Sleep Sufficiency Factor Analysis

Xinxiang Gao

University of Toronto
xinxiang.gao@mail.utoronto.ca

1 Introduction

Sleep is a fundamental biological need that is pivotal in promoting physical and mental well-being. In today's fast-paced society, where demands often encroach upon the time allocated for rest, the importance of sufficient and quality sleep cannot be overstated. Theoretically, as proposed by Krueger et al [4], sleep initiation is driven by altered neuronal activity patterns mediated by growth factors, leading to synaptic remodelling, and coordinated by both sleep regulatory circuits and activational-projection systems throughout the brain. In other words, sleep is crucial for the body to repair, recharge, and regulate brain remodelling. According to [5], Adults between 18 to 70 years old are suggested at least seven hours of sleep on average, highlighting the importance of sufficient sleep time.

Despite sleep takes a crucial role in our health, insufficient sleep caused by sleeping disorders is becoming a huge concern in modern life. NHLBI estimates that between 50 million and 70 million individuals are currently affected by persistent sleep disorders. These disorders impact an individual's quality of life and carry substantial societal and economic impacts. According to the work of [3], sleep deficiency can lead to a range of impairments including decreased cognitive function, emotional disturbances, increased risk of chronic diseases, and performance deficits comparable to alcohol intoxication. Furthermore, the randomized control trial proposed by Freeman et al. [2], provides strong evidence that insomnia is a causal factor in the occurrence of psychotic experiences and other mental health problems.

This study aims to investigate the factors that correlate a sufficient sleep time for adults. We initially hypothesize that sleeping quality, alcohol consumption, mental depression and heart disease are connected to whether adults have a sufficient sleep duration on average. Understanding these relationships is crucial for informing interventions aimed at promoting healthier sleep habits and mitigating the adverse effects of sleep disorders.

2 Methods

2.1 Data Retrieval

This study is conducted utilizing the NHANES data collected from 2017-March to 2020 Pre-pandemic. The National Health and Nutrition Examination Survey

(NHANES) is a program conducted by the Centers for Disease Control and Prevention (CDC) to assess the health and nutritional status of adults and children in the United States. NHANES collects data through interviews, physical examinations, and laboratory tests, making it a valuable resource for studying various health-related outcomes.

From the variety of datasets available, we retrieved 5 datasets of interest:

- `P_DEMO`: Demographics data of the participants. Document link
- `P_SLQ`: Sleep Disorders Questionnaire response. Document link
- `P_ALQ`: Alcohol use Questionnaire response. Document link
- `P_DPQ`: Mental Health - Depression Screener Questionnaire response. Document link
- `P_CDQ`: Cardiovascular Health Questionnaire response.Document link

The datasets are retrieved using `nhanesA` package in R, ensuring reproductivity and relief in data preprocessing.

2.2 Preprocessing

The documentation provided with each dataset on the website was utilized to understand the variables and their meanings. Variable names were modified from question ID to a summary of the content of the corresponding questions. while retaining some numerical values to represent the extent of the variables. More specifically, in the Mental Health-Depression Screener, a scale of 0 to 3 indicates the severance of having little interest in doing things, from ”not at all” to ”almost every day”, and their values are irreplaceable for further feature calculations.

For some of the questions, answers that indicate ”refuse to answer” or ”do not know” are modified to NA values. The missing value amount is huge by the nature of a questionnaire and mutation from negligible answers to NA values. To tackle this challenge, rather than gathering all available data, we opt to exclude questions with a high missing value rate and merge the sleep data with the relevant factors individually. Subsequently, we conduct our analysis while either excluding observations with missing values in the key variables or setting missing values as a separate category.

The potential outliers are well addressed by the datasets. For example, if a person drinks more than 15 drinks (in which a drink represents the quantity of a 12 oz. beer, a 5 oz. glass of wine, or one and a half ounces of liquor) every day in the past 12 months, this observation value would be a static 15.

Then besides the overall data cleaning, for each dataset, we modify as follows:

For the Sleep Disorders dataset, the average sleeping duration is calculated from their sleeping duration on weekdays and weekends with a weight of 5/7 and 2/7, respectively, and the answer to the question ”How often feel overly sleepy during the day?” is kept as it serves as an indicator of the sleeping quality. Then, as mentioned in [5], we define a cut point of a participant having enough sleep by verifying whether their average sleep duration is greater than or equal to 7 hours.

For the Alcohol Use dataset, we found that some variables were too specific and didn't contribute significantly to our analysis. Additionally, we observed that certain variables were already captured within broader variables. For instance, the question "Past 12 months how often drink alcoholic beverages" exhibits a similar statistic to "Average number of alcoholic drinks per day in the past 12 months." Therefore, to streamline our analysis and avoid redundancy or multicollinear problems, we excluded these specific variables and retained either the broader ones that encompassed their information or the more complete ones that had fewer NA values.

For the Mental Health-Depression Screener, a total score ranging from 0 to 27 was calculated based on complete responses to symptom questions. [1]

In the Cardiovascular Health dataset, angina classification into grade 1 and grade 2 is determined using the Rose questionnaire criteria as in [6]. For our analysis, individuals who have angina in either grade are categorized as having angina. [2]

For the demographic data, education level, age and gender are retrieved for further analysis. As our hypothesis focuses on adults, we opt to exclude observations with an age of less than 18 or greater than 60.

The combined dataset has an observation count of 3604, with 11 variables of interest.

2.3 Model

For the model methodology, variable selection methods for binary classification algorithms such as stepAIC for logistic regression, which choose variables according to AIC, random forest and extreme gradient boosting are used. Instead of choosing the best model according to accuracy metrics, we propose to infer or obtain the relative importance of each variable from the results of the model, and a mutually important variable would suggest a strong correlation between enough sleep with its value. In the case of gradient boosting, the grid search explored a range of hyperparameters, including learning rates from 0.01 to 0.2, maximum tree depths of 3, 6, and 9, and optimizing model performance using cross-validation with 5 folds.

[1] Pre-defined cut-points were applied to assess depression severity, with scores less than 10 indicating rare occurrence of major depressive episodes, and scores of 15 or greater suggesting the presence of a major depressive episode. [1]. However, numerical scores tend to perform more significantly than the extent in all models, indicating the cut-points are less applicable in this analysis.

[2] Adding the dataset to the current analysis had unforeseen consequences. The updated GLM model revealed statistically insignificant p-values for all variables, prompting us to exclude the dataset altogether. This outcome may stem from the extreme imbalance in the data, with few participants exhibiting signs of angina. Future research could focus solely on analyzing angina and sleep duration. Additionally, oversampling the data may help address the imbalance issue and provide more reliable results.

3 Results

3.1 Data

As depicted in Figure 1, the plot illustrates the distribution of some significant factors. A balanced distribution is observed across gender and age groups. Furthermore, approximately one-quarter of the participants experience sleep-related issues. Additionally, the majority of participants have attained some level of college education, report occasionally feeling sleepy, and have a depression score of 0. Notably, the distribution of depression scores appears to be imbalanced, which may be attributed to several factors such as responses marked as "Refused" or "Don't know" being treated as NA values, or a significant proportion of participants reporting little symptoms of depression across all questions.

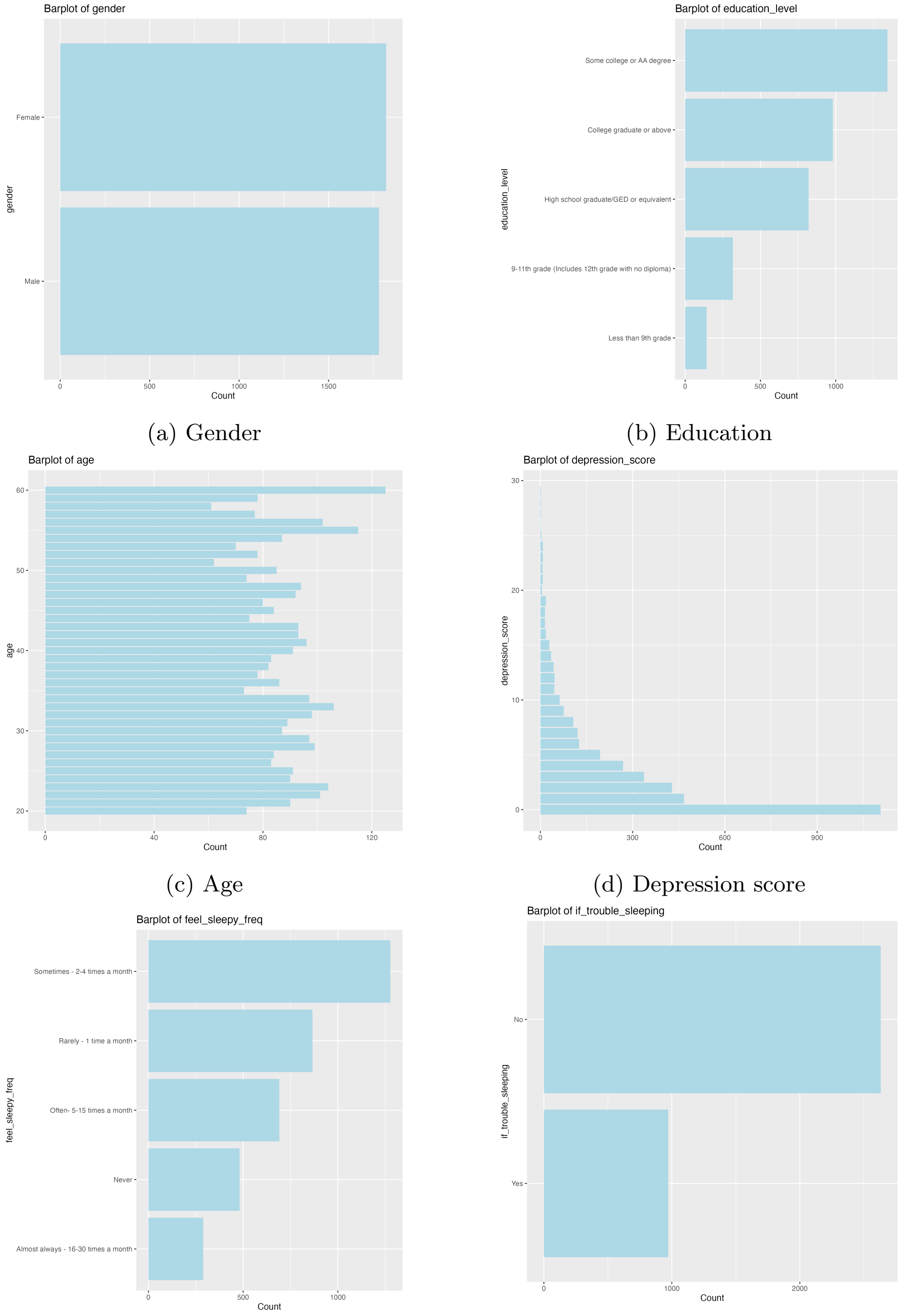
As shown in Figure 2a, the plot illustrates the distribution of whether individuals had enough sleep. Notably, this distribution bears resemblance to that shown in Figure 1f. Upon further investigation through both exploratory analysis and modeling, it becomes apparent that the presence of sufficient average sleep duration appears indifferent, as evidenced by the similar ratios observed in Figure 2b. Surprisingly, knowing that a participant has sleep problems contributes little to predicting the sufficiency of sleep duration.

3.2 Model

Generalized Linear Model (GLM): The GLM analysis revealed several significant predictors of sufficient sleep duration among adults. Specifically, gender was found to be a significant predictor, with males exhibiting lower odds of having enough sleep compared to females. Additionally, education level emerged as a significant factor, with individuals who attained a college degree or higher education level showing increased odds of sufficient sleep duration. Conversely, no significant associations were observed for individuals with a high school diploma or equivalent, less than a 9th-grade education, or some college or an associate's degree.

Age was identified as a significant predictor, indicating that for each one-unit increase in age, the log-odds of having enough sleep decreased by 0.014276, holding all other variables constant. This finding suggests that as individuals age, the likelihood of experiencing sufficient sleep decreases. Furthermore, the frequency of snoring and feeling sleepy were found to be significant predictors of sufficient sleep duration, with individuals who reported never snoring or feeling sleepy exhibiting higher odds of having enough sleep compared to those who reported occasional or frequent occurrences.

Moreover, depression score was found to be negatively associated with sufficient sleep duration, indicating that higher depression scores were associated with lower odds of having enough sleep. Overall, these findings underscore the multifaceted nature of factors influencing sleep duration among adults, encompassing demographic characteristics, mental health indicators, and sleep-related behaviors.



(a) Gender

(b) Education

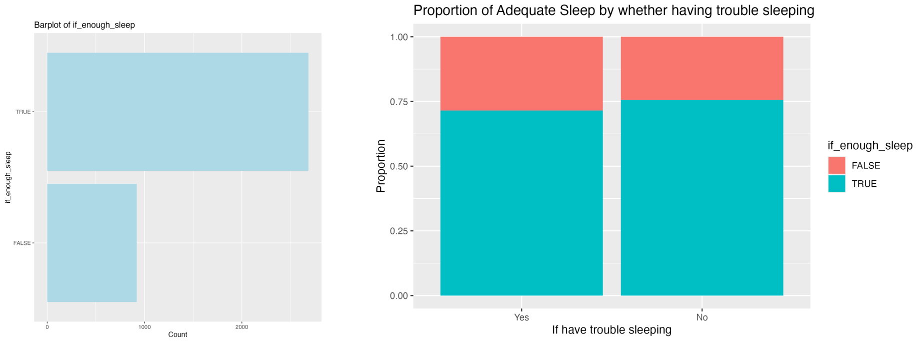
(c) Age

(d) Depression score

(e) Feel sleepy frequency

(f) Have trouble sleeping

Fig. 1: Overall distribution



(a) If enough sleep

(b) If enough sleep by having trouble sleeping

Random Forest: In the Random Forest model, the top six variables identified as most important predictors of sufficient sleep duration were age, frequency of alcohol consumption in the last year, depression score, frequency of consuming more alcohol, education level, and frequency of feeling sleepy. These variables collectively provide insight into the multifaceted nature of factors influencing sleep duration, encompassing both demographic and behavioral aspects, as well as mental health indicators.

XGBoost: Based on the best tune obtained from the grid search, the optimal hyperparameters for the XGBoost model are as follows:

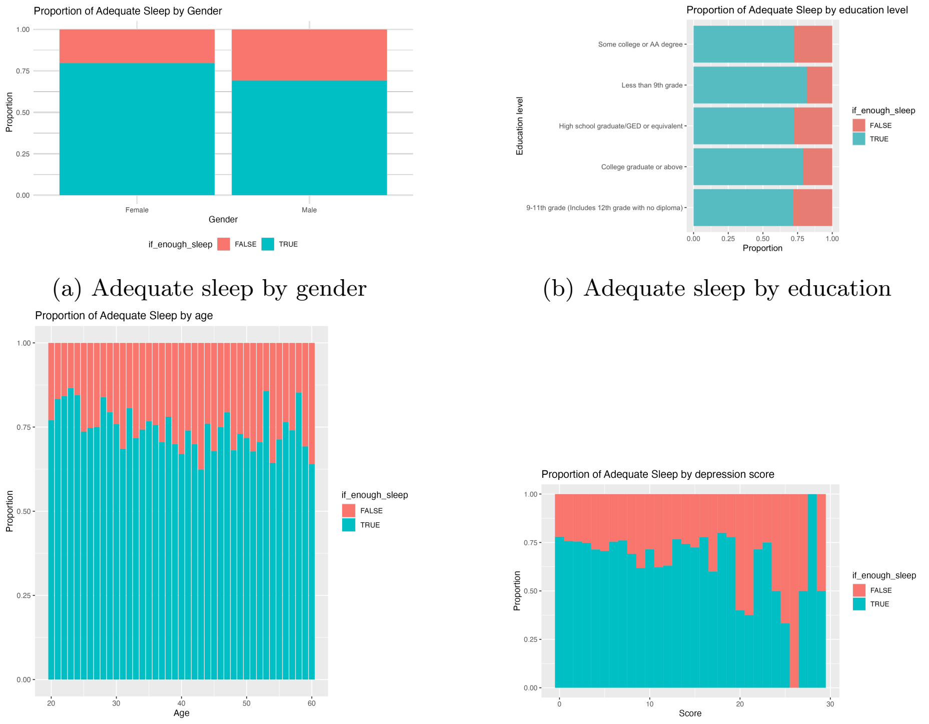
- Number of boosting rounds (`nrounds`): 500
- Maximum tree depth (`max_depth`): 3
- Learning rate (`eta`): 0.01
- Gamma: 0
- Column subsampling ratio by tree (`colsample_bytree`): 0.6
- Minimum sum of instance weight needed in a child (`min_child_weight`): 1
- Subsample ratio of the training instances (`subsample`): 1

These hyperparameters represent the configuration that yielded the highest performance for the XGBoost model in predicting sufficient sleep duration among adults. By utilizing these optimized settings, the XGBoost model can effectively capture the complex relationships between predictor variables and the likelihood of having enough sleep, contributing to a more accurate and reliable predictive model.

Similarly, in the XGBoost model, age emerged as a prominent predictor of sufficient sleep duration, along with gender, depression score, frequency of consuming more alcohol in the last 30 days, education level, and frequency of feeling sleepy. These variables highlight the importance of considering demographic characteristics, mental health status, and lifestyle behaviors in understanding sleep patterns and behaviors among adults.

Overall, the results from the three models underscore the complex interplay of various factors in influencing sleep duration among adults, emphasizing the need for a comprehensive approach that accounts for both individual characteristics and behavioral patterns.

4 Conclusions and Summary



(a) Adequate sleep by gender

(b) Adequate sleep by education

(c) Adequate sleep by age

(d) Adequate sleep by depression score

Fig. 3: Overall adequate sleep distribution

As proposed by the variable selection from some of the three models which aligns with our initial hypothesis, we take five variables: `gender`, `education`, `age`, `depression_score`, 3 `feel_sleepy` 4 and compare their association with whether the participants obtain a sufficient sleep duration on average.

- `gender`: In our sample, individuals who identify as male tend to sleep less, which aligns with the interpretation of the age coefficient in the GLM model.
- `education`: Our sample suggests that participants who have completed less than grade 9 or have a college degree tend to obtain more sufficient sleep.

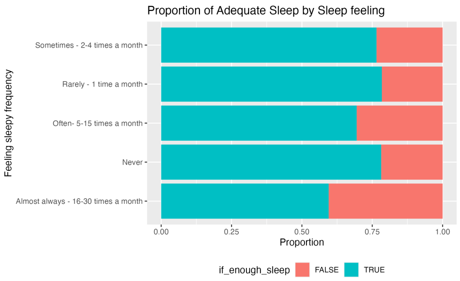


Fig. 4: Proportion of adequate sleep by feeling sleepy frequency

- `age`: We observe variation in sufficient sleep across different age groups. While it is challenging to identify general patterns, it is notable that individuals around the ages of 40 and 26 seem to have difficulty obtaining sufficient sleep.
- `depression_score`: We notice a significant decrease in sufficient sleep when the depression score is around 20, which could indicate a major depressive episode as defined in [1].
- `feel_sleepy`: People who often or almost always feel sleepy during the day tend to correlate with the lack of sufficient sleep as well.

These observations highlight the importance of various factors that correlate with the likelihood of obtaining sufficient sleep. Understanding these factors can help identify individuals who may be at risk of sleep deprivation and inform interventions to promote better sleep hygiene and overall well-being.

However, it's essential to acknowledge a notable limitation of this study, which pertains to the potential biases associated with self-report data, particularly as questionnaire data constituted a significant portion of our dataset. Self-reported sleep duration and related variables are susceptible to recall bias, social desirability bias, and inaccuracies, potentially leading to measurement error. Future research incorporating objective measures of sleep, such as actigraphy or polysomnography, could provide complementary insights and enhance the validity and reliability of sleep-related assessments.

References

1. Ehde, D.M.: Patient Health Questionnaire, pp. 1883–1885. Springer New York, New York, NY (2011). https://doi.org/10.1007/978-0-387-79948-3$_2$002, `https://doi.org/10.1007/978-0-387-79948-3_2002`
2. Freeman, D., Sheaves, B., Goodwin, G.M., Yu, L.M., Nickless, A., Harrison, P.J., Emsley, R., Luik, A.I., Foster, R.G., Wadekar, V., et al.: The effects of improving sleep on mental health (oasis): a randomised controlled trial with mediation analysis. The Lancet Psychiatry **4**(10), 749–758 (2017)
3. Killgore, W.D.: Effects of sleep deprivation on cognition. Progress in brain research **185**, 105–129 (2010)

4. Krueger, J.M., Obál, F., Fang, J.: Why we sleep: a theoretical view of sleep function. Sleep Medicine Reviews **3**(2), 119–129 (1999). https://doi.org/https://doi.org/10.1016/S1087-0792(99)90019-9, `https://www.sciencedirect.com/science/article/pii/S1087079299900199`
5. Panel, C.C.: Recommended Amount of Sleep for a Healthy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society. Sleep **38**(6), 843–844 (06 2015). https://doi.org/10.5665/sleep.4716, `https://doi.org/10.5665/sleep.4716`
6. Rose, G.A.: The diagnosis of ischaemic heart pain and intermittent claudication in field surveys. Bulletin of the World Health Organization **27**(6), 645 (1962)