Decoding Sleep Disorders Through Self-Reported Patterns*

A Logistic Regression Approach to the NHANES 2017-March 2020 Sleep Data

Terry Tu

April 14, 2024

This study investigates the potential relationship between individuals' sleep patterns and the self-reported incidence of sleep troubles as confirmed by a medical professional. Using logistic regression analysis on data from the NHANES 2017-March 2020 Sleep Disorders dataset, we examine how various factors such as sleep duration on weekdays and weekends, frequency of snoring, and daytime sleepiness are associated with the likelihood of reporting sleep problems to a doctor. The findings aim to shed light on the predictive value of self-monitored sleep behaviors in identifying individuals who may require medical attention for sleep-related issues. This analysis contributes to the broader understanding of sleep health and its complex interactions with daily functioning.

Table of contents

*Code and data supporting this analysis are available at https://github.com/TEJMaster/Sleep-Disorder-Analysis

1 Introduction

The rhythm of our nightly rest is more than a personal habit; it's a public health pulse that captures the essence of well-being in our fast-paced society. The increasing prevalence of sleep disorders and their impact on daily life and overall health has become a point of societal concern, akin to the growing conversation around mental health and lifestyle diseases. The intimate link between quality of sleep and the vitality of an individual's life prompts a closer examination of sleep patterns within the populace. Drawing on the rich dataset from the National Health and Nutrition Examination Survey (NHANES) for the years 2017 to March 2020, our research delves into the self-reported instances of sleep disturbances and their association with various sleep behaviors.

This study is an explorative journey into the silent epidemic of sleep disorders that plague modern society, affecting productivity, mental health, and long-term wellbeing. The aim is to uncover the underlying patterns of sleep behavior that correlate with the reports of sleep troubles to medical professionals, thereby piecing together the nocturnal puzzle of restless societies. We explore the quantitative relationship between self-reported snoring frequency, feelings of daytime sleepiness, and the regularity of sleep hours during weekdays and weekends with the likelihood of reporting sleep issues to a doctor (CDC 2021).

Our analysis hinges on the application of logistic regression to the NHANES dataset, which presents a comprehensive view of American sleep habits. By interpreting the nuances of this rich dataset, we aspire to illuminate the factors that signal the need for medical attention in the domain of sleep health. The outcome of our investigation is poised to provide a scaffold for healthcare professionals and policymakers to base early intervention strategies, aiming to cultivate a well-rested population.

Following this introduction, the structure of the paper is laid out to facilitate a coherent flow of information and analysis. Section 2 (Data) provides a meticulous breakdown of the NHANES dataset, elucidating the data cleaning process and offering a descriptive overview of the key variables. Section 3 (Model) details the logistic regression model's design and the rationale behind the choice of predictors. Section 4 (Result) presents the findings, interpreted with precision and caution, alongside graphical representations for clarity. Concluding the paper, Section 5 (Discussion) reflects on the broader implications of the study, acknowledging limitations and proposing avenues for future research.

2 Data

2.1 Raw Data

The dataset underpinning this analysis is derived from the National Health and Nutrition Examination Survey (NHANES), spanning from 2017 to March 2020. This public dataset in-

cludes responses from participants regarding their sleep patterns, incorporating 10,195 records initially. After meticulous data cleaning, the dataset for analysis stands at 10,031 records, encapsulating variables critical to our research: the respondent's ID, usual sleep and wake times on both weekdays and weekends, total sleep duration, frequency of snoring, incidence of breathing pauses during sleep, and self-reported communication of sleep troubles to a health professional. The dataset provides a snapshot of Americans' sleep behaviors before the disruption caused by the COVID-19 pandemic (CDC 2021).

The survey participants' ages range widely, reflecting the diversity of the American population. Variables are finely tuned to capture the multifaceted nature of sleep, encompassing aspects such as duration, disruptions, and subjective experiences of daytime sleepiness. The NHANES protocol ensures that this dataset is a robust and reliable source of information, adhering to stringent ethical standards and data collection methods, as detailed in the NHANES Analytic Guidelines. For further information on the data cleaning specifics and validation checks, please see the supplementary material in Section 6.1.

2.2 Data Analysis Tools

Our statistical exploration was conducted within the R programming environment (R Core Team 2022), leveraging its comprehensive ecosystem for data analysis. We utilized the tidyverse collection of R packages (Wickham et al. 2019) to streamline our data processing tasks. The ggplot2 package (Wickham 2016) was instrumental in crafting insightful visualizations that articulated the intricate relationships within our data. The dplyr package (Wickham et al. 2022) provided a syntax that facilitated the manipulation and transformation of our dataset, enabling us to prepare the data effectively for logistic regression analysis. Data importation was efficiently handled by the readr package (Wickham, Hester, and Bryan 2022), known for its quick and user-friendly approach to reading tabular data. Navigational simplicity within our project's directories was achieved with the here package (Müller 2020), which reliably managed file paths without the need for manual path setting. The reproducibility of our research was enhanced by the knitr package (Xie 2014), which seamlessly wove R code into our report, ensuring that our findings are transparent and replicable. For tabular data presentation, kableExtra(Zhu 2021) offered a suite of customization options that enhanced the readability and aesthetic appeal of our tables. The logistic regression model was developed using core functions in R, which provide robust methods for estimating the effects of various predictors on a binary outcome.

2.3 Variable Description

Weekday Sleep Duration (SLD012): This variable measures the total number of hours respondents usually sleep on weekdays or workdays, with values ranging from 2 to 14 hours. It provides insight into their sleep patterns during the typical workweek.

Weekend Sleep Duration (SLD013): Similar to the weekday sleep duration, this variable represents the total number of hours respondents usually sleep on weekends or non-workdays, also ranging from 2 to 14 hours. It helps in understanding the variation in sleep patterns during days off from work.

Snoring Frequency (SLQ030): This variable records how often respondents snore while sleeping, with responses ranging from 0 (Never) to 3 (Frequently). Snoring is a common symptom of sleep disorders such as obstructive sleep apnea, making this variable relevant to the study of sleep health.

Overly Sleep Frequency (SLQ120): This variable assesses how often respondents feel excessively or overly sleepy during the day, with values ranging from 0 (Never) to 4 (Almost always). It is an indicator of sleep quality and quantity, as well as potential sleep disorders.

Reported Sleep Trouble (SLQ050): The dependent variable in this study, it indicates whether respondents have ever told a doctor or other health professional that they have trouble sleeping. It is treated as a binary outcome variable, with values of 0 (No report of sleep trouble) and 1 (Reporting sleep trouble).

2.4 Sample of Cleaned Sleep Disorder Data

Table 1: Sample of Sleep Disorder Data

Respondent ID	Weekday Sleep Duration (hrs)	Weekend Sleep Duration (hrs)	Snoring Frequency	Breathing Pause Frequency	Overly Sleep Frequency	Reported Sleep Trouble
109266	7.5	8.0	1	0	0	0
109267	8.0	8.0	0	0	2	0
109268	8.5	8.0	0	0	1	0
109271	10.0	13.0	0	0	3	1
109273	6.5	8.0	0	0	2	1
109274	9.5	9.5	1	0	0	0

Table 1 represents a subset of the broader NHANES sleep disorder dataset. Each row in the table corresponds to an individual participant, uniquely identified by their Respondent ID. The "Weekday Sleep Duration (hrs)" and "Weekend Sleep Duration (hrs)" columns quantify the number of hours slept during the weekdays and weekends, respectively, providing a snapshot of the individual's sleep patterns. "Snoring Frequency" and "Breathing Pause Frequency" are categorical measures that reflect how often the respondents experience snoring and breathing pauses during sleep, common indicators of sleep disturbances such as sleep apnea. The "Overly Sleep Frequency" column indicates the frequency at which respondents report feeling overly

sleepy during the day, a sign that can be indicative of inadequate sleep quality or quantity. Lastly, the "Reported Sleep Trouble" column is a binary measure showing whether the respondent has reported having sleep troubles to a health professional, with 0 signifying no reported trouble and 1 indicating reported trouble.

2.5 Measurement:

In this study, we utilized data from the National Health and Nutrition Examination Survey (NHANES), specifically focusing on the sleep disorders component which includes data collected between 2017 and March 2020. The NHANES program, a longstanding project conducted by the National Center for Health Statistics (NCHS), plays a critical role in assessing the health and nutritional status of adults and children in the United States. This dataset is pivotal in understanding public health and informs policy decisions through scientifically reliable data (CDC 2021).

The sleep disorders dataset within NHANES is enriched by questions adapted from the Munich ChronoType Questionnaire (Roenneberg, Wirz-Justice, and Merrow 2003), targeting various aspects of sleep behavior and disorders. The inclusion of these questions is instrumental in exploring the complex dynamics of sleep patterns among the U.S. population. Due to disruptions caused by the COVID-19 pandemic, the 2019-2020 data collection cycle was prematurely halted in March 2020, leading to its combination with the 2017-2018 cycle to ensure national representativeness and analytical robustness.

This combined dataset referred to as the NHANES 2017-March 2020 pre-pandemic data, offers valuable insights into the sleep habits of Americans before the pandemic. It is instrumental for researchers and public health officials aiming to understand baseline sleep behaviors and potential disturbances across a broad demographic spectrum.

To prepare the dataset for analysis, extensive data cleaning and processing were conducted. This included removing entries with missing, refused, or 'don't know' responses for critical variables such as sleep duration and frequency of snoring. Additionally, to address issues with data reliability and consistency, about 3% of the data underwent verification through audio recordings of the interviews. Moreover, for variables capturing sleep duration on weekdays (SLD012) and weekends (SLD013), reported times were meticulously reviewed, with outliers adjusted and rounded to the nearest half-hour, enhancing the data's accuracy and usability.

Detailed descriptions of the variables used in this study, along with the specific adjustments made to the dataset, are available in Section 2.3. This section is designed to provide a comprehensive understanding of the origins, processing, and analytical framework applied to each variable relevant to this study.

2.6 Data Exploration:

In this section, we explore the distributions of key variables related to sleep patterns and disorders in the NHANES dataset. Histograms provide visual insights into the frequency of reported sleep duration, snoring, breathing pauses, daytime sleepiness, and reported sleep troubles.

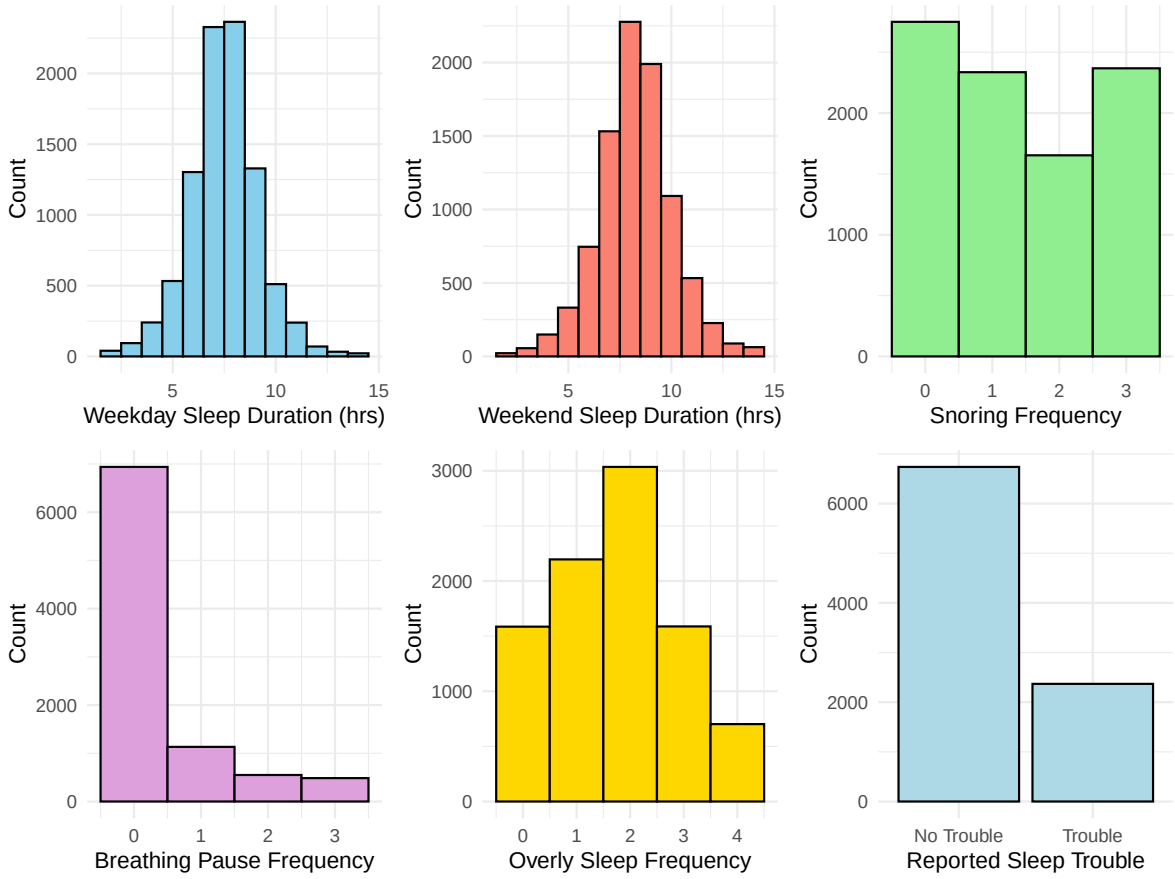

Figure 1: Distributions of sleep-related variables

Weekday Sleep Duration (hrs): The histogram for weekday sleep duration exhibits a unimodal distribution, peaking around 7 to 8 hours, which aligns with the generally recommended sleep duration for adults. The symmetrical shape suggests a commonality in sleep patterns among individuals during the workweek.

Weekend Sleep Duration (hrs): Similarly unimodal, the weekend sleep duration distribution also peaks around the same range as weekday sleep but shows a noticeable tendency for slightly longer durations. This may suggest that individuals take the opportunity to sleep more on weekends, possibly compensating for the workweek.

Snoring Frequency: The snoring frequency histogram reveals a broad distribution, with the majority of respondents indicating they never or rarely snore. A significant count of individuals also reports frequent snoring, suggesting that the experience of snoring is quite varied among the respondents.

Breathing Pause Frequency: The distribution for breathing pause frequency is steeply skewed towards 'Never', suggesting that breathing pauses during sleep are infrequently experienced or reported by the majority of respondents. The scant instances of frequent breathing pauses may indicate a lower occurrence or a lack of self-awareness of these events.

Overly Sleep Frequency: The distribution of daytime sleepiness frequency is right-skewed, with 'Sometimes (2-4 times a month)' emerging as the modal category. However, a notable proportion of respondents experience sleepiness during the day at least once a month, suggesting that daytime sleepiness is a common issue.

Reported Sleep Trouble: The binary distribution for reported sleep trouble shows a larger proportion of individuals reporting no sleep trouble compared to those reporting trouble, approximately threefold. This ratio highlights that while sleep trouble is present, it is not reported by the majority of participants in this sample.

3 Model

The aim of our model is to explore the relationship between various sleep-related factors and the self-reported incidence of sleep troubles. We employ a Bayesian logistic regression model to analyze the data from the National Health and Nutrition Examination Survey (NHANES). Further details and diagnostics of this model are provided in Section 6.2.1.

3.1 Model set-up

Let y_i denote the binary outcome indicating whether an individual has reported sleep troubles to a doctor, with $y_i = 1$ for reported troubles and $y_i = 0$ otherwise. The predictors include weekday sleep duration (x_{i1}), weekend sleep duration (x_{i2}), snoring frequency (x_{i3}), breathing pause frequency (x_{i4}), and daytime sleepiness (x_{i5}). The logistic regression model is formulated as follows:

$$
\begin{aligned}
y_i|p_i &\sim \text{Bernoulli}(p_i) && (1)\\
\text{logit}(p_i) &= \beta_0 + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + \beta_3 \times x_{i3} + \beta_4 \times x_{i4} + \beta_5 \times x_{i5} && (2)\\
\beta_0 &\sim \text{Normal}(0, 1) && (3)\\
\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 &\sim \text{Normal}(0, 1) && (4)\\
& && (5)
\end{aligned}
$$

The logistic regression model is defined using a Bayesian framework, implemented in R with the `rstanarm` package. This approach allows us to incorporate prior knowledge about the parameters and to estimate their posterior distributions based on the observed data.

In our model, the probability of an individual reporting sleep troubles, denoted by p_i, follows a Bernoulli distribution. The log odds of reporting sleep troubles are modeled as a linear combination of the predictors, with β_0 representing the intercept, and β_1 through β_5 representing the slopes for the respective predictors.

The priors for the intercept (β_0) and the slopes (β_1, β_2, β_3, β_4, β_5) are deliberately chosen to reflect initial neutrality or skepticism regarding the impact of each predictor on the likelihood of reporting sleep troubles. Specifically, β_0 is assigned a Normal(0, 1) prior, indicating our initial uncertainty about the base rate of sleep trouble in the population.

For β_1 and β_2, which correspond to the effects of weekday and weekend sleep duration, we posit Normal(0, 0.5) priors. These priors are centered around zero, reflecting the number of sleep hour can be either positively or negatively contribute to sleep troubles, with a modest expectation of effect size.

The coefficients β_3, β_4, β_5 relating to snoring frequency, breathing pause frequency, and overly sleep frequency are each given a Normal(0, 1) prior. These priors imply that, prior to analyzing the data, we do not have strong expectations about the direction or scale of these effects.

3.2 Model Justification

Our Bayesian logistic regression model is designed to investigate the associations between various sleep-related factors and the likelihood of reporting sleep troubles. The priors for the coefficients are set with specific hypotheses in mind, informed by existing literature and plausible biological mechanisms:

Weekday Sleep Duration (β_1): We initially hypothesize that there may be a complex relationship between sleep duration on weekdays and sleep troubles. Long sleep durations could indicate an attempt to compensate for poor quality sleep, potentially related to sleep disturbances. As such, the prior for this coefficient is centered around zero, allowing the data to indicate the direction of association.

Weekend Sleep Duration (β_2): For sleep duration on weekends, we consider the possibility of a recuperative effect, where catching up on sleep may reduce the likelihood of reporting troubles. This hypothesis is tentative, hence a prior centered around zero with a conservative standard deviation, permitting data-driven insights into this relationship.

Snoring Frequency (β_3): Frequent snoring is a recognized symptom of sleep disorders like obstructive sleep apnea, which can lead to disrupted sleep. We posit a positive association between snoring frequency and reported sleep troubles, reflected in the priors for this coefficient.

Breathing Pause Frequency (β_4): Breathing pauses during sleep are indicative of sleep apnea, which is directly related to sleep troubles. We hypothesize a positive association for this coefficient, which is incorporated into the prior setting.

Daytime Sleepiness (β_5): Excessive daytime sleepiness is often the result of inadequate or disrupted nighttime sleep. We hypothesize a positive relationship between daytime sleepiness and the reporting of sleep troubles, expecting that those experiencing higher levels of sleepiness will be more likely to report troubles.

These hypotheses are structured into our Bayesian model as informed priors, which, while reflecting our initial expectations, are sufficiently flexible to be shaped by the empirical data. This integration of prior beliefs and observed information will yield posterior distributions that convey a robust interpretation of how sleep behaviors relate to the reporting of sleep troubles.

4 Result

4.1 Model Coefficients Interpretation

Table 2: Summary Statistics for the Coefficients of the Logistic Model

Term	Estimate
(Intercept)	-1.396
WeekdaySleepDuration	0.051
WeekendSleepDuration	-0.131
SnoringFrequency	0.032
BreathingPauseFrequency	0.412
OverlySleepFrequency	0.425

Table 2 presents the key coefficients for understanding the influence of various sleep-related variables on the likelihood of reporting sleep troubles:

Intercept (β_0): The intercept of the model, β_0, is estimated at -1.396. This value represents the log odds of reporting sleep troubles when all the predictors are held at zero. It suggests that, in the absence of any sleep issues or without considering any sleep behaviors, the baseline log odds of reporting sleep troubles to a doctor are negative, indicating a lower probability of reporting sleep troubles.

Weekday Sleep Duration (β_1): For the weekday sleep duration, the coefficient β_1 is 0.051, suggesting a slight increase in the log odds of reporting sleep troubles for each additional hour of sleep during the weekdays. This could be indicative of a situation where those who sleep

more on weekdays might be doing so due to sleep issues that have led them to consult with a doctor.

Weekend Sleep Duration (β_2): Conversely, the coefficient for weekend sleep duration, β_2, is -0.131. This implies that for each additional hour of sleep during the weekends, there is a decrease in the log odds of reporting sleep troubles. It could be interpreted as those who manage to sleep more on weekends are less likely to report sleep problems, possibly because catching up on sleep may alleviate some of their weekday sleep deficits.

Snoring Frequency (β_3): The coefficient for snoring frequency, β_3, has a value of 0.032. While this is a positive value, suggesting that an increase in snoring frequency is associated with an increase in the likelihood of reporting sleep troubles, the effect size is relatively small.

Breathing Pause Frequency (β_4): The coefficient for breathing pause frequency, β_4, is 0.412, indicating a more substantial positive association with reporting sleep troubles. This aligns with clinical understanding, as breathing pauses are often associated with sleep disorders like sleep apnea, which can be a significant concern prompting medical consultation.

Overly Sleep Frequency (β_5): Lastly, the coefficient for overly sleep frequency, β_5, is 0.425, suggesting a strong relationship with the reporting of sleep troubles. This result is intuitive, as feeling excessively sleepy can be a direct symptom of poor sleep quality or a sleep disorder, leading to discussions with a healthcare provider.

Interpretation: The model indicates that various factors are related to the reporting of sleep troubles. While longer weekday sleep duration slightly increases the log odds of reporting sleep troubles, longer weekend sleep can reduce it. Frequent snoring has a smaller effect compared to breathing pauses and daytime sleepiness, which both have significant positive associations with reporting sleep troubles. These findings underline the complexity of sleep behavior and its impact on sleep quality and health consultations.

4.2 Model Equation

$$\text{logit}(p_i) = -1.396 + 0.051 \times x_{i1} - 0.131 \times x_{i2} + 0.032 \times x_{i3} + 0.412 \times x_{i4} + 0.425 \times x_{i5} + \varepsilon_i \tag{6}$$

5 Discussion

6 Appendix

6.1 Additional Data Details

6.2 Model Details

6.2.1 Posterior Predictive Check

References

CDC. 2021. "P_SLQ." *National Health and Nutrition Examination Survey*. Centers for Disease Control; Prevention. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_SLQ.htm.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Roenneberg, Till, Anna Wirz-Justice, and Martha Merrow. 2003. "Life Between Clocks: Daily Temporal Patterns of Human Chronotypes." *Journal of Biological Rhythms* 18: 80–90.

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

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, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.