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.

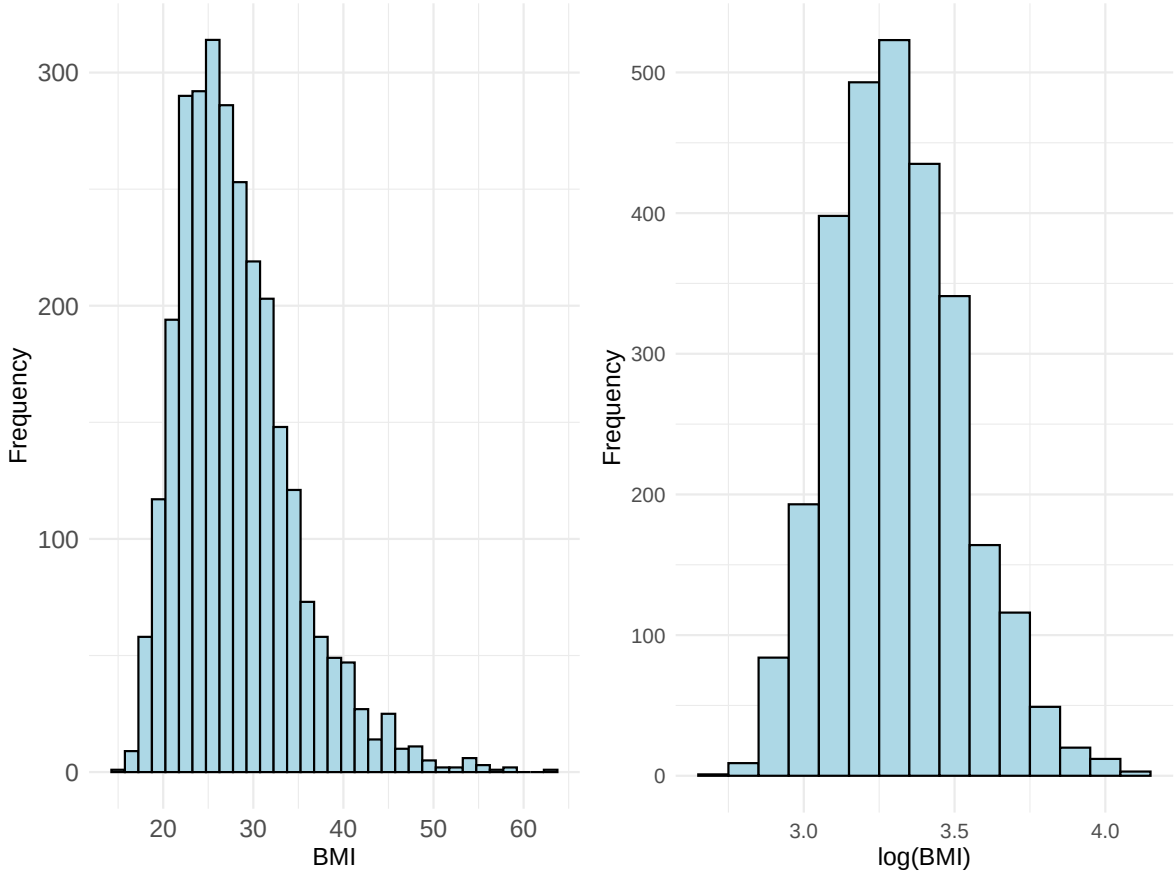
A summary table of cleaned data is shown in Table 1.

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

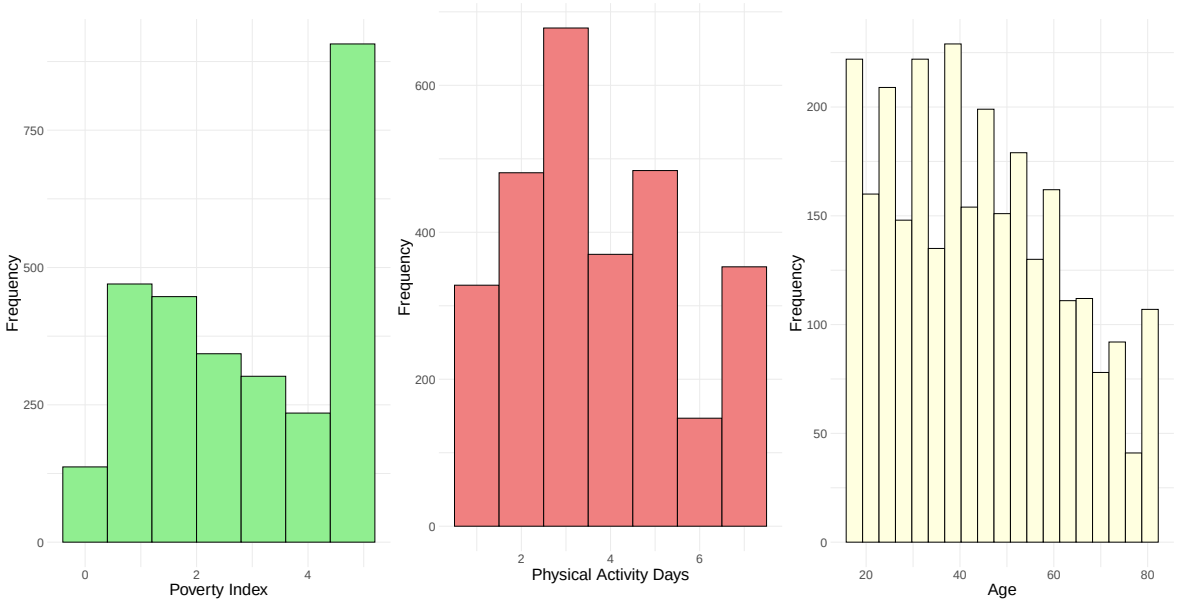
The NHANES dataset provides a comprehensive view of the health and nutritional status of the U.S. population, transforming real-world phenomena into structured data entries. The data collection process begins with the selection of a sample from the U.S. population using a complex, multistage probability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



(a) Distribution of BMI

(b) Distribution of log(BMI)

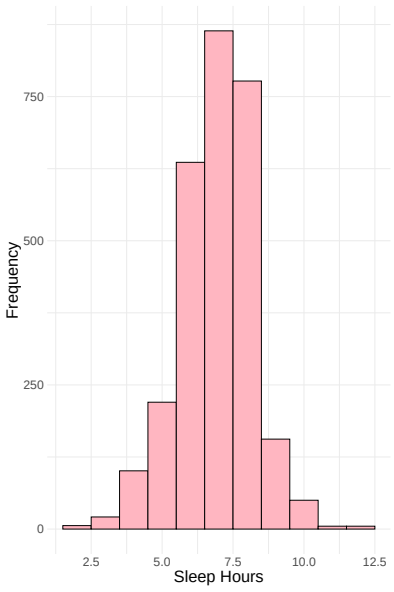
Figure 1: Distributions of BMI and log(BMI) in the NHANES dataset.



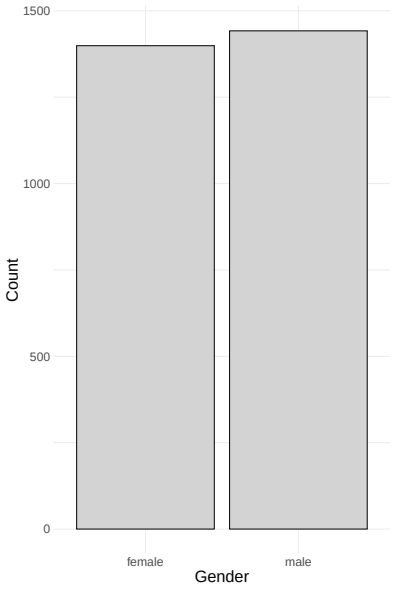
(a) Distribution of Poverty Index

(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

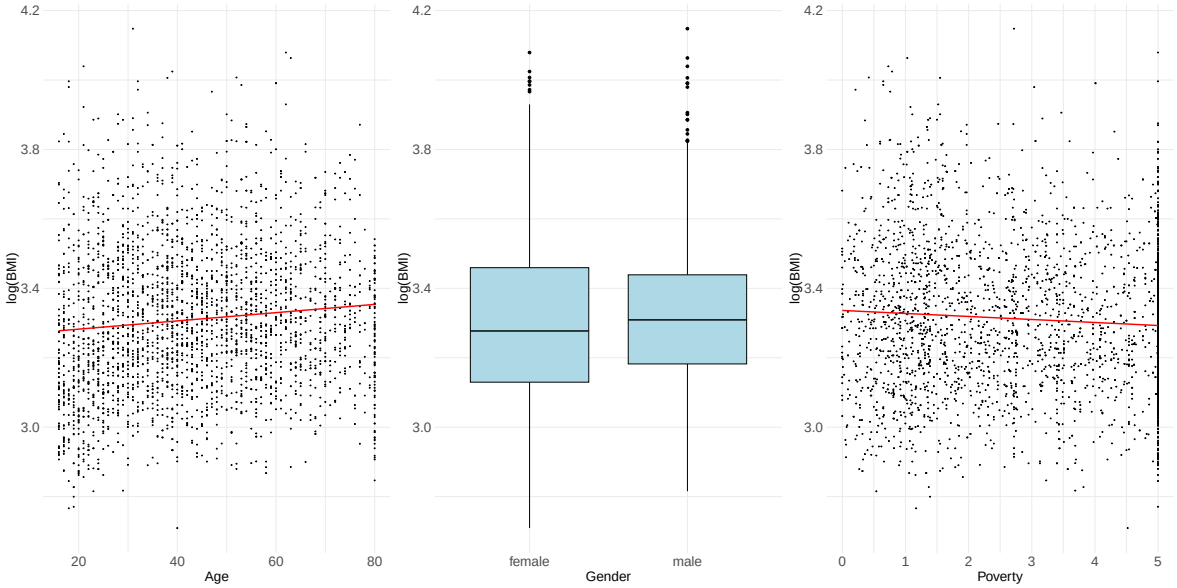
5 Discussion

5.1 What is done in this paper?

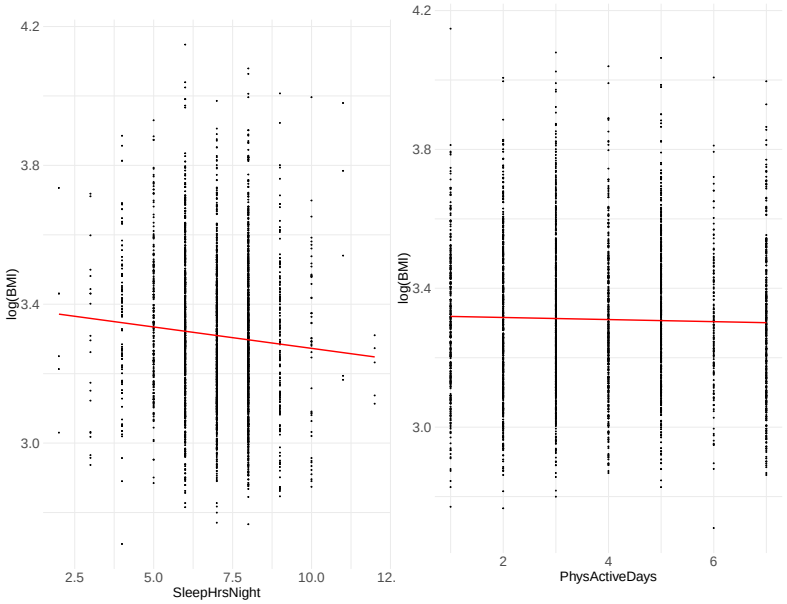
page This paper explores the relationship between Body Mass Index (BMI) and several key lifestyle and demographic factors, including poverty, physical activity, age, sleep duration, and gender. The analysis is performed using a linear regression model, which allows us to estimate how each of these predictors 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 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 is not without 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. This result is consistent with a growing body of literature that links lower socioeconomic status to obesity and poor health outcomes. 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.

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. Finally, 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. 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 {{#sec-append}}

B Additional data details

B.1 Data Justication

The NHANES_raw dataset provides a comprehensive collection of data from the American National Health and Nutrition Examination Surveys (NHANES), including 20,293 observations spanning 75 variables. This dataset includes additional variables that describe the sample weighting scheme employed to account for the complex survey design. However, despite its breadth and richness, NHANES_raw introduces complexities in analysis due to the oversampling of certain subpopulations, such as racial minorities. This sampling approach necessitates the use of survey weights and other design-specific parameters to obtain unbiased estimates and valid statistical inferences.

Naïve analysis of NHANES_raw—without accounting for the sample design and weights—can lead to significant biases and erroneous conclusions. For instance, the racial composition in NHANES_raw does not represent the true demographic distribution of the U.S. population due to oversampling adjustments. As a result, researchers must account for these weights in any analysis to avoid overstating significance levels and generating inaccurate results. This makes NHANES_raw unsuitable for straightforward analysis in contexts where the primary goal is to demonstrate simple statistical relationships, especially for educational purposes.

In contrast, the NHANES dataset, which is a resampled subset of NHANES_raw, addresses the complexities associated with the oversampling design. With 10,000 observations and the same 75 variables, NHANES is simplified to approximate a simple random sample from the U.S. population. This resampling process removes the need for researchers to manually account for the oversampling effects, enabling analyses that are easier to interpret and less prone to methodological errors for educational purposes.

The decision to use NHANES in this paper was driven by its appropriateness for demonstrating relationships between body mass index (BMI) and lifestyle, poverty, age, and gender. NHANES allows for clear and reliable analyses without requiring specialized expertise in survey design or sampling weights, making it ideal for this study's educational focus. The dataset enables meaningful exploration of public health questions while minimizing the risk of misinterpretation due to unaddressed biases in the sampling framework.

It is important to acknowledge that NHANES is adapted for educational purposes and is not intended for formal research. Researchers conducting advanced studies should rely on the original NHANES_raw data and adhere to the prescribed sampling and weighting procedures to ensure the validity of their findings. Future research could extend the current analysis by using NHANES_raw in combination with survey weighting techniques to explore population-level health trends and disparities with greater accuracy.

C Surveys, Sampling, and Observational Data

The National Health and Nutrition Examination Survey (NHANES) represents one of the most comprehensive and influential health datasets in the United States, providing a valuable resource for understanding the health and nutritional status of the U.S. population. Conducted by the National Center for Health Statistics (NCHS), NHANES gathers both interview and physical examination data, which are combined to produce estimates of the general population's health, nutritional status, and related behaviors. The methodology behind the NHANES survey is complex and carefully designed to ensure that the data collected are representative of the U.S. civilian, non-institutionalized population.

The target population of NHANES consists of the civilian, non-institutionalized population of the United States, including individuals of all ages. This broad and inclusive definition ensures that the data captured by NHANES reflect the health and nutritional status of nearly every demographic group in the country. However, as a complex survey design, NHANES also uses a stratified, multistage sampling approach to select participants, ensuring that certain subgroups, particularly those with greater health disparities, are adequately represented. To achieve this goal, NHANES selects a representative sample from all counties in the U.S. Each year, a new group of PSUs (Primary Sampling Units) is chosen, which typically consist of counties or groups of contiguous counties. A total of about 12,000 people participate in each two-year cycle. This sampling method incorporates differential probabilities of selection, ensuring that groups at higher risk of health issues or underrepresentation (e.g., racial minorities, elderly individuals) are more likely to be sampled. As a result, NHANES oversamples certain subgroups, such as non-Hispanic Black individuals, Hispanic individuals, and elderly individuals over the age of 80, to improve the precision and reliability of health estimates for these populations.

The recruitment process for NHANES is critical to ensuring that the sample is representative and that response rates are maximized. The process begins by selecting households through a multi-stage sampling design. In the first stage, PSUs are chosen using probability proportional to size (PPS), ensuring that more populous areas are selected with higher likelihood. In the second stage, segments within the PSUs are selected—typically representing census blocks or other local geographic units. In the third stage, households within the segments are listed, and a sample of households is chosen. Finally, individuals are selected within these households, using a method that ensures the representation of key demographic characteristics such as age, sex, and race/ethnicity. On average, two participants are chosen per eligible household, which allows the survey to gather a substantial amount of data on family structures, as well as individual health outcomes.

The sampling approach used by NHANES is probabilistic, which means that each person in the target population has a known, non-zero probability of being selected. This is in contrast to deterministic approaches, which might introduce bias by excluding certain groups or individuals from the sample. By employing a probabilistic method, NHANES ensures that its data can be generalized to the entire U.S. population, although care must be taken when analyzing smaller subgroups, given the smaller sample sizes in these categories. One of the trade-offs of

this probabilistic sampling approach is the increased complexity of the survey design. Given the multi-stage sampling procedure, survey weights and variance estimation methods must be carefully applied during the data analysis phase to adjust for the unequal probabilities of selection and the clustering of respondents within geographic regions. These adjustments are necessary to avoid biased estimates, particularly when estimating health indicators for subgroups or when combining data across multiple survey cycles.

Non-response is a common challenge in large-scale surveys like NHANES, and its handling is critical to ensuring that the data remain unbiased and representative. Non-response in NHANES can occur at both the unit (or sample person) level and the component (or item) level. Unit non-response refers to individuals who are selected to participate in the survey but either refuse to participate or fail to respond to any part of the survey (e.g., household interview or medical examination). In the 2009-2010 cycle, for example, of the 13,272 persons eligible to participate, 21% failed to complete the household interview, and a further 2% of those who completed the interview did not participate in the medical examination. These non-responses were adjusted for using sampling weights, which are designed to reflect the number of individuals in the population represented by each sample respondent.

Component non-response, on the other hand, refers to missing data for specific questions or tests within the survey. For example, a respondent may refuse to have their blood pressure measured but complete all other examination components. NHANES addresses component non-response by treating missing values as missing and adjusting the weights accordingly. When analyzing the data, researchers are encouraged to evaluate the extent of missing data for their variables of interest and to apply reweighting or imputation techniques if necessary, particularly when more than 10% of the data for a given variable is missing. While NHANES employs rigorous methods to handle non-response, it remains a potential source of bias. Non-respondents may differ in key characteristics from respondents, such as age, gender, or health status. This is particularly true for elderly individuals or those with more severe health issues, who may be less likely to participate in the survey. Therefore, the results of NHANES should always be interpreted with caution, especially when analyzing specific subgroups or when drawing conclusions about less-represented populations.

The NHANES questionnaire and data collection process are designed to capture a comprehensive range of health, nutritional, and demographic information. The questionnaire consists of multiple sections, including detailed questions on demographic characteristics (e.g., age, sex, race, income), health behaviors (e.g., smoking, alcohol consumption, physical activity), and medical history (e.g., chronic diseases, medication use). The questionnaire also covers dietary intake, which is crucial for understanding the population's nutritional status.

The questionnaire is administered during a household interview, which is followed by a physical examination at the Mobile Examination Centers (MEC). During the MEC visit, participants undergo a range of tests, including anthropometric measurements (e.g., height, weight, BMI), blood pressure measurements, and laboratory tests (e.g., blood, urine samples). The tests performed in the MEC allow for objective measurements of participants' health, complementing the self-reported information collected during the interview.

The NHANES questionnaire is continuously updated to reflect new public health concerns and emerging health trends. For example, over time, the survey has included questions related to emerging issues such as mental health, sleep patterns, and environmental exposures. However, the questionnaire also has limitations. For example, self-reported data on health behaviors like physical activity or dietary intake may be subject to recall bias, with participants underestimating or overestimating their behaviors. Additionally, some participants may be reluctant to report sensitive information, such as smoking or alcohol consumption, which can further skew the data.

One of the major strengths of the NHANES methodology is its ability to provide nationally representative estimates of health and nutritional status across a wide range of demographic and geographic subgroups. By combining interviews with objective physical examinations and laboratory tests, NHANES offers a rich, multi-dimensional view of health in the U.S. population. The rigorous sampling and data collection procedures ensure that the data can be used for a wide variety of public health analyses, from assessing chronic disease prevalence to tracking health trends over time.

However, NHANES also has several weaknesses. The most notable is the issue of non-response, particularly in certain subgroups (e.g., elderly individuals, people with poor health). While the survey adjusts for non-response using weights, there is always the possibility that the non-respondents differ systematically from the respondents, which could lead to biased estimates. Additionally, the reliance on self-reported data can introduce biases such as social desirability bias or recall bias, which can affect the accuracy of responses on health behaviors and lifestyle factors.

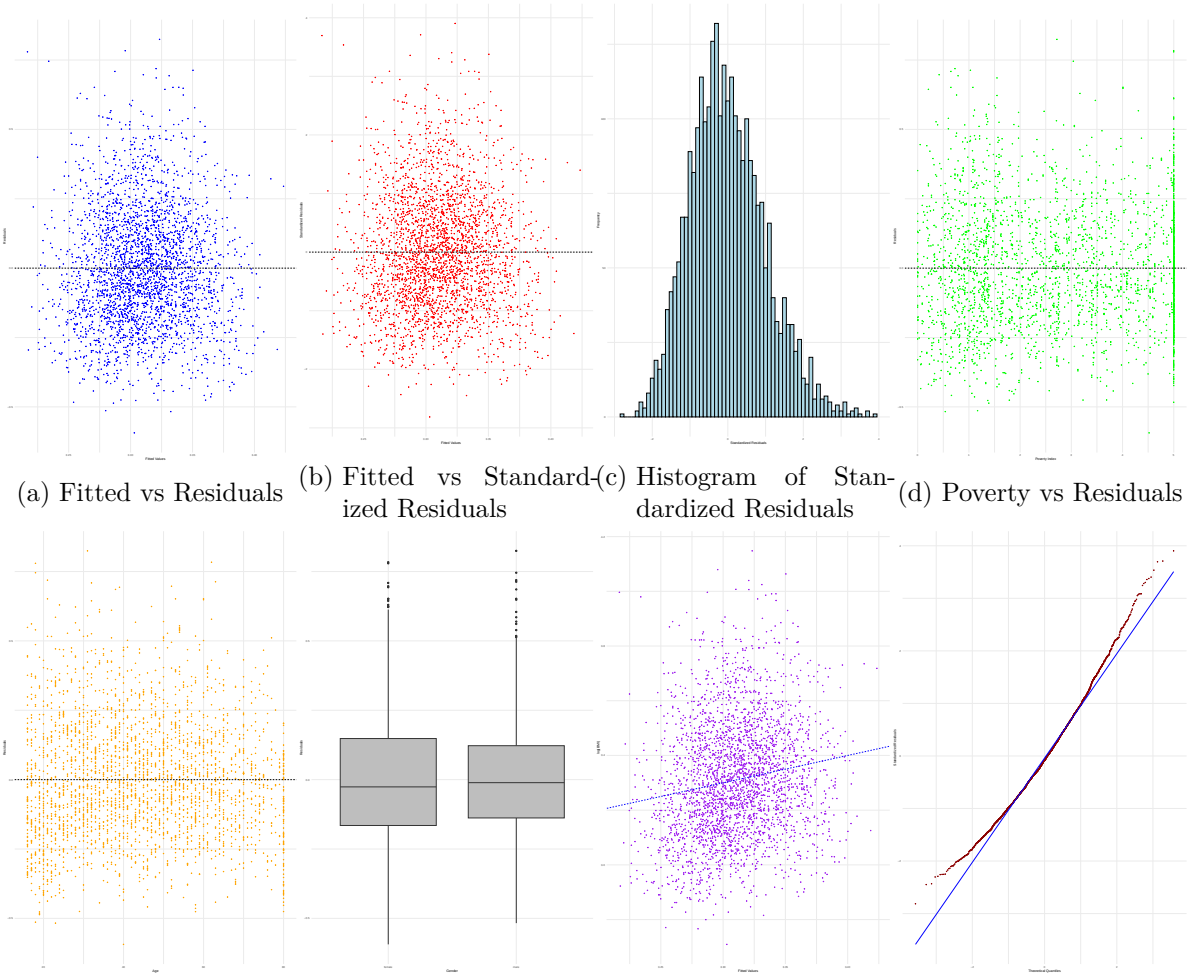
Moreover, the fact that NHANES data are released in two-year cycles means that estimates for some subgroups may not be stable, particularly for rare health conditions or smaller demographic groups. Combining data across multiple years can help mitigate this issue, but it also introduces complexities in terms of data harmonization and analytic guidelines.

In conclusion, NHANES provides a valuable and unique resource for understanding the health and nutritional status of the U.S. population. Its robust sampling methodology, comprehensive questionnaire, and use of both self-reported and objective health measures make it an indispensable tool for public health research. While the survey has several strengths, including its representativeness and ability to assess health trends over time, it also faces challenges related to non-response, data quality, and biases inherent in self-reported information. Researchers must be mindful of these issues when analyzing NHANES data, and apply appropriate statistical methods to account for potential biases. Nonetheless, NHANES remains a cornerstone of U.S. public health surveillance, providing critical insights into the health and well-being of the nation's diverse population.

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 and Predictor Relationships in the Linear Model for log(BMI).

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.