STATISTICAL ANALYSIS OF DETERMINANTS OF SLEEP DURATION

Miriam Helena Hudák – Jakub Danko

Abstract

This article presents a statistical analysis to identify and understand the factors that influence sleep duration. In the introduction, basic information about sleep is provided, which will help to better understand its functions and role in our life. Subsequently, the statistical model is theoretically described, specifically the logistic model, which is used in the analytical part of this study. This model allows us to evaluate the influence of selected factors on the length of sleep, while the analyzed data comes from the National Research Resource. The goal of this analysis is to gain a deeper understanding of how individual factors, namely gender, age, race, and body mass index, affect individuals' sleep duration. The results of this analysis provide important insights into how these factors interact and influence sleep. The collected data and statistical methods pave the way for further research in the field of sleep and health, with the aim of improving the quality of life for all individuals. This statistical analysis offers insight into the complex nature of sleep and its determinants.

Key words: sleep duration, logistic regression, REM sleep phase

JEL Code: C10, I10

Introduction

Sleep is one of the biological needs of the human body. It is a fundamental aspect of our lives, occupying a substantial portion of our daily routines. Beyond its restorative properties, sleep plays a crucial role in maintaining overall health and well-being. The quantity and quality of sleep a person receives can significantly impact their physical, mental, and emotional state. Many scientists are trying to understand the secret of sleep, but despite a lot of research, it still has many unexplained and unknown sides.

In the presented article, we decided to look at selected factors that mainly relate to the length and quality of sleep. Using statistical methods and models, we will perform an analysis and try to clarify the results based on theory.

1 Data description and methodology

Presented analysis is based on the data from the National Sleep Research Resource (NSRR). Currently, they have at their disposal almost 30 different data sets, which contain data on respondents of different age categories with a number of variables that, in addition to sleep, also relate to other sociodemographic and health data on individual respondents. In the article presented by us, we analyze the dataset marked as SHHS. It's an acronym from the Sleep Heart Health Study, which is a multi-center cohort study implemented by the National Heart Lung & Blood Institute to determine the cardiovascular and other consequences of sleep-disordered breathing.

In order to quantify individual variables and their impact on the respondent's sleep duration, we used a binary logistic regression model. Logistic regression predicts the probability that an observation falls into one of two categories of a dichotomous dependent variable based on one or more independent variables that can be either continuous or categorical.

We model the conditional probability of one categorical value of the variable Y from the explanatory variables X1, X2, . . . , Xk. The simplest case is when the explained variable is alternative (dichotomous, binary), but we can also consider an ordinal or categorical variable with more than two categories. Based on this, we also know more complex models such as multinomial logistic regression or ordinal logistic regression, but we will not deal with these in our work.

If we assume that the probability that a randomly selected respondent has sufficient sleep duration (or normal sleep efficiency) is π, and the probability that a randomly selected respondent does not is 1-π, then the chance of the occurrence of agreement with this statement (denote as odds) can be calculated as follows:

$$odds = \frac{\pi}{1-\pi} \tag{1}$$

And the following formula holds for the probability π:

$$\pi = \frac{odds}{1+odds} \tag{2}$$

Logit is defined as $\ln(odds) = \ln\frac{\pi}{1-\pi}$ with values from -∞ to ∞. If we use logit as the explained variable, the logistic regression function will be of the form:

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{3}$$

Odds and the probability is then obtained by reverse transformation:

$$\frac{\pi}{1-\pi} = \mathrm{e}^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k} \tag{4}$$

We interpret the coefficients β_j estimated by logistic regression as a change in logit caused by a unit change in the value of the explanatory variable X_j under the assumption of ceteris paribus (otherwise unchanged conditions). However, the value of e^{β_j} is better interpreted: it is the multiple by which the odds changes if the value of the explanatory variable increases by one unit, assuming ceteris paribus. When this value is higher than 1, the odds will increase, and if this value is lower than 1, the odds will decrease.

When interpreting, one must pay attention to categorical variables, where this interpretation is not for a unit change but in relation to the so-called reference category.

As binary dependent variable, we defined variable named *durationBinary* based on the recommendations of the National Sleep Foundation. The panel agreed that, for healthy individuals with normal sleep, the appropriate sleep duration for newborns is between 14 and 17 hours, infants between 12 and 15 hours, toddlers between 11 and 14 hours, preschoolers between 10 and 13 hours, and school-aged children between 9 and 11 hours. For teenagers, 8 to 10 hours was considered appropriate, 7 to 9 hours for young adults and adults, and 7 to 8 hours of sleep for older adults. From the point of view of age, the sample of respondents analyzed by us can be considered as adults, respectively older adults (see Table 1 below where you can see that the range of age of respondents is between 39 to 90). Based on this, according to the variable *Total Sleep Duration* which is defined as the interval between sleep onset and sleep offset while the participant is asleep from type II polysomnography in minutes, we calculated the binary variable *durationBinary*, which is equal to the value 1 if this duration is at least 7 hours and to the value 0 if this duration is less than 7 hours.

After cleaning the data for missing observations, we had available data on 5044 respondents, while initially more than 1000 different attributes about these respondents were available. The basic descriptive characteristics of the selected continuous variables, which will also appear in the model, are presented in the Table 1.

Tab. 1: Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Total Sleep Duration	5044	34,50	519,00	361,1064	63,65227	4051,611
Sleep Efficiency	5044	11,31	104,15	82,8462	10,42777	108,738
Wake after Sleep Onset	5044	,00	338,50	61,5498	43,94822	1931,446
Age of participant	5044	39,00	90,00	63,6009	10,97465	120,443
Body mass index	5044	18,00	50,00	28,1647	5,06369	25,641
Percentage of total sleep duration in REM	5044	,00	43,30	19,8152	6,21237	38,593
Valid N (listwise)	5044					

Source: own processing using IBM SPSS

From the point of view of the representation of some binary or categorical characters, 47.6% of the respondents are men, 52.4% of the respondents are women. The vast majority of respondents are white (86.2%), 7.6% are black, and 6.2% are of a different skin color. About a third of respondents (30.2%) had trouble falling asleep last night.

2 Analysis of sleep duration

After analyzing a number of variables, we selected those that statistically significantly affect the respondent's odds of sleeping 7 or more hours. The resulting logit model has the form:

$$\begin{aligned} logit(\text{durationBinary}) &= -41.206 - 0.672\ Gender\ (Male) - 0.089\ Age\ category \\ &- 0.457\ Race\ (White) - 0.621\ Race\ (Black) + 0.431\ Sleep\ Efficiency \\ &+ 0.064\ Wake\ after\ Sleep\ Onset + 0.312\ Difficulty\ falling\ asleep\ (No) \\ &+ 0.029\ REM\ percentage - 0.019\ BMI \end{aligned}$$

Tab. 2: Comparison of Logit Models

Variable in Logit Model	B	Sig.	Exp(B)	Exp(B) - 1	Exp(B) - 1 [*100 %]
Gender of the participant (1)	-0,672	0,000	0,511	-0,489	-48,91 %
Age category	-0,089	0,029	0,915	-0,085	-8,54 %
Race of the participant		0,010			
Race of the participant (1)	-0,457	0,005	0,633	-0,367	-36,69 %
Race of the participant (2)	-0,621	0,007	0,537	-0,463	-46,25 %
Sleep Efficiency	0,431	0,000	1,539	0,539	53,91 %

Wake after Sleep Onset	0,064	0,000	1,066	0,066	6,60 %
Did you have difficulty falling asleep last night? (1)	0,312	0,004	1,366	0,366	36,56 %
Percentage of total sleep duration in REM	0,029	0,000	1,030	0,030	2,96 %
Body mass index	-0,019	0,026	0,981	-0,019	-1,92 %
Constant	-41,206	0,000	0,000		

Source: own processing using IBM SPSS

We performed model diagnostics using the Hosmer-Lemeshow. The Hosmer-Lemeshow test is a statistical test for goodness of fit for the logistic regression model. A large value of Chi-squared (with small p-value < 0.05) indicates poor fit and small Chi-squared values (with larger p-value closer to 1) indicate a good logistic regression model fit. In our case we have large p-value (0.45) so it means that our logistic model is good enough.

Tab. 3: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7,829	8	,450

Source: own processing using IBM SPSS

The model can explain about a third of the total variability, we quantify it using pseudo R square characteristics (Cox & Snell and Negelkerke R Square) shown in Table 4.

Tab. 4: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3430,354	,222	,367

Source: own processing using IBM SPSS

We used the Wald test to determine the statistical significance of the regression parameters of the model. The Wald test is a parametric statistical measure to confirm whether a set of independent variables are collectively 'significant' for a model or not. It is also used for confirming whether each independent variable present in a model is significant or not.

Let's move on to the interpretation of the regression parameters. First of all, we must realize that categorical variables are interpreted in relation to the so-called reference category. From the table below, we can see that in the case of the participant's race, we considered a race *other* than white and black as the reference category, in the case of difficulty falling asleep, the reference category was the answer *yes*, and from the point of view of gender, the reference category is *female*.

Tab. 5: Categorical Variables Codings

			Parameter coding	
		Frequency	(1)	(2)
Race of the participant	White	4346	1,000	,000
	Black	383	,000	1,000
	Other	315	,000	,000
Did you have difficulty falling asleep last night?	No	3521	1,000	
	Yes	1523	,000	
Gender of the participant	Male	2400	1,000	
	Female	2644	,000	

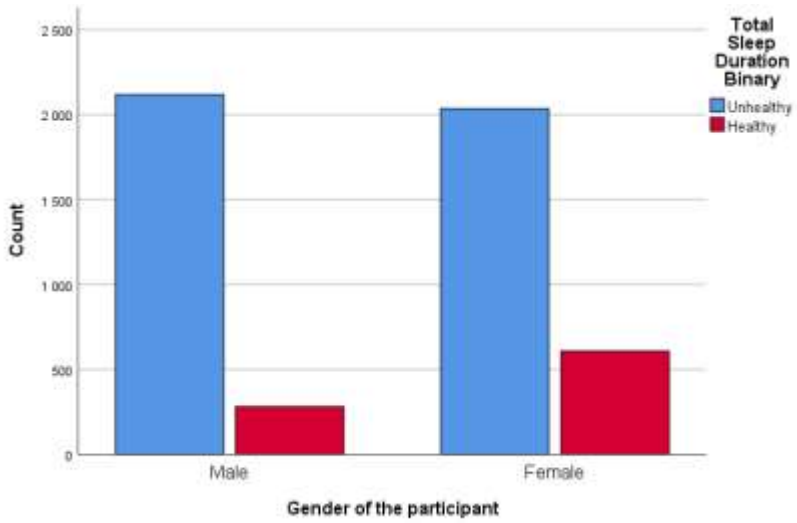
Source: own processing using IBM SPSS

2.1 The relationship between sleep duration and gender

Men are 48.91% less likely to sleep at least 7 hours compared to women, ceteris paribus. In the context of our analysis, we consider sleeping at least 7 hours as a "healthy sleep duration". It can also be seen on the basis of the graphs below, in which we can see that unfortunately the "unhealthy" sleep duration dominates in both genders, but that women sleep longer than men.

There are several reasons why women need more sleep than men. Hormones play a significant role in regulating sleep needs and the sleep-wake cycle. For women, these hormonal changes can occur monthly (within the menstrual cycle) or throughout life (for example, during pregnancy and menopause). These changes can affect their circadian rhythms, which can lead to a greater need for sleep. In addition, the female brain has a more complex structure and requires a longer time for its regeneration.

Fig. 1: Comparison of Total Sleep Duration between gender



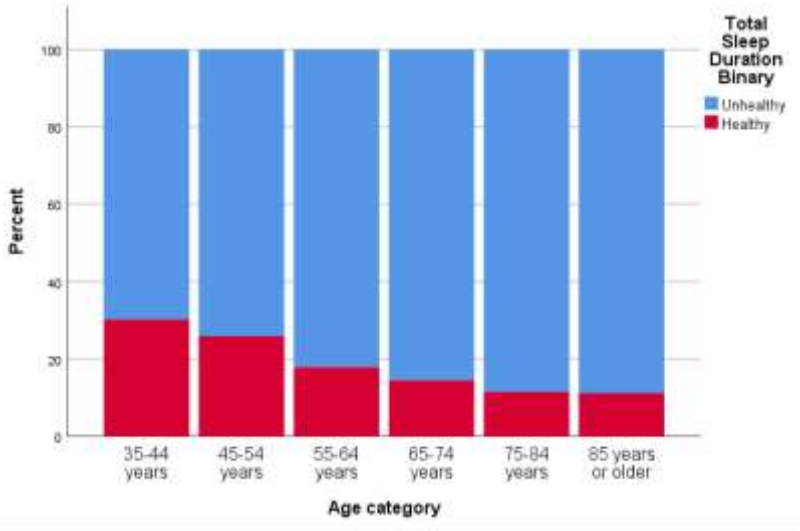
Source: own processing using IBM SPSS

2.2 The relationship between sleep duration and age category

With increasing age, the odds that the respondent's sleep duration will be longer than 7 hours decreases by an average of 8.54% ceteris paribus. Age brings physiological changes in the body, which also include changes in sleep patterns.

Elderly individuals often tend to sleep less and wake up more often at night. At the same time, there may be a risk of the occurrence of various health problems that can affect the quality of sleep. These include, for example, heart problems, arthritis, insomnia, and breathing problems (such as sleep apnea).

Fig. 2: Stacked Bar Percent of Age category by Total Sleep Duration



Source: own processing using IBM SPSS

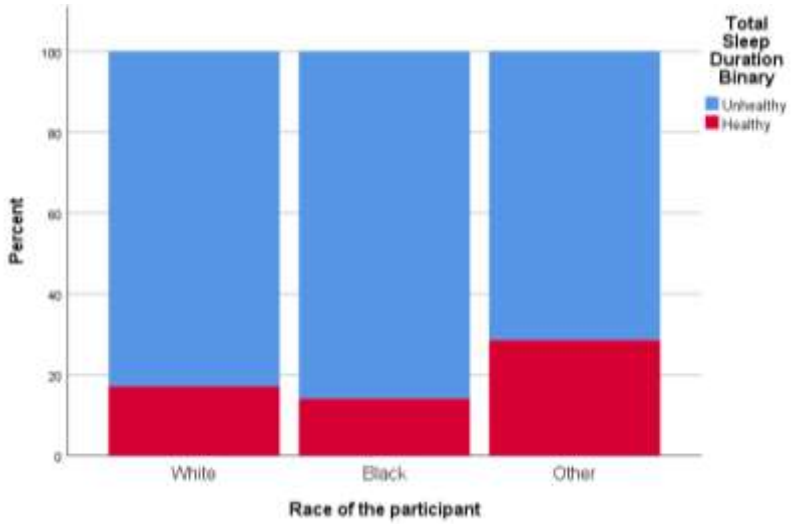
2.3 The relationship between sleep duration and race

Different cultures and environments may have different approaches to sleep and sleep habits. For example, family habits, working hours, or social norms affect the length of sleep. Socioeconomic status or genetic factors are also closely related to this.

The results of our analysis show that the reference category of a race other than whites and blacks fares better than these two basic races in terms of sleep duration. We see that, compared to the reference category, the odds of healthy sleep duration decrease by an average of 36.69% for whites, and even by 46.25% for blacks, ceteris paribus.

It is important to emphasize that these differences in sleep duration between different racial groups are the result of a complex set of interactions between these factors and cannot be easily reduced to a single simple reason.

Fig. 3: Stacked Bar Percent of Race of the participant by Total Sleep Duration



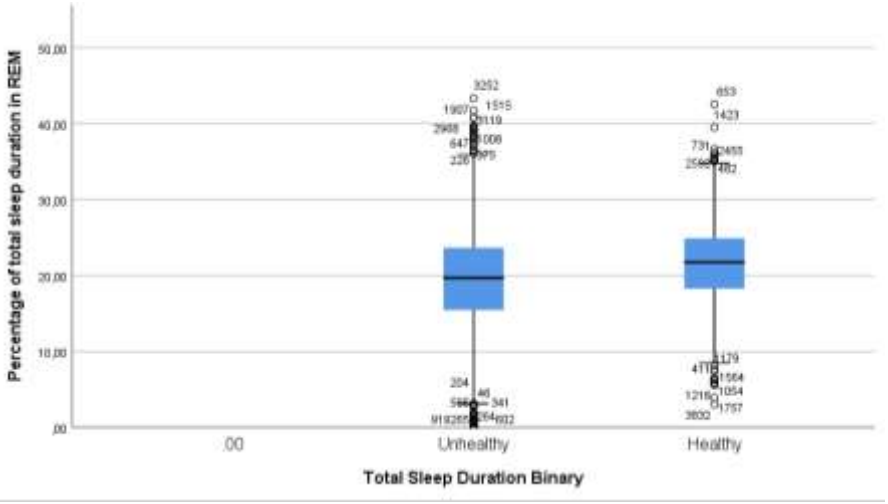
Source: own processing using IBM SPSS

2.4 The relationship between sleep duration and percentage of total sleep duration in REM

Rapid eye movement (REM) sleep is one of four stages the brain moves through while sleeping. In REM sleep, the eyes move rapidly in various directions and dreams can occur. REM sleep typically starts within 90 minutes of falling asleep.

Our analysis shows that with an increase in the proportion of REM sleep by one percentage point, the odds of a healthy sleep duration increase by almost three percent, ceteris paribus. The REM phase of sleep is important for restoring brain functions and emotional balance. During this phase, cognitive processing and memory consolidation take place. Thus, improved quality and duration of REM sleep can lead to better psychological and cognitive well-being. When this phase is well balanced and lasts long enough, it can contribute to better rest during the night.

Fig. 4: Boxplot of Percentage of Total Sleep Duration in REM



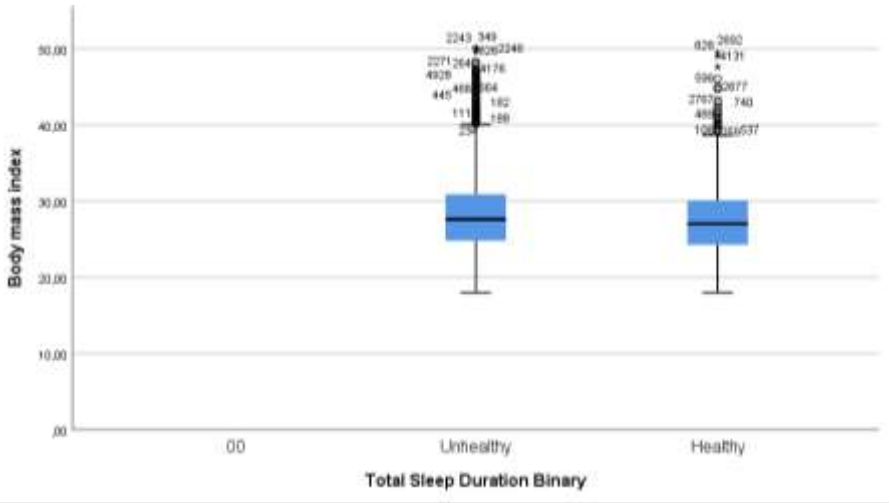
Source: own processing using IBM SPSS

2.5 The relationship between sleep duration and Body Mass Index

The body mass index (BMI) is a measure that uses your height and weight to work out if your weight is healthy. The BMI calculation divides an adult's weight in kilograms by their height in meters squared. In the context of sleep, BMI has an impact on several factors that can affect the likelihood of long healthy sleep.

Our analysis shows that with an increase in the proportion of BMI by one unit, the odds of a healthy sleep duration decrease by less than two percent, ceteris paribus. Those with a higher BMI can often face physical discomfort, such as louder snoring or discomfort in bed. At the same time, they have a higher risk of developing sleep apnea. This disease can lead to frequent awakenings at night and reduce the quality of sleep, as we mentioned above.

Fig. 8: Boxplot of Body Mass Index by Total Sleep Duration



Source: own processing using IBM SPSS

Conclusion

Sleep is one of the most important needs in our life. Even though we are not aware of it, it brings us many beneficial effects on our health. In this article, we focused on the analysis of selected factors that affect the length of sleep. We can observe differences not only between men and women, but also between people of different ages, races, or people with different BMI. However, it is important to remember that every human body is unique, and we cannot generalize these results. Sleep affects several other aspects that were not explored in this article.

References

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., & Redline, S. (2018). The National Sleep Research Resource: Towards a Sleep Data Commons. Journal of the American Medical Informatics Association, 25(10), 1351–1358. https://doi.org/10.1093/jamia/ocy064

Quan SF;Howard BV;Iber C;Kiley JP;Nieto FJ;O'Connor GT;Rapoport DM;Redline S;Robbins J;Samet JM;Wahl PW; (n.d.). The sleep heart health study: Design, rationale, and methods. Sleep. https://pubmed.ncbi.nlm.nih.gov/9493915/

Hirshkowitz, M., Whiton, K., Albert, S. M., Alessi, C. A., Bruni, O., DonCarlos, L. L., Hazen, N., Herman, J. B., Katz, E. S., Kheirandish-Gozal, L., Neubauer, D., O'Donnell, A., Ohayon, M. M., Peever, J. H., Rawding, R., Sachdeva, R., Setters, B., Vitiello, M. V., Ware, J. S., & Hillard, P. J. A. (2015). National Sleep Foundation's sleep time duration recommendations: methodology and results summary. Sleep Health, 1(1), 40–43. https://doi.org/10.1016/j.sleh.2014.12.010

U.S. Department of Health and Human Services. (n.d.). Brain basics: Understanding sleep. National Institute of Neurological Disorders and Stroke. https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-understanding-sleep

Pacheco, D. (2023, August 1). Do women need more sleep than men?. Sleep Foundation. https://www.sleepfoundation.org/women-sleep/do-women-need-more-sleep-than-men

Kráľ, P., Kanderova, M., Kaščáková, A., & Bojdova, V. (2009). Viacrozmerné štatistické metódy so zameraním na riešenie problémov ekonomickej praxe. ResearchGate. https://www.researchgate.net/publication/265785993_Viacrozmerne_statisticke_metody_so_zameranim_na_riesenie_problemov_ekonomickej_praxe

Contact

Miriam Helena Hudák
Prague University of Economics and Business
W. Churchill Sq. 1938/4, 130 67 Prague, Czech Republic
hudm07@vse.cz

Jakub Danko
Prague University of Economics and Business
W. Churchill Sq. 1938/4, 130 67 Prague, Czech Republic
jakub.danko@vse.cz