensure that the final dataset accurately reflects the population's health status (Curtin et al. 2013). This process involves using the complex survey design parameters, such as the primary sampling unit (PSU) and strata, to adjust the data accordingly, so that analyses can account for oversampling of certain subpopulations and non-response during data collection.

For this analysis, a resampled version of the NHANES dataset, referred to as NHANES, was used. NHANES includes 10,000 resampled observations, and it is derived from the original NHANESraw dataset, which contains 20,293 observations. The resampling process helps reduce biases associated with the complex survey design in the raw dataset, particularly biases arising from oversampling specific subpopulations. This simplified dataset eliminates the effects of oversampling and allows for more straightforward analysis.

The resampled NHANES dataset is structured to reflect a simplified, but still representative, version of the U.S. population, addressing potential sampling imbalances that might arise in the raw data due to overrepresentation of certain groups (Pruim 2015). It is particularly important in this analysis because it minimizes biases related to the complex survey design, making it more suitable for modeling relationships between different factors and log(BMI). By resampling and adjusting for the complex survey design parameters, NHANES ensures that the relationships observed between variables are more accurate and generalizable to the larger population.

Thus, every entry in the NHANES dataset is an outcome of rigorous data collection methods, reflecting a specific health measure or lifestyle factor transformed into a quantifiable variable. These variables have been adjusted to account for complex survey design effects and missing data, ensuring the dataset used in the analysis is robust, representative, and suitable for informing public health research.

2.3 Outcome variables

The primary outcome variable in this study is log-transformed BMI, a measure of body mass index adjusted to normalize its distribution. BMI is widely used to assess healthy weight relative to height, and its relevance to health outcomes such as cardiovascular disease and diabetes has been well-established. For this analysis, BMI was log-transformed to address its skewed distribution shown in Figure 1, providing a better fit for statistical modeling.

2.4 Predictor variables

The following predictor variables were examined to assess the potential lifestyle and socio-economic factors influencing log(BMI) across different gender and age groups:

- `Poverty`: This variable represents the ratio of a family's income to the federal poverty guidelines, with lower values indicating higher levels of poverty.
- `Physical Activity Days (PhysActiveDays)`: The number of days in a typical week that a participant engages in moderate or vigorous physical activity. This variable is recorded for individuals aged 12 years and older.
- `Sleep Hour (SleepHrsNight)`: The self-reported average number of hours of sleep a participant receives on weekdays or workdays. This variable is recorded for individuals aged 16 years and older.
- `Gender`: The gender of the participant, categorized as male or female.
- `Age`: The participant's age at the time of screening, recorded in years. For participants aged 80 years or older, the age was recorded as 80.

2.4.1 Distribution of Predictor Variables

The summary statistics presented in Table 1 and the histograms of predictor variables shown in Figure 2 provide insights into the distribution of these variables:

- `Poverty`: The poverty index ranges from 0 to 5, with a mean of 3.077. The histogram indicates a marked left skew, suggesting that a significant proportion of participants fall into lower income categories.

- `Physical Activity Days (PhysActiveDays)`: The number of days participants engage in physical activity ranges from 2 to 7, with a mean value of 3.7 days per week.
- `Sleep Hour (SleepHrsNight)`:The number of hours participants sleep each night ranges from 2 to 12 hours, with a mean of 6.96 hours. The distribution of this variable approximates a normal curve.
- `Gender`: The gender distribution is nearly balanced, with 1,814 male participants and 1,759 female participants.
- `Age`: The age of participants spans from 16 to 80 years, with a mean age of 43.61 years. The histogram shows a slight right skew, with a concentration of participants aged between 29 and 56 years.

3 Model

The goal of this analysis is to use a multiple linear regression model to investigate the relationship between the log-transformed Body Mass Index (log(BMI)) and several predictors: poverty level, physical activity days, sleep duration, gender, and age. These variables were selected based on their known or hypothesized influence on BMI, as supported by prior research and health literature. The log-transformation of BMI is applied to address skewness in the data and to better approximate a normal distribution, which is an assumption for linear regression models.

Further background details and diagnostics are included in Appendix D.

3.1 Model set-up

Let $log(BMI)_i$ represent the log-transformed BMI for the i-th individual. The multiple linear regression model is defined as:

$$\log(\text{BMI}_i) = \beta_0 + \beta_1 \cdot \text{Poverty}_i + \beta_2 \cdot \text{PhysActiveDays}_i + \beta_3 \cdot \text{Age}_i + \beta_4 \cdot \text{SleepHrsNight}_i + \beta_5 \cdot \text{Gender}_i + \epsilon_i$$

where:

- $log(BMI)_i$ is the log-transformed body mass index for the i-th individual,
- β_0 is the intercept, representing the baseline log(BMI) when all predictors are zero,
- β_1 through β_5 are the coefficients associated with each predictor, representing the change in $log(BMI)_i$ for a one-unit increase in the corresponding predictor, while holding all other predictors constant,
- ϵ_i is the error term, assumed to follow a normal distribution: $\epsilon_i \sim \text{Normal}(0, \sigma^2)$.

This model was fitted using the lm() function in R (R Core Team 2023), with data manipulation performed using the dplyr package and data storage managed by the arrow package. As this is an ordinary least squares (OLS) regression model, no specific priors were applied, and the coefficients were estimated directly from the data.

3.1.1 Model justification

We chose a multiple linear regression model because it is well-suited for estimating the relationship between a continuous dependent variable (log-transformed BMI) and multiple independent variables, including both numerical and categorical predictors (e.g., physical activity days, age, poverty, sleep hours, and gender). This method is interpretable and provides straightforward estimates of the effect of each predictor on BMI.

The decision to log-transform BMI was driven by the variable's right-skewed distribution. Log-transforming BMI improves the normality of its distribution, which is an important assumption for linear regression. It also helps stabilize the variance across levels of BMI, reducing the potential for heteroscedasticity, and enables a more linear relationship between BMI and the predictors.

`Poverty` was included as a predictor to account for socio-economic status, as prior studies (Webber et al. 2023) have shown that lower income is often associated with higher BMI. `PhysActiveDays` captures the number of days an individual engages in physical activity each week, with increased physical activity generally linked to a healthier BMI (Webber et al. 2023). `Age` is included because BMI naturally changes with age, with older individuals often having higher BMI values due to changes in metabolism and lifestyle. `SleepHrsNight` reflects the amount of sleep an individual gets, as insufficient sleep is known to contribute to weight gain and higher BMI (Ekstedt et al. 2013). Lastly, `Gender` is included because research consistently finds gender differences in BMI, with men generally having lower BMI than women (Longo-Silva et al. 2023). The choice to treat gender as a categorical variable is justified because gender is a nominal variable with distinct, non-ordinal categories. For this analysis, gender was encoded as a binary variable (male, female), which allows us to estimate the effect of gender on BMI.

The model is based on the assumption that the relationship between the predictors and log(BMI) is linear, and that the errors are normally distributed with constant variance (homoscedasticity). While linear regression provides a reliable and interpretable method for analyzing the relationship between BMI and the chosen predictors, it does not account for potential interactions between variables. This model assumes that the effect of each predictor is independent of the others, which may not fully capture the complexities of the relationships between the predictors and BMI.

Some other models that were considered were models including interaction terms, such as between poverty and physical activity. However, when these interaction terms were added, the model's fit decreased (as evidenced by a lower R-squared value) and the residual plot indicated potential model misspecification. Given these results, we decided to exclude interaction terms to avoid overfitting and maintain a more parsimonious model. Future analyses with larger datasets could revisit this approach, potentially including interactions or testing alternative modeling techniques to capture more complex relationships.

The model was implemented using R, and all results were validated through diagnostic checks. Specifically, we examined the residual plots to assess the assumptions of linearity, normality of errors, and homoscedasticity. Additional validation techniques, including out-of-sample testing, could be incorporated in future iterations to further assess the model's predictive performance.

4 Results

Our results are summarized in Table 2, and the correlations between predictos and log(BMI) are plotted in Figure 3. The findings indicate that the selected factors collectively have a minimal influence on BMI, as evidenced by the low R-squared value (0.0246), This differs from the conclusions of prior studies such as Ekstedt et al. (2013) and Longo-Silva et al. (2023), which reported stronger effects, likely due to the inclusion of interaction terms that are not part of this analysis. Nevertheless, individual predictors still show significant associations with log(BMI), as indicated by p-values below 0.05 for all predictors except `PhysActiveDays` are lower than 0.05 shows that these predictos are significant. For example, being male increases the average log(BMI) by 0.0172, which is consistent with the findings of Ekstedt et al. (2013). Poverty is negatively associated with BMI, where a one-unit increase in the poverty index is associated with a decrease of 0.0109 in the average log(BMI). This result contradicts the findings of Webber et al. (2023), which reported a positive relationship between poverty and BMI. The discrepancy may stem from differences in data sources, as this analysis uses NHANES data from the 2009–2010 cycle, whereas Webber et al. (2023) examined data from 2017 onwards. We also observed a positive relationship between age and BMI, with the average log(BMI) increasing by 0.0014 for each one-unit increase in age. Additionally, our findings confirm that increased sleep duration and physical activity are associated with lower BMI. Specifically, each additional day of physical activity in a week decreases the average log(BMI) by 0.0035, while each additional hour of sleep per night reduces the average log(BMI) by 0.0108. These observations align with the conclusions of Ekstedt et al. (2013) and Longo-Silva et al. (2023), supporting the hypothesis that healthier lifestyles contribute to lower BMI levels.

Table 2: This table presents the estimated coefficients, standard errors, test statistics, and p-values for the predictors included in the multiple linear regression model analyzing the relationship between log-transformed BMI and poverty level, physical activity days, sleep duration, gender, and age. The R-squared value indicates the proportion of variance in log-transformed BMI explained by the model.

term	estimate	std.error	statistic	p.value
(Intercept)	3.3609829	0.0256909	130.823921	0.0000000
Poverty	-0.0108643	0.0023967	-4.532989	0.0000061
PhysActiveDays	-0.0035154	0.0021651	-1.623650	0.1045617
Age	0.0013841	0.0002270	6.097981	0.0000000
SleepHrsNight	-0.0107754	0.0030139	-3.575183	0.0003558
Gendermale	0.0172291	0.0079590	2.164742	0.0304906
R-squared	0.0246352	NA	NA	NA

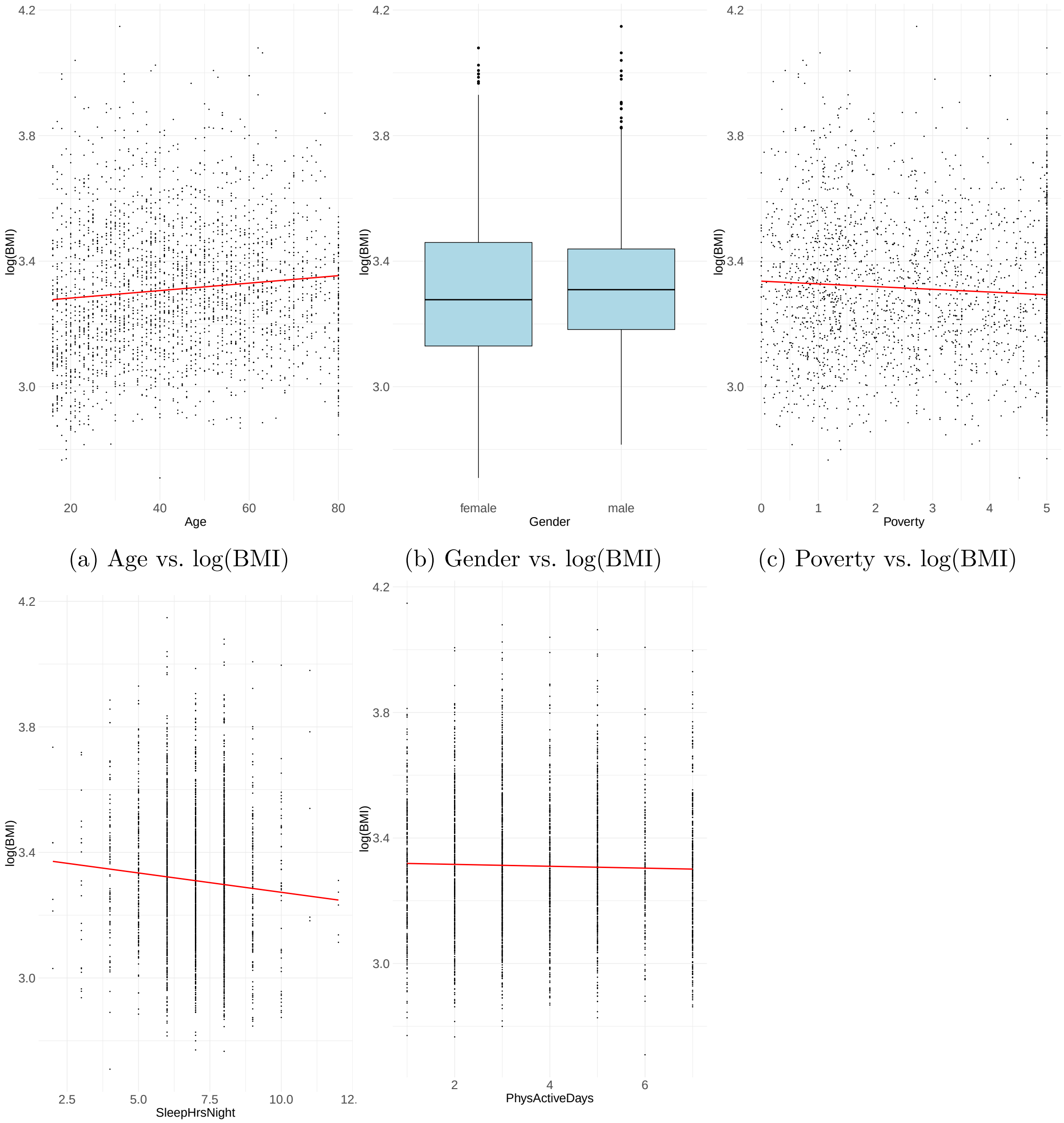
5 Discussion

5.1 What is done in this paper?

This paper explores the relationship between log(BMI) and several key lifestyle and demographic factors, including poverty, physical activity, age, sleep duration, and gender. The analysis is performed using a linear regression model, which allows us to estimate how each of these predictors influences BMI, as measured by its log-transformed values. The dataset used for this study is the cleaned data from the US National Health and Nutrition Examination Survey (NHANES), which contains a broad range of health-related information from a representative sample of the U.S. population.

The model aims to quantify the impact of each predictor on BMI while controlling for the influence of other variables. The regression analysis reveals significant relationships between BMI and all the predictors considered, with some factors having a negative impact (e.g., poverty, physical activity) and others having a positive one (e.g., age, sleep hours, gender). This model sheds light on the ways in which various lifestyle choices and demographic factors may contribute to BMI variation across different segments of the population.

The model used in this study, a multiple linear regression model, is a fundamental yet powerful tool that helps to identify the strength and direction of the relationships between predictors and BMI. The results provide a valuable starting point for understanding how changes in lifestyle and demographic factors may affect BMI. However, while the model is useful, it has its limitations. For instance, it assumes that the relationships between the predictors and BMI are linear, which may not always be the case. Additionally, the low R-squared value suggests that other important factors influencing BMI may not have been included in the model.



(a) Age vs. log(BMI)

(b) Gender vs. log(BMI)

(c) Poverty vs. log(BMI)

(d) Sleep Duration vs. log(BMI)

(e) Physical Active Days vs. log(BMI)

Figure 3: This figure presents the relationships between log-transformed BMI and the predictors in the model. `Panel a` shows the possitive relationship between age and log(BMI). `Panel b` compares the distribution of log(BMI) across genders. `Panel c` illustrates the negative correlation between poverty level and log(BMI), while `Panel d` depicts the negative relationship between sleep duration (in hours) and log(BMI). Finally, `Panel e` shows the negative correlation between the number of physically active days per week and log(BMI)

5.2 What is something that we learn about the world?

One of the key findings from this paper is that poverty is negatively correlated with BMI, and the relationship is statistically significant. The negative coefficient for the poverty variable suggests that individuals in poverty tend to have higher BMI, on average, than those with higher income levels. It may seem counterintuitive at first—since poverty is often associated with limited access to healthy foods and exercise—but the correlation can be explained by several factors.

In many low-income communities, the availability of affordable healthy food options is limited, leading individuals to rely on cheaper, calorie-dense foods that contribute to weight gain. Additionally, individuals in poverty may face multiple barriers to physical activity, including the lack of access to safe recreational spaces or the time and energy needed to engage in exercise, due to long working hours or other stressors. Furthermore, the high cost of healthcare and medical services in lower-income populations may prevent individuals from seeking preventative care and managing conditions like obesity. Thus, this finding underscores the importance of addressing socioeconomic disparities as part of efforts to combat obesity and improve public health.

Moreover, this finding highlights the need for policy interventions that promote access to healthy food and physical activity in low-income communities. It also suggests that any interventions aimed at reducing BMI must consider not just individual behaviors, but the broader socioeconomic context in which these behaviors occur. Policymakers should prioritize interventions that reduce the barriers to healthy living for individuals in poverty, such as improving access to nutritious food, increasing public health education, and providing community resources for physical activity.

5.3 What is another thing that we learn about the world?

Another key takeaway from this paper is that sleep duration has a significant negative correlation with BMI, suggesting that individuals who get more sleep tend to have lower BMI. This relationship is in line with previous research that has examined the link between sleep and obesity. The coefficient for sleep hours per night shows that for each additional hour of sleep, BMI decreases slightly, indicating that sleep plays a role in regulating body weight. This finding has important implications for public health, as sleep duration is a modifiable lifestyle factor that could be targeted in weight management programs.

The relationship between sleep and BMI can be explained through several mechanisms. Poor sleep has been shown to disrupt the body's metabolism, leading to an increase in appetite, particularly for high-calorie foods. Sleep deprivation also affects the hormones that regulate hunger and satiety, increasing the production of ghrelin (the hunger hormone) while decreasing the production of leptin (the satiety hormone). These changes can lead to overeating and, consequently, weight gain. Furthermore, insufficient sleep can lead to fatigue, reducing the likelihood of physical activity and further contributing to weight gain. Therefore, improving sleep hygiene and encouraging sufficient sleep may serve as an effective strategy for reducing BMI and addressing obesity.

This result also emphasizes the importance of taking a holistic approach to health. In addition to focusing on diet and physical activity, interventions aimed at improving sleep quality could be a key component in the prevention and treatment of obesity. Public health campaigns that educate individuals about the importance of sleep, along with strategies to improve sleep hygiene (e.g., creating a conducive sleep environment, avoiding caffeine, establishing regular sleep routines), could complement existing efforts to promote healthy eating and physical activity.

5.4 Weaknesses and next steps

While the model results provide valuable insights into the relationship between log(BMI) and various predictor variables, there are some limitations that should be addressed in future analyses. One significant weakness is the relatively low R-squared value of 0.02335, indicating that the model only explains a small portion of the variance in BMI. This suggests that there are other factors, potentially unaccounted for in the model, that influence BMI. Although several predictors show statistically significant relationships with BMI, the explanatory power of the model remains limited. It is possible that additional variables, such as dietary habits, genetics, or environmental factors, could improve the model's ability to explain the variation in BMI.

Another limitation is the potential for omitted variable bias. Despite including a range of relevant predictors, there may be other important factors that have not been incorporated into the model. For instance, socioeconomic status, mental health, or access to healthcare may play significant roles in determining BMI but were not included in the current analysis. Future models could benefit from a more comprehensive selection of predictor variables to provide a more holistic understanding of BMI variation.

Additionally, the model assumes linear relationships between the predictors and BMI, which may not fully capture the complexities of the data. Interaction effects, such as between poverty and physical activity, could potentially reveal more nuanced relationships, but these were not included in the current model. The absence of interaction terms may have led to the underestimation of the impact of certain predictors, as the effect of one variable may depend on the level of another. Future analyses could explore interaction terms to assess whether these improve model fit and offer deeper insights into the factors that influence BMI.

Another avenue for improvement involves the application of model selection criteria, such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), to systematically compare different model specifications. By testing multiple models with varying combinations of predictors and interaction terms, AIC or BIC could help identify a model that balances explanatory power with parsimony. This approach would allow for the identification of additional variables or interactions that may influence log(BMI) while avoiding overfitting. Incorporating model selection techniques could also provide a clearer understanding of which variables contribute meaningfully to BMI variation, guiding the development of a more robust and interpretable model.

Moving forward, there are several steps that could be taken to address these weaknesses. First, the inclusion of additional predictors could improve the model's explanatory power. Collecting more detailed data on lifestyle, diet, or mental health factors would help provide a more complete picture of the determinants of BMI. Second, using a non-linear model or exploring transformations of the predictors might better capture the relationships between BMI and the explanatory variables. Third, addressing the possibility of interaction effects and testing different model specifications could enhance the robustness of the results. This could include experimenting with non-parametric models, such as random forests or gradient boosting machines, to better handle the complexities of the data. Finally, employing model selection criteria like AIC or BIC during the exploratory phase would ensure that the final model is both parsimonious and comprehensive. By expanding the scope of the analysis and refining the model, future work could yield more precise estimates and offer more actionable insights for public health interventions.

A Appendix

B Additional data details

B.1 Data Justication

The NHANES_raw dataset represents a rich and expansive collection of data from the American National Health and Nutrition Examination Surveys (NHANES). It encompasses 20,293 observations across 75 variables, including data on health, nutrition, and demographic factors. Additionally, NHANES_raw incorporates variables describing the sample weighting scheme employed to account for its complex, multistage survey design. This design includes oversampling of certain subpopulations, such as racial and ethnic minorities, to improve the precision of estimates for these groups. While this approach enhances its utility for population-level research, it introduces complexities for analysis, particularly for researchers unfamiliar with survey design methodologies.

A critical limitation of NHANES_raw is its reliance on survey weights to account for the oversampling of subpopulations and ensure representativeness of the U.S. population. Naïve analyses that disregard these weights can yield misleading results. For example, the raw racial composition in NHANES_raw does not reflect the true demographic distribution of the U.S. population due to intentional oversampling adjustments. Failure to account for the survey weights risks overstating statistical significance, generating biased estimates, and drawing inaccurate conclusions. As such, analyses based on NHANES_raw require expertise in survey statistics and careful implementation of weighting techniques to produce valid results.

For the purposes of this study, the NHANES dataset, a simplified resampled subset of NHANES_raw, was selected. This subset comprises 10,000 observations and retains the same 75 variables but approximates a simple random sample from the U.S. population. The resampling process mitigates the challenges associated with the original survey design by removing the need to manually account for oversampling effects and survey weights. This simplification makes NHANES particularly suitable for educational applications, as it allows for straightforward statistical analyses without requiring advanced knowledge of survey methodologies. Furthermore, the subset retains sufficient variability to explore meaningful relationships between variables while reducing the risk of misinterpretation arising from unaddressed biases in the original dataset.

The decision to use NHANES in this paper was guided by its alignment with the study's educational focus and analytical goals. Specifically, it provides a clean and interpretable framework for investigating the relationships between body mass index (BMI) and factors such as lifestyle, poverty, age, and gender. By simplifying the sampling complexities of the original dataset, NHANES ensures that the analyses presented in this paper are accessible to a broader audience, including those less familiar with survey weighting procedures.

It is important to recognize that while NHANES is tailored for educational purposes, it is not suitable for formal research aimed at generating population-level inferences. Researchers conducting such studies should rely on NHANES_raw and adhere strictly to the prescribed sampling and weighting procedures to maintain the validity and accuracy of their findings. Future extensions of this analysis could utilize NHANES_raw in conjunction with survey weighting techniques to examine population-level health trends and disparities with greater precision. Additionally, incorporating variables specific to the survey design—such as strata and cluster identifiers—could enable a more comprehensive exploration of the complexities underlying the relationships between health outcomes and sociodemographic factors. This approach would provide a more robust framework for addressing nuanced public health questions at the national level.

C Surveys, Sampling, and Observational Data

C.1 Methodology Analysis

C.1.1 Introduction

The National Health and Nutrition Examination Survey (NHANES) is a critical tool for understanding the health and nutritional status of the U.S. population. Managed by the National Center for Health Statistics (NCHS), NHANES gathers comprehensive data through interviews and physical examinations, providing invaluable insights into the health behaviors and chronic conditions affecting diverse groups. This essay aims to explore NHANES's methodology in-depth, including its target population, sampling strategy, recruitment process, non-response handling, and data collection techniques.

C.1.2 Target Population and Sampling Methodology

NHANES aims to represent the civilian, non-institutionalized U.S. population, encompassing individuals of all ages, genders, and racial/ethnic groups. The survey's sampling strategy is designed to be representative of the entire U.S. population and includes specific oversampling of groups that are underrepresented or have health disparities. This methodology allows researchers to obtain reliable estimates of health outcomes across different demographic groups.

C.1.2.1 Stratified, Multistage Sampling

NHANES employs a **stratified, multistage sampling design** to select participants, ensuring diverse representation and reducing sampling bias. The process begins with the selection of Primary Sampling Units (PSUs), which are typically counties or groups of contiguous counties.

These PSUs are chosen using a probability proportional to size (PPS) method, ensuring that more populous areas have a higher likelihood of being selected.

The sampling process proceeds through several stages: 1. **PSU Selection**: Using PPS, counties or groups of counties are selected. 2. **Segment Selection**: Segments, often representing census blocks, are chosen within selected PSUs. 3. **Household Selection**: Households within selected segments are identified and sampled. 4. **Individual Selection**: Individuals are selected within households, ensuring balanced representation across demographics such as age, gender, and race.

This **probabilistic design** allows each person in the target population to have a known, nonzero probability of selection, making it a powerful tool for generalizing the results to the U.S. population.

C.1.2.2 Oversampling and Stratification

One of the key strengths of NHANES's sampling design is its **oversampling of certain subgroups**. Specifically, NHANES oversamples non-Hispanic Black individuals, Hispanic individuals, and elderly individuals, especially those over the age of 80. This approach ensures that the survey yields sufficiently accurate and reliable estimates for these populations, which may otherwise be underrepresented in a simple random sample.

This stratified sampling approach ensures that key demographic groups are adequately represented, which is particularly important when assessing health disparities in the U.S. population.

C.1.2.3 Variance Estimation and Survey Weights

Given the multi-stage, stratified design, NHANES data require careful handling during analysis. **Survey weights** are applied to adjust for unequal probabilities of selection. These weights correct for the fact that individuals in more populous areas have a higher chance of selection, and they help adjust for the over- or underrepresentation of certain subgroups.

Additionally, the sampling clusters within PSUs introduce **intra-cluster correlation**, meaning that individuals within the same PSU are more likely to share similar characteristics than individuals across PSUs. Variance estimation methods, such as Taylor Series Linearization or Balanced Repeated Replication, are used to account for this correlation in order to generate accurate confidence intervals and significance tests.

C.1.3 Recruitment Process

The recruitment process for NHANES is designed to maximize participation rates while ensuring the sample remains representative of the U.S. population. Each stage in the sampling process is carefully designed to engage diverse populations, reduce bias, and address barriers to participation.

C.1.3.1 Stage 1: PSU Selection

Primary Sampling Units (PSUs), typically counties or groups of contiguous counties, are selected using PPS. This ensures that larger counties with higher populations have a higher probability of being selected.

C.1.3.2 Stage 2: Segment and Household Selection

Within each PSU, geographic segments are selected, often representing census blocks. From these segments, households are selected, ensuring a mix of rural, suburban, and urban areas.

C.1.3.3 Stage 3: Individual Selection

Within each household, individuals are chosen to participate in the survey, with selection methods that account for key demographic variables such as age, gender, and ethnicity. By choosing individuals based on these criteria, NHANES ensures that the sample is balanced across relevant characteristics.

C.1.4 Non-Response and Data Adjustments

Non-response is a common issue in large-scale surveys, and NHANES employs multiple strategies to minimize its impact on the results. Non-response can occur at two levels: unit non-response (refusals to participate) and item non-response (failure to respond to specific questions).

C.1.4.1 Unit Non-Response

Unit non-response occurs when selected individuals refuse to participate or are unable to participate for other reasons. In the 2009-2010 cycle, for instance, 21% of eligible participants did not complete the household interview, and 2% of those who completed the interview refused the medical examination. Non-response at the unit level is adjusted by applying **sampling weights** that reflect the demographic characteristics of non-respondents.

C.1.4.2 Item Non-Response

Item non-response happens when respondents skip or refuse to answer specific questions or tests, such as blood pressure measurements or questions about smoking habits. NHANES addresses item non-response through the use of imputation techniques or by applying weights that account for missing data.

C.1.5 Data Collection and Questionnaire Design

NHANES collects an extensive range of health and nutrition data through interviews and physical examinations, providing a rich dataset for public health research. The survey includes a variety of sections, such as:

1. **Demographic Information**: Age, sex, race, income, education level.
2. **Health Behaviors**: Smoking, alcohol consumption, physical activity levels.
3. **Medical History**: Chronic diseases, medication use, and past health diagnoses.
4. **Dietary Intake**: Detailed dietary recalls, including information about calorie and nutrient intake.

C.1.5.1 Physical Examinations

Participants undergo a series of medical tests at Mobile Examination Centers (MECs), which include: - **Anthropometric Measurements**: Height, weight, and BMI. - **Blood Pressure**: Both systolic and diastolic blood pressure measurements. - **Laboratory Tests**: Blood and urine samples to assess biochemical markers of health.

These objective measurements complement the self-reported data from the interviews, providing a more complete picture of participants' health status.

C.1.5.2 Limitations of Self-Reported Data

While NHANES includes objective measurements, it also relies heavily on **self-reported data**, which can be prone to biases such as recall bias or social desirability bias. For instance, respondents may underreport behaviors like alcohol consumption or smoking, or overestimate their physical activity levels. This introduces potential inaccuracies, especially for health behaviors that are sensitive or socially stigmatized.

C.1.6 Strengths and Limitations of NHANES

C.1.6.1 Strengths

NHANES provides robust, nationally representative data that can be used to track health trends, monitor disease prevalence, and guide public health policy. The **multi-dimensional data collection**—which includes both subjective interviews and objective examinations—offers a comprehensive understanding of health issues across the U.S. population.

Additionally, the **oversampling of subgroups** ensures that groups typically underrepresented in national surveys are adequately studied, which enhances the generalizability of findings to diverse populations.

C.1.6.2 Limitations

Despite its strengths, NHANES faces several limitations: 1. **Non-response**: Even with the application of weights, non-response can introduce bias, especially among older adults or individuals in poor health. 2. **Self-Report Bias**: The reliance on self-reported data for behaviors such as diet, physical activity, and substance use is subject to significant recall and social desirability biases. 3. **Data Cycles**: NHANES data is collected in two-year cycles, which can make it challenging to estimate trends for smaller subgroups or rare health conditions. Combining data from multiple cycles may be necessary, but this introduces additional complexities in analysis.

C.1.7 Conclusion

NHANES is a powerful and invaluable resource for public health research, offering detailed, nationally representative data on a wide range of health and nutritional topics. The survey's complex, stratified, multistage sampling design ensures that the data are reflective of the U.S. population. Although challenges such as non-response and the limitations of self-reported data exist, NHANES remains a critical tool for understanding public health issues and guiding policy decisions. Researchers must apply appropriate statistical techniques to address these challenges, ensuring the robustness and accuracy of their findings.

C.2 Idealized survey

The survey will be structured as follows:

C.2.1 Introductory Section

The survey will begin with a clear introduction outlining the purpose of the survey, ensuring participants understand the goals of the study. This section will also describe the confidentiality measures, emphasizing that personal information will not be shared. Contact details for questions or concerns will be provided.

C.2.2 Demographics and Personal Information (Section 1)

This section will collect basic information, such as age, gender, ethnicity, and socio-economic status, which are known to impact BMI.

C.2.3 Health and Lifestyle Factors (Section 2)

This section will inquire about physical activity levels, dietary habits, alcohol consumption, smoking, and sleep patterns. It will include validated scales such as the International Physical Activity Questionnaire (IPAQ) and a food frequency questionnaire (FFQ).

C.2.4 Genetic and Family History (Section 3)

Participants will be asked if they have any family history of obesity, diabetes, or cardiovascular disease, which can help identify genetic predispositions.

C.2.5 Psychological and Social Factors (Section 4)

This section will include questions on mental health, stress levels, and social support, as these factors can also affect BMI.

C.2.6 BMI Measurement (Section 5)

Participants will be asked to self-report their weight and height, and their BMI will be calculated. This will be followed by a prompt for participants to either confirm their BMI or opt to take part in a physical measurement of weight and height.

C.2.7 Closing Section

A thank-you message will be included, and respondents will be reminded that their participation is voluntary and can be withdrawn at any time.

C.3 3. Sampling Approach

A stratified random sampling approach will be used to ensure that the survey sample is representative of the general population. The stratification will be based on key demographic factors, such as:

- **Age**: Different age groups will be targeted to examine how BMI and its predictors vary across the lifespan.
- **Gender**: Both men and women will be surveyed to investigate gender differences in BMI-related factors.
- **Socio-economic Status**: Income and education levels will be used to examine how socio-economic factors influence BMI.
- **Geographic Location**: Participants will be sampled from urban, suburban, and rural areas to capture regional variations in BMI and health behaviors.

The sample size will be approximately 10,000 respondents, ensuring statistical power for the analysis. The sampling frame will be drawn from public voter registration lists, with additional recruitment through social media and community outreach to reach diverse populations.

C.4 4. Recruitment Strategy

Recruitment will be conducted through multiple channels:

- **Online Surveys**: A link to the survey will be distributed via social media platforms, targeting a broad audience. A paid advertisement campaign will be used to ensure a diverse and representative sample.

- **Community Outreach**: Partnerships with community organizations will allow for in-person recruitment, especially for populations that may be underrepresented in online surveys.

- **Incentives**: Participants will receive a small financial incentive (e.g., $5) to encourage participation.

C.5 5. Non-Response Handling

To handle non-response bias, several strategies will be employed:

- **Follow-up Invitations**: Non-respondents will receive two email reminders and a phone call if their contact information is available.
- **Incentive Adjustment**: A larger incentive (e.g., $10) will be offered for participants who complete the survey within a certain time frame.

C.6 6. Data Validation

Data quality will be ensured through the following measures:

- **Data Consistency Checks**: Automatic consistency checks will be built into the survey platform to flag implausible responses (e.g., extreme self-reported height or weight).
- **Randomized Validation Sample**: A subset of respondents will be asked to visit a local clinic for physical measurements of height and weight, allowing for validation of self-reported data.

C.7 7. Survey Questions

The questions will be constructed to minimize biases and ensure reliability. The following are examples of question types:

- **Demographic Questions:**
  - "What is your age?"
  - "What is your gender?"
- **Physical Activity Questions:**
  - "On average, how many days per week do you engage in at least 30 minutes of moderate physical activity?"
  - "How many hours of sedentary activity (e.g., watching TV, using a computer) do you engage in per day?"
- **Dietary Habits:**
  - "How often do you consume fruits and vegetables on a typical day?"
  - "How many servings of sugary drinks do you consume per week?"
- **Psychological Factors:**
  - "On a scale from 1 to 10, how would you rate your current stress level?"
  - "Do you feel supported by your family and friends?"

C.8 8. Aggregation and Statistical Methods

The collected data will be aggregated and analyzed using multiple regression techniques to identify significant predictors of BMI. The following methods will be employed:

- **Descriptive Statistics** will be used to summarize demographic data and BMI-related variables.

- **Regression Analysis** will help identify key predictors of BMI (e.g., physical activity, diet, genetics).
- **Interaction Terms** will be included to examine how combinations of factors (e.g., physical activity and diet) jointly affect BMI.

C.9 9. Questionnaire Design and Order

The survey will be structured to minimize response bias. Sensitive questions (e.g., those about weight and diet) will be placed at the end of the survey, once rapport has been built. Demographic questions will appear first to quickly categorize participants without triggering discomfort.

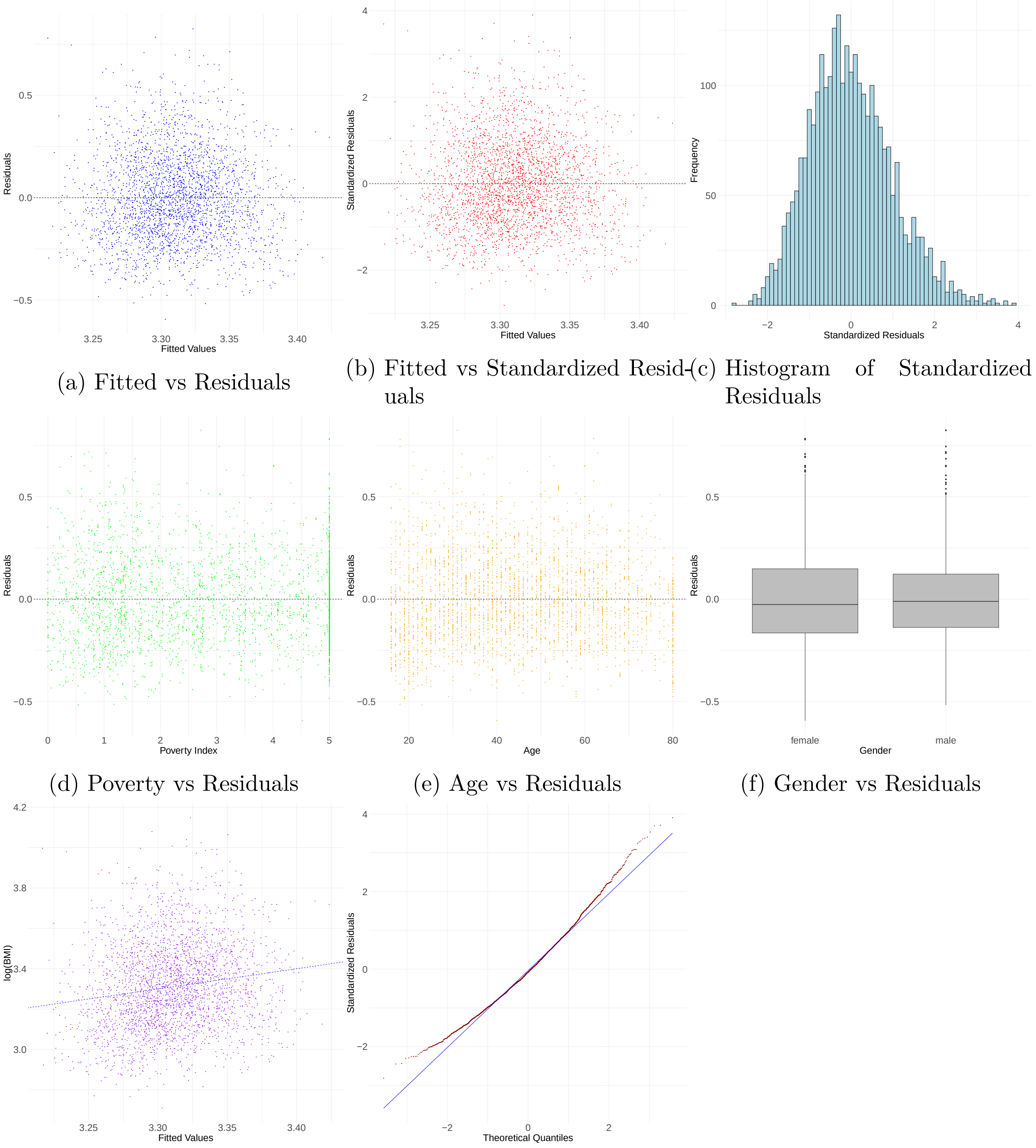
C.10 10. Survey Platform

The survey will be implemented using a robust platform like Google Forms or Qualtrics, allowing for automated data collection and validation.

D Model details

D.1 Diagnostics

The analysis of linear regression assumptions shows that all conditions are met. The fitted versus residual plot displays a null plot centered around zero, indicating that the linearity assumption is satisfied, as no discernible pattern exists. Additionally, the consistent spread of residuals across the fitted values confirms the constant error variance assumption. The absence of correlation or patterns suggests that the independence of errors is also satisfied. Both the fitted versus standardized residuals plot and the predictor variables versus residuals plots support these conclusions. The QQ plot shows standardized residuals following a straight diagonal line, while the histogram indicates a distribution close to ~N(0,1), confirming the normality assumption of errors. Lastly, the response variable versus fitted values plot shows that observed values align with predicted values, suggesting that the model accurately predicts the response variable. Although the residual plots indicate that all regression assumptions are met, the predictor versus response plots show weak correlations between predictor variables and log(BMI).



(a) Fitted vs Residuals

(b) Fitted vs Standardized Residuals

(c) Histogram of Standardized Residuals

(d) Poverty vs Residuals

(e) Age vs Residuals

(f) Gender vs Residuals

(g) Fitted vs log(BMI)

(h) Normal Q-Q Plot

Figure 4: Diagnostic plots assessing the assumptions and performance of the linear regression model for predicting log(BMI). The plots evaluate the residuals and fitted values to ensure the validity of the linear model. (1) Residuals vs. Fitted Values tests the assumption of homoscedasticity. (2) Standardized Residuals vs. Fitted Values examines the presence of outliers and systematic patterns. (3) Distribution of Standardized Residuals evaluates the normality of residuals. (4) Residuals vs. Poverty Index and (5) Residuals vs. Age assess the independence of residuals concerning key predictors. (6) Residuals vs. Gender compares residuals across categorical groups. (7) Fitted Values vs. Observed log(BMI) visualizes model accuracy, and (8) Q-Q Plot assesses the normality of residuals against theoretical quantiles. Together, these plots ensure the appropriateness of the model and identify potential violations of its assumptions.

References

Curtin, Lester R, Leila K Mohadjer, Suzanne M Dohrmann, et al. 2013. “National Health and Nutrition Examination Survey: Sample Design, 2007–2010.” *Vital Health Stat 2* 160.

Ekstedt, M., G. Nyberg, M. Ingre, Ö. Ekblom, and C. Marcus. 2013. “Sleep, Physical Activity and BMI in Six to Ten-Year-Old Children Measured by Accelerometry: A Cross-Sectional Study.” *International Journal of Behavioral Nutrition and Physical Activity* 10 (1): 82. https://doi.org/10.1186/1479-5868-10-82.

Iannone, Richard, Mauricio Vargas, and June Choe. 2024. *Pointblank: Data Validation and Organization of Metadata for Local and Remote Tables*. https://CRAN.R-project.org/package=pointblank.

Johnson, Claudia L, Rupa Paulose-Ram, Cynthia L Ogden, et al. 2013. “National Health and Nutrition Examination Survey: Analytic Guidelines, 1999–2010.” *Vital Health Stat 2* 161.

Longo-Silva, G., A. K. P. Pedrosa, P. M. B. de Oliveira, J. R. da Silva, R. C. E. de Menezes, P. de M. Marinho, and R. S. Bernardes. 2023. “Beyond Sleep Duration: Sleep Timing Is Associated with BMI Among Brazilian Adults.” *Sleep Medicine* X: 100082. https://doi.org/10.1016/j.sleepx.2023.100082.

Müller, Kirill, Lorenz Walthert, and Indrajeet Patil. 2024. *Styler: Non-Invasive Pretty Printing of r Code*. https://cran.r-project.org/web/packages/styler/index.html.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

Pruim, Randall. 2015. *NHANES: Data from the US National Health and Nutrition Examination Study*. https://CRAN.R-project.org/package=NHANES.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Webber, B. J., D. B. Bornstein, P. A. Deuster, F. G. O'Connor, S. Park, K. M. Rose, and G. P. Whitfield. 2023. “BMI and Physical Activity, Military-Aged u.s. Population 2015–2020.” *American Journal of Preventive Medicine* 64 (1): 66–75. https://doi.org/10.1016/j.amepre.2022.08.008.

Wickham, Hadley. 2011. “Testthat: Get Started with Testing.” *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2024. *David Robinson and Alex Hayes and Simon Couch*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

Zipf, George, Michele Chiappa, Kimberly S Porter, et al. 2013. "National Health and Nutrition Examination Survey: Plan and Operations, 1999–2010." *Vital Health Stat 1* 56.