Decoding Sleep Disorders Through Self-Reported Patterns*

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

Terry Tu

April 13, 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:

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:

$$y_i|p_i \sim \text{Bernoulli}(p_i) \tag{1}$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + \beta_3 \times x_{i3} + \beta_4 \times x_{i4} + \beta_5 \times x_{i5} \tag{2}$$

$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$

$$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 \sim \text{Normal}(0, 2.5) \tag{4}$$

The model is implemented in R using the rstanarm package, enabling the incorporation of prior beliefs about the parameters and estimation of their posterior distributions. The stan_glm function from rstanarm is used to perform the Bayesian logistic regression, integrating both the data and prior knowledge for inference.

Priors for the intercept and coefficients are chosen to reflect a neutral stance before observing the data, centered around zero with a moderate spread to allow flexibility in the estimates. These priors convey an expectation of no substantial effect from each predictor, with the data informing deviations from this baseline assumption.

The logistic regression framework models the log odds of reporting sleep troubles as a linear combination of the predictors, with the coefficients indicating the direction and magnitude of each predictor's effect. The stan_glm function combines these priors with the data to estimate the posterior distributions of the model parameters, providing a comprehensive view of the relationships between sleep-related factors and reported sleep troubles.

3.2 Model Justification

In our analysis, we hypothesize that certain sleep-related factors will have significant associations with the likelihood of reporting sleep troubles. Specifically:

Weekday Sleep Duration (β_1): We hypothesize that this coefficient may be positive, indicating that longer sleep durations on weekdays could be associated with a higher likelihood of reporting sleep troubles. This could reflect the possibility that individuals with sleep disturbances may attempt to compensate by sleeping longer.

Weekend Sleep Duration (β_2): Conversely, we hypothesize that this coefficient may be negative, suggesting that longer sleep durations on weekends might be associated with a lower likelihood of reporting sleep troubles. This could indicate that catching up on sleep during weekends may alleviate some sleep-related issues.

Snoring Frequency (β_3): We expect this coefficient to be positive, as frequent snoring is often a symptom of sleep disorders such as obstructive sleep apnea, which could lead to reporting sleep troubles.

Breathing Pause Frequency (β_4): Similarly, we hypothesize that this coefficient will be positive, as pauses in breathing during sleep are a hallmark of sleep apnea, a condition commonly associated with sleep disturbances.

Daytime Sleepiness (β_5): We anticipate that this coefficient will also be positive, as excessive sleepiness during the day is a common consequence of poor sleep quality or insufficient sleep, which may prompt individuals to report sleep troubles.

The Bayesian framework allows us to incorporate these hypotheses as priors in our model, providing a structured way to integrate our substantive expectations with the data analysis. By doing so, we can obtain a more informed and nuanced interpretation of the relationships between sleep-related factors and the likelihood of reporting sleep troubles.

4 Result

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.