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 2, 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	28.059	27.000	15.020	63.300	23.500	31.49
log(BMI)	3.311	3.296	2.709	4.148	3.157	3.45
Poverty Index	2.940	2.910	0.000	5.000	1.340	5.00
Physical Activity Days	3.723	3.000	1.000	7.000	2.000	5.00
Age	43.850	43.000	16.000	80.000	29.000	57.00
Sleep Hours	6.926	7.000	2.000	12.000	6.000	8.00

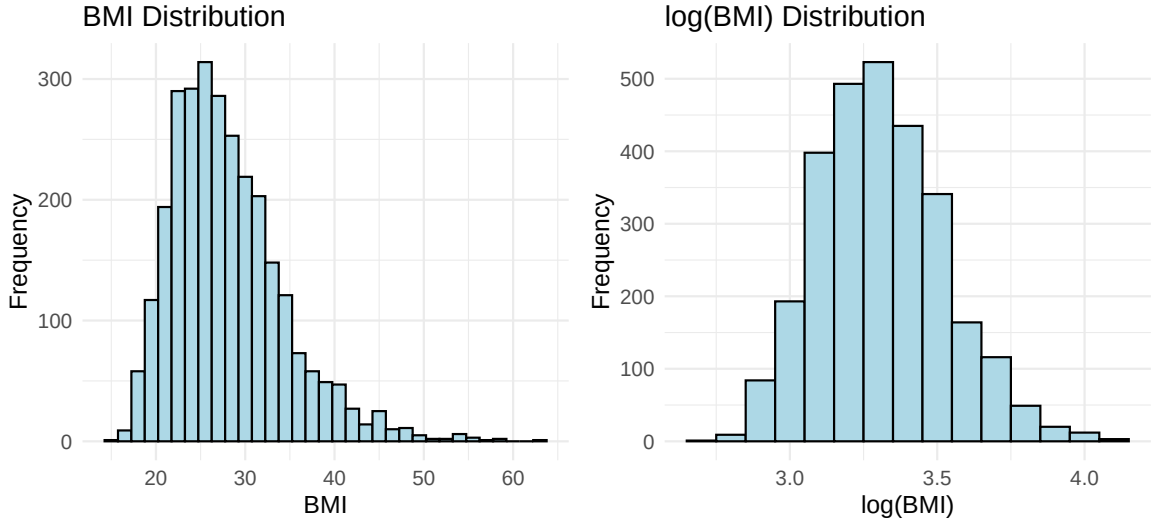
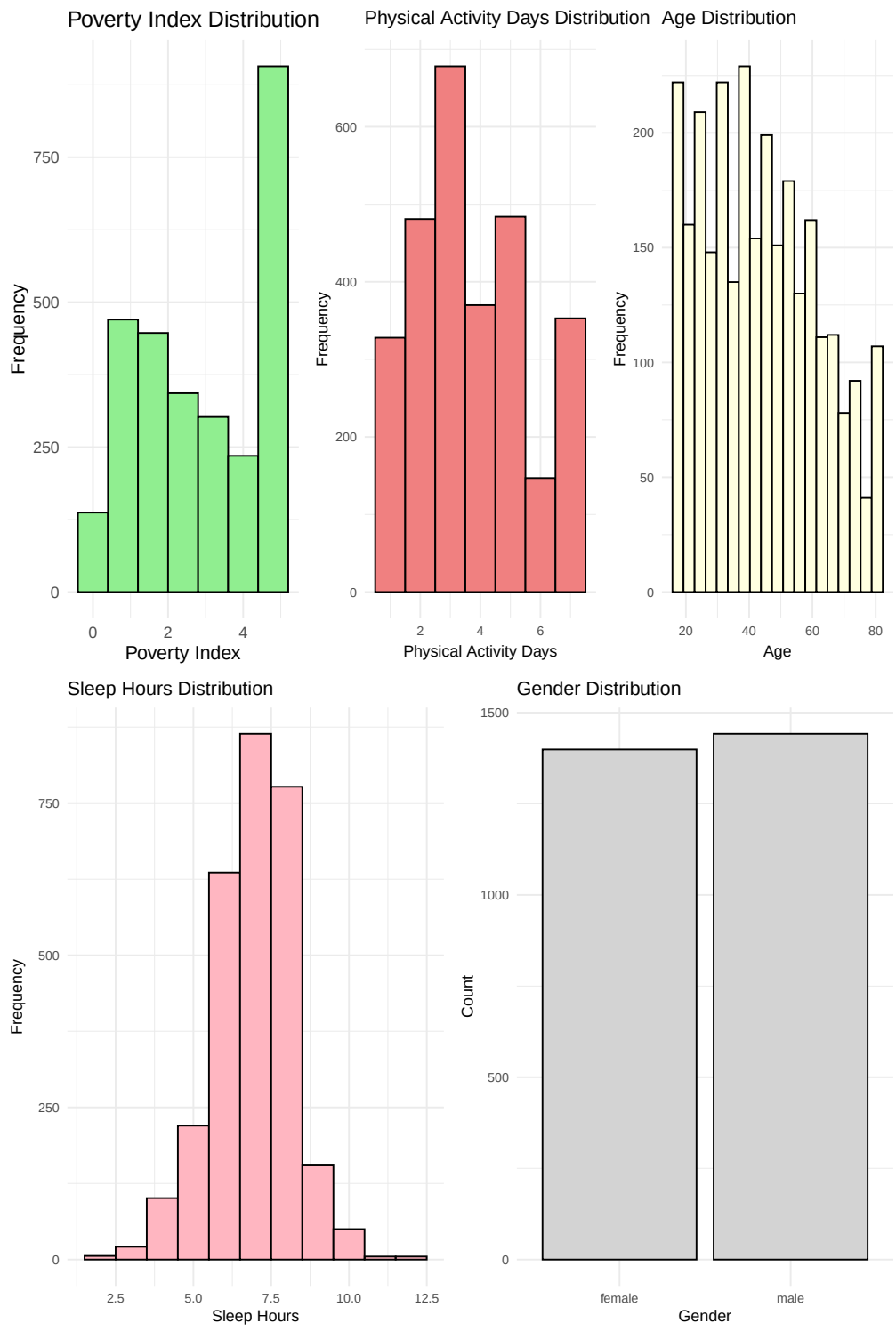


Figure 1: Distributions of BMI and log(BMI) in the NHANES dataset.

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 use a multiple linear regression model to estimate the relationship between body mass index (BMI) and various predictors, including poverty index, physical activity days, age, sleep hours, and gender. The model is designed to predict log-transformed BMI, as this transformation helps normalize the distribution of BMI and address skewness. This approach allows us to explore how these lifestyle and socio-economic factors influence BMI across the sample, using high-quality survey data from the NHANES dataset. We employ predictors like poverty, physical activity, age, sleep hours, and gender to account for key factors that are known to influence BMI, based on existing research.

We use linear regression because it is appropriate for modeling the relationship between a continuous outcome variable (log(BMI)) and predictors physical activity, sleep hours, age,

poverty, and gender. This method is straightforward to interpret, and the results could quantify the effect of predictors on BMI, making it a suitable choice for analyzing how these factors relate to BMI.

We chose to log-transform the BMI variable due to its right-skewed distribution. Log-transforming BMI allows us to better meet the assumptions of linear regression by stabilizing variance and making the relationship between BMI and predictors more linear. This is particularly important for ensuring that the model's estimates are valid and interpretable.

We included the variable 'Poverty' to account for socio-economic status, as previous research has suggested that lower-income individuals tend to have higher BMI levels (Webber et al., 2023). 'PhysActiveDays' was included to capture the effect of physical activity on BMI, as increased physical activity is typically associated with a lower BMI(Webber et al., 2023). Age was included as a predictor to account for the natural changes in BMI that occur as individuals age. Sleep hours were added because there is evidence that insufficient sleep can contribute to weight gain (Ekstedt et al., 2013). Gender was included as a predictor due to well-established gender differences in BMI, with males typically having lower BMI than females (Longo-Silva et al., 2023).

We used the default settings in lm() from the stats package, which assumes normally distributed errors and does not apply any specific prior distributions, as this method does not require priors in the same way that Bayesian methods do. This approach provides a reliable way to assess the linear relationship between the predictors and BMI. However, we acknowledge that other methods, such as Bayesian regression or generalized linear models, could be used for more complex modeling, but this approach was selected for its simplicity and interpretability given the goals of the analysis.

Initially, we considered including interaction terms between predictors (such as between poverty and physical activity). However, when these interaction terms were added, the R-squared value decreased, indicating a reduced fit to the data. Additionally, the residual plot showed a noticeable concentration in the residuals, suggesting potential model misspecification or overfitting. As a result, we decided to exclude interaction terms to maintain a more interpretable and well-fitted model. Future analyses could revisit this approach with a larger dataset or alternative modeling strategies to better account for potential interactions.

By using this model, we aim to better understand the factors that influence BMI and provide insights into potential public health interventions that could target lifestyle changes and socio-economic factors to reduce BMI and related health risks.

4 Results

Our results are summarized in Table 2.

Table 2: Summary statistics for variables in the NHANES dataset.

Table 2: Summary of the Linear Model for Log(BMI)

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

```
# Scatter plot for Age vs. log(BMI)
plot_age_bmi <- ggplot(nhanes_data, aes(x = Age, y = log(BMI))) +
  geom_point(size = 0.3) +
  geom_smooth(method = "lm", col = "red", se = FALSE) +
  labs(title = "Age vs. log(BMI)", x = "Age", y = "log(BMI)") +
  theme_minimal() +
  theme(plot.title = element_text(size = 12), axis.title = element_text(size = 10), axis.text

# Scatter plot for Gender vs. log(BMI)
plot_gender_bmi <- ggplot(nhanes_data, aes(x = Gender, y = log(BMI))) +
  geom_point() +
  geom_smooth(method = "lm", col = "red", se = FALSE) +
  labs(title = "Figure17: Gender vs. log(BMI)", x = "Gender", y = "log(BMI)") +
  theme_minimal() +
  theme(plot.title = element_text(size = 14), axis.title = element_text(size = 12), axis.text

# Scatter plot for Poverty vs. log(BMI)
plot_poverty_bmi <- ggplot(nhanes_data, aes(x = Poverty, y = log(BMI))) +
  geom_point() +
  geom_smooth(method = "lm", col = "red", se = FALSE) +
  labs(title = "Figure18: Poverty vs. log(BMI)", x = "Poverty", y = "log(BMI)") +
  theme_minimal() +
  theme(plot.title = element_text(size = 14), axis.title = element_text(size = 12), axis.text

# Scatter plot for SleepHrsNight vs. log(BMI)
plot_sleep_bmi <- ggplot(nhanes_data, aes(x = SleepHrsNight, y = log(BMI))) +
  geom_point() +
  geom_smooth(method = "lm", col = "red", se = FALSE) +
  labs(title = "Figure19: SleepHrsNight vs. log(BMI)", x = "SleepHrsNight", y = "log(BMI)") +
```

```r
  theme_minimal() +
  theme(plot.title = element_text(size = 14), axis.title = element_text(size = 12), axis.text

# Scatter plot for PhysActiveDays vs. log(BMI)
plot_physactive_bmi <- ggplot(nhanes_data, aes(x = PhysActiveDays, y = log(BMI))) +
  geom_point() +
  geom_smooth(method = "lm", col = "red", se = FALSE) +
  labs(title = "Figure20: Physical Active Days vs. log(BMI)", x = "PhysActiveDays", y = "log
  theme_minimal() +
  theme(plot.title = element_text(size = 14), axis.title = element_text(size = 12), axis.text

# Combine all plots into a single figure using patchwork
library(patchwork)
combined_plot <- (plot_age_bmi | plot_gender_bmi) /
                 (plot_poverty_bmi | plot_sleep_bmi) /
                 plot_physactive_bmi +
  plot_annotation(
    title = "Diagnostic and Predictor Relationships in the Linear Model",
    caption = "Figure: Diagnostic and variable relationships from the linear model for log(BM
    theme = theme(plot.caption = element_text(hjust = 0.5, size = 10))
  )

# Display the combined plot
combined_plot
```

```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```

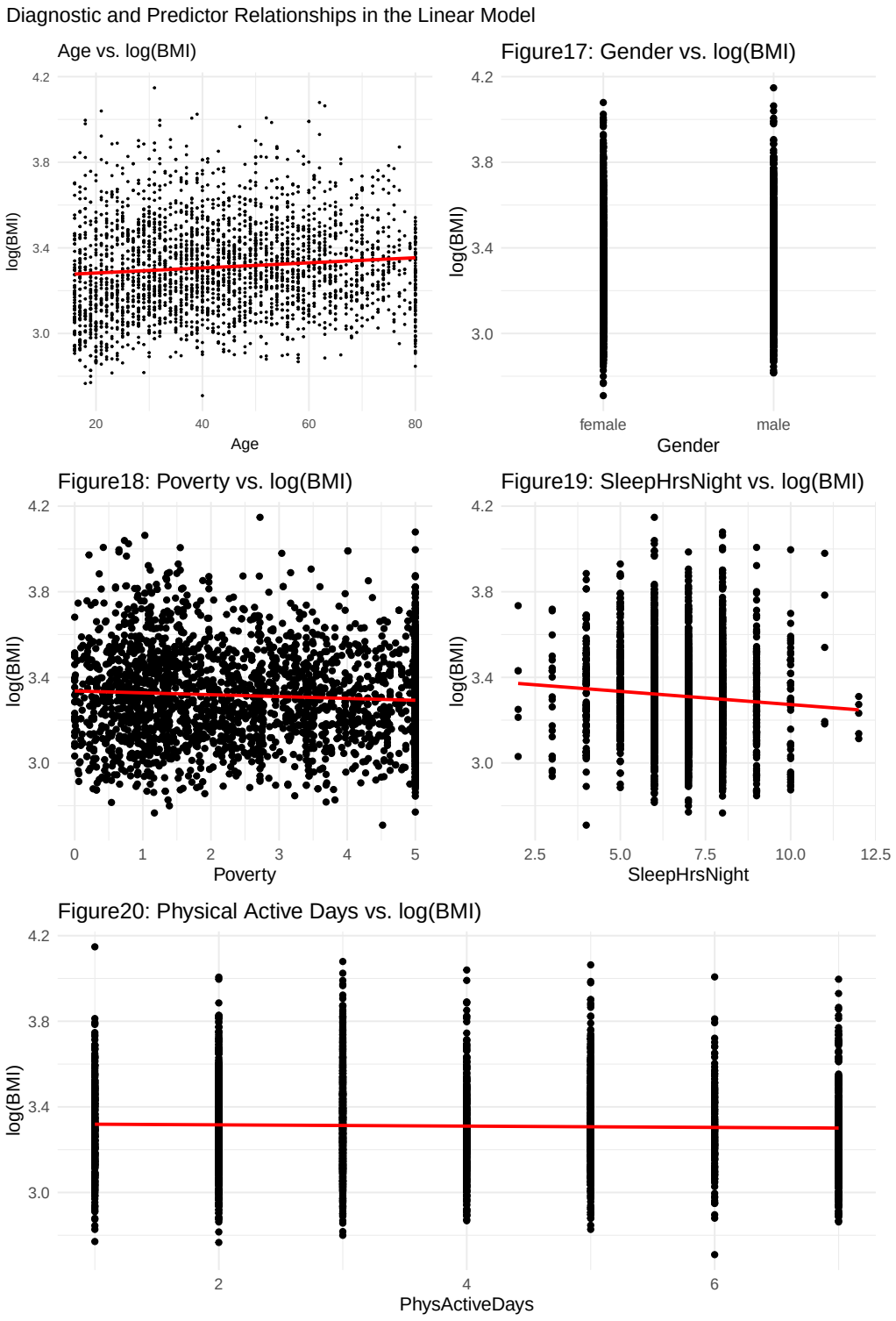


Figure: Diagnostic and variable relationships from the linear model for log(BMI).

Figure 3: Scatter plots of predictors vs. log(BMI) with fitted regression lines.

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

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.

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

The analysis of linear regression assumptions shows that all conditions are met. The fitted versus residual plot (Figure 8) 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 (Figure 9) and the predictor variables versus residuals plots (Figures 11 to 13) support these conclusions. The QQ plot (Figure 15) shows standardized residuals following a straight diagonal line, while the histogram (Figure 10) indicates a distribution close to ~N(0,1), confirming the normality assumption of errors. Lastly, the response variable versus fitted values plot (Figure 14) 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 (Figures 16 to 20) show weak correlations between predictor variables and log(BMI), with log(BMI) approximated by

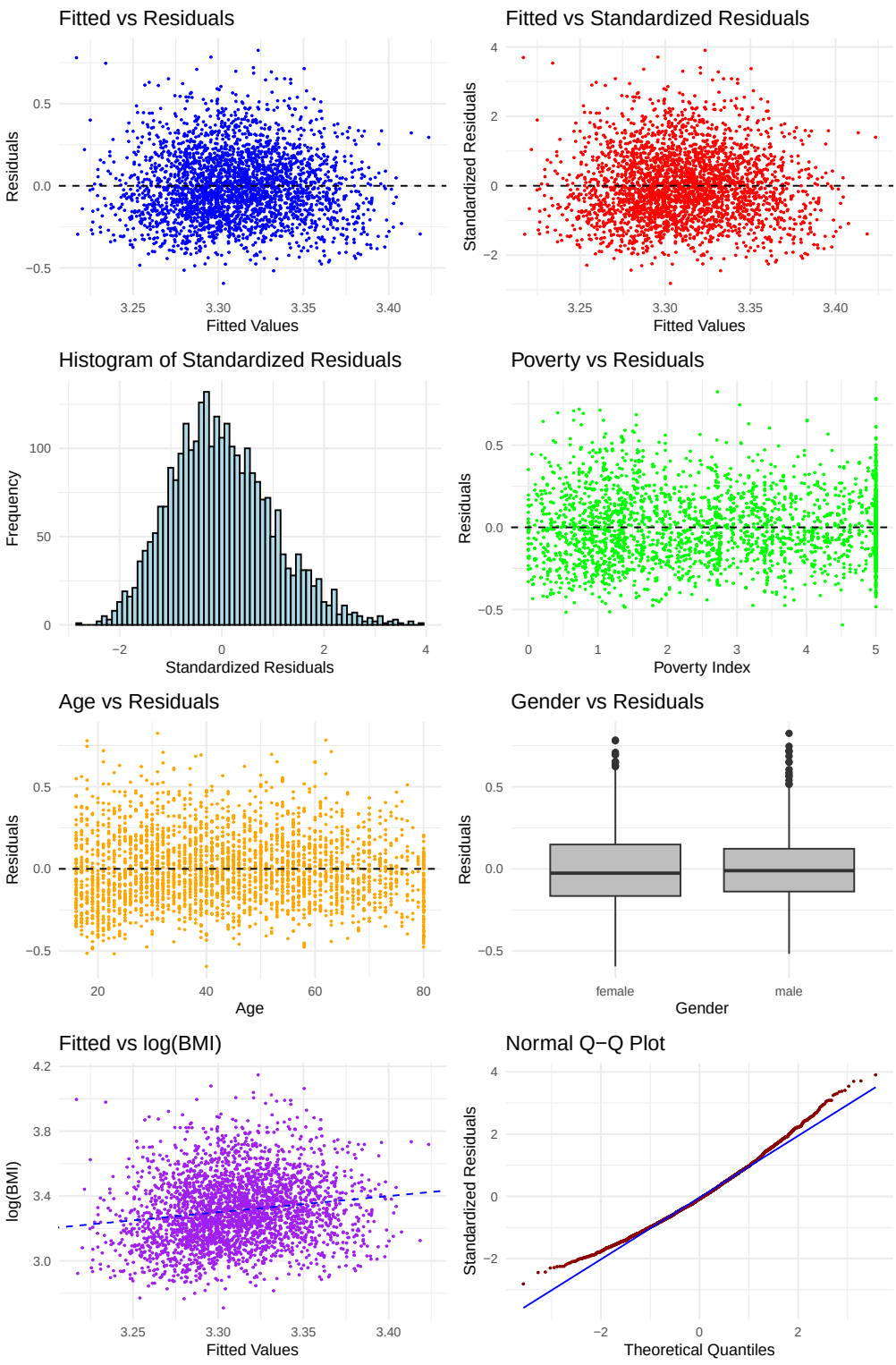


Figure 4: Diagnostic and Predictor Relationships in the Linear Model for log(BMI).

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/.