

Decoding Sleep Disorders Through Self-Reported Patterns*

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

Terry Tu

April 2, 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

1	Introduction	2
2	Data	2
3	Model	4
4	Result	4
5	Discussion	4
6	Appendix	4
	References	4

*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 (Appendix).

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. It is derived from their reported usual bedtime and wake time, providing an insight into their sleep patterns during the typical work-week.

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. 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 ‘Never’ to ‘Frequently’. Snoring is a common symptom of sleep disorders such as obstructive sleep apnea, making this variable relevant to the study of sleep health.

Daytime Sleepiness (SLQ120): This variable assesses how often respondents feel excessively or overly sleepy during the day. 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, where ‘1’ signifies reporting sleep trouble and ‘2’ signifies no report of sleep trouble.

3 Model

4 Result

5 Discussion

6 Appendix

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/>.
- 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 Golemund, 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>.