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 1, 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, age, sleep duration, and gender interact to influence BMI outcomes.

Using 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

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

activity frequency (measured in days), age, sleep duration, and gender. A multilinear regression approach was employed to quantify these relationships and identify patterns within the dataset.

The results indicate that while the correlations between 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….

2 Data

2.1 Overview

The dataset used in this analysis is derived from the US National Health and Nutrition Examination Survey (NHANES), version 2.1.0, published in July 2015. 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). The dataset contains 10,000 observations and 76 variables. 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 in the dataset, 3,573 observations remained. The dataset was prepared, cleaned, and analyzed using R (R Core Team, 2022) with the following libraries: opendatatoronto (Gelfand, 2022) for accessing the data, tidyverse (Wickham et al., 2019), dplyr (Wickham et al., 2023) for data manipulation, and

ggplot2 (Wickham, 2016) for visualizations. Additionally, knitr (Xie, 2023a) was used for report generation, and styler (Müller et al., 2024) ensured the R code was properly styled.

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

Table 1: Summary statistics for variables in the NHANES dataset.

Variable	Mean	Median	Min	Max	1st Quantile	3rd Quantile
BMI	27.975	26.950	15.020	63.300	23.500	31.300
log(BMI)	3.308	3.294	2.709	4.148	3.157	3.444
Poverty Index	3.077	3.170	0.000	5.000	1.450	5.000
Physical Activity Days	3.709	3.000	1.000	7.000	2.000	5.000
Age	43.607	43.000	16.000	80.000	29.000	56.000
Sleep Hours	6.956	7.000	2.000	12.000	6.000	8.000

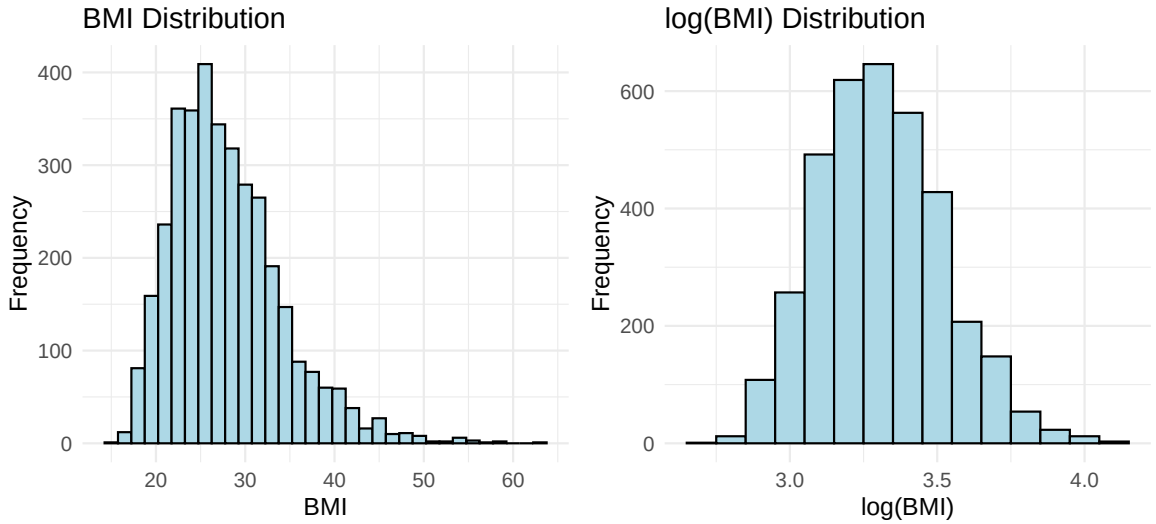
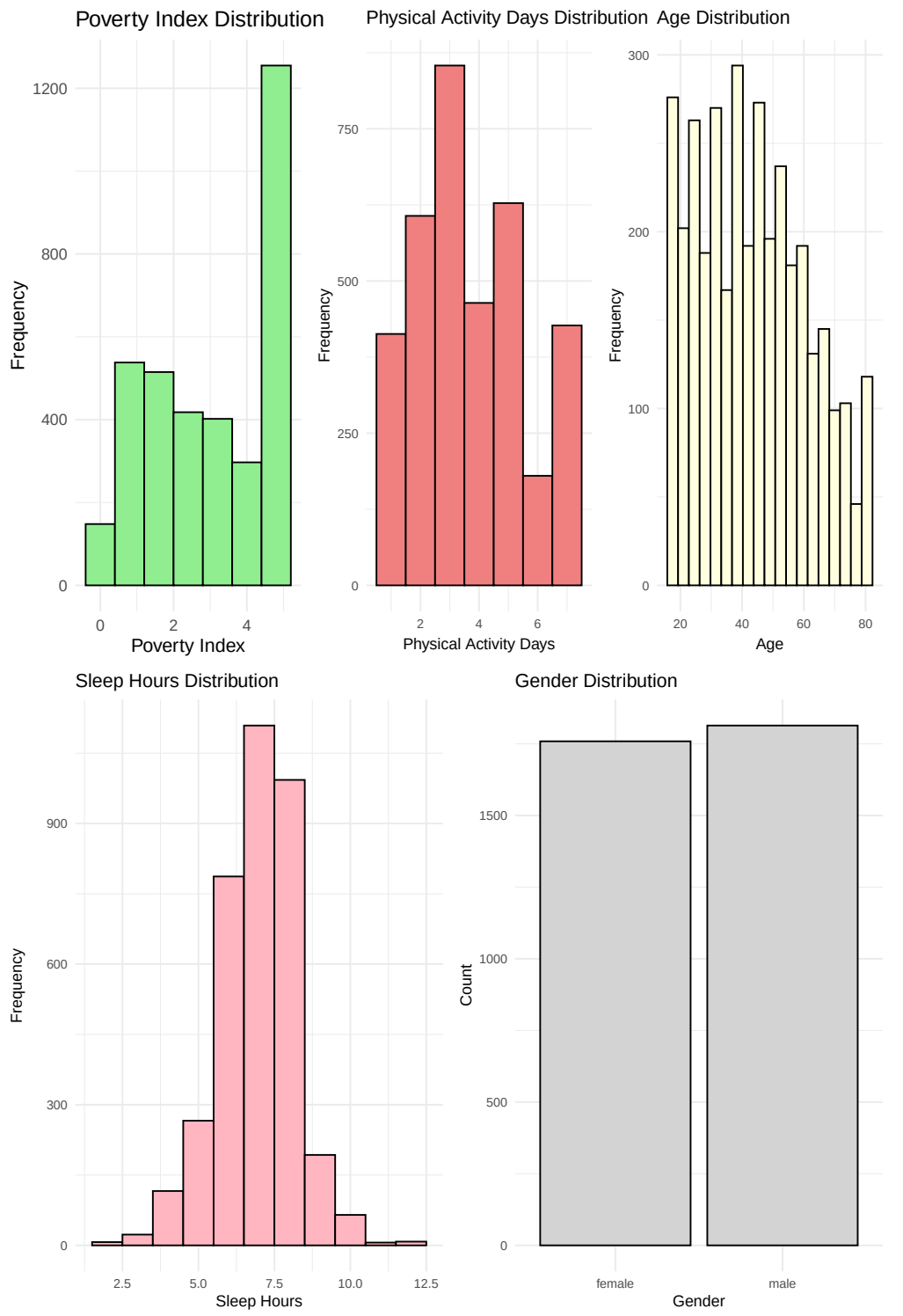


Figure 1: Distributions of BMI and log(BMI) in the NHANES dataset.

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

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.



ıtions of predictor variables in the NHANES dataset, including Poverty, Physical Activity Days, Age, Sleep Hoı

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

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 BMI:

- `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 Duration (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 Figure 1 and the histograms of predictor variables shown in Figure 3 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 Duration (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 our modelling strategy is to use multilinear regression model to investigate the relationship between log(BMI) and poverty, Physical Activity Days, Sleep Duration, gender and age. Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define $log(BMI)_i$ as the log(BMI). Then β_i are the coefficients associated with each predictor variable, which represent the change in $log(BMI)_i$ for a one-unit change in the corresponding predictor, while holding all other predictors constant.

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

$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$

We ran the model in R (R Core Team 2023) using the lm() function for linear regression, with data manipulation performed using the dplyr package and data reading via the arrow package. No specific priors were applied, as this model relies on ordinary least squares (OLS) estimation, which assumes no prior distributions for the coefficients.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

References

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