UKRAINIAN CATHOLIC UNIVERSITY

Applied Sciences Faculty

Signal analysis course project:

FACTORS AFFECTING SLEEP QUALITY

Khrystyna Kubatska

Lviv - 2020

Abstract

The paper explores the impact of different factors on sleep quality, measured as the Pittsburg Sleep Quality Index. Data was taken from the Multilevel Monitoring of Activity and Sleep in Healthy people dataset, which provides various dimension information for 22 adult people. Different regression models with OLS estimators were made, and 4 of them were chosen to show a possible connection between various factors and PSQI. The conclusion was made using the results of the analysis, and suggestions to improve are also provided.

1 Introduction

Sleep is an essential function that allows a human body and mind to recharge, leaving it refreshed when it wakes up. Healthy sleep also helps the body be healthy and stave off diseases. Without it, the brain cannot function properly. This can impact concentration, clear thinking, and memory processing. For different age categories, there are different sleep durations needed for proper cognitive and behavioral functions. However, this is not a single reason that can have a significant impact. Besides duration, different factors can impact sleep quality, and it is essential to understand the relationship between sleep and daily life since it can provide insights into a healthy lifestyle.

In this research, I wanted to establish a specific explanation for the impact of particular heart data, triaxal accelerometer data, sleep quality, physical activity, psychological characteristics, and salivary sample factors on sleep quality, using statistical methods. The project aim is to investigate the relationship between different factors and sleep quality, measured as PSQI, for further improvement/development of sleep techniques. I hope that the paper's models will give a clear picture of the impact of individual factors on sleep quality.

2 Database description*

2.1 Database

Multilevel Monitoring of Activity and Sleep in Healthy people (MMASH) is the project's main dataset. It provides 24 hours of continuous beat-to-beat heart data, triaxal accelerometer data, sleep quality, physical activity, psychological characteristics and salivary samples.

2.2 Methods

The data were collected and provided by BioBeats in collaboration with researchers from the University of Pisa. The data were recorded by sport and health scientists, psychologists and chemists.

22 healthy young adult males were recruited. The study was approved by the Ethical Committee of the University of Pisa (0077455/2018).

At the start of the data recording, anthropomorphic characteristics (i.e. age, height and weight) of the participants were recorded. At the same time, participants filled in a set of initial questionnaires that provide information about participants psychological status:

- Morningness-Eveningness Questionnaire (MEQ)
- Pittsburgh Sleep Quality Questionnaire Index (PSQI)
- Behavioural avoidance/inhibition (BIS/BAS).

During the test, participants wore two devices continuously for 24 hours:

- a heart rate monitor (Polar H7 heart rate monitor - Polar Electro Inc., Bethpage, NY, USA) to record heartbeats and beat-to-beat interval
- an actigraph (ActiGraph wGT3X-BT - ActiGraph LLC, Pensacola, FL, USA) to record actigraphy information such as accelerometer data, sleep quality and physical activity

Also, the perceived mood (Positive and Negative Affect Schedule - PANAS) were recorded at different times of the day (i.e. 10, 14, 18, 22 and 9 of the next day).

Additionally, participants filled in Daily Stress Inventory (DSI) before going to sleep, to summarize the stressful events of the day.

Twice a day (before going to bed and when they woke up) the subjects collected saliva samples at home in appropriate vials. Saliva samples were used to extract RNA and measure the induction of specific clock genes, and to assess specific hormones.

A washout period from drugs of at least a week was required from the participants in the study.

2.3 Data description

MMASH consists of seven files for each participant (22):

- user_info.csv - anthropocentric characteristics of the participant:
  - *gender*: M and F refer to Male and Female, respectively;
  - *height* is expressed in centimetres (cm);
  - *weight* is expressed in kilograms (kg);
  - *age* is expressed in years.
- sleep.csv - information about sleep duration and sleep quality of the participant:
  - *In Bed Date*: 1 and 2 refer to the first and second day of data recording, respectively;
  - *In Bed Time*: time of the day (hours:minutes) when the user went to the bed
  - *Out Bed Date*: 1 and 2 refer to the first and second day of data recording, respectively;
  - *Out Bed Time*: time of the day (hours:minutes) when the user went out of the bed;
  - *Onset Date*: 1 and 2 refer to the first and second day of data recording, respectively;
  - *Onset Time*: time of the day (hours:minutes) when the user falls asleep;
  - *Latency Efficiency*: percentage of sleep time on total sleep in bed;
  - *Total Minutes in Bed*: minutes spent in the bed per night;
  - *Total Sleep Time (TST)*: length of the sleep per night expressed in minutes;
  - *Wake After Sleep Onset (WASO)*: time spent awake after falling asleep the first time;
  - *Number of Awakenings* during the night;
  - *Average Awakening Length*: time in seconds spent awakening during the night;
  - *Movement Index*: number of minutes without movement expressed as a percentage of the movement phase (i.e., number of period with arm movement)
  - *Fragmentation Index*: number of minutes with movement expressed as a percentage of the immobile phase (i.e., the number of period without arm movement);
  - *Sleep Fragmentation Index*: ratio between the Movement and Fragmentation indices.
- RR.csv - beat-to-beat interval data:
  - *ibi_s*: time in seconds between two consecutive beats;
  - *day*: 1 and 2 refer to the first and second day of data recording, respectively;
  - *time*: day time when the heartbeat happened (hours:minutes:seconds).
- questionnaire.csv - scores for all the questionnaires:
  - *MEQ*: Morningness-Eveningness Questionnaire value. The chronotype score is ranging from 16 to 86: scores of 41 and below indicate Evening types, scores of 59 and above indicate Morning types, scores between 42-58 indicate intermediate types;
  - *STAI1*: State Anxiety value obtained from State-Trait Anxiety Inventory. The results are range from 20 to 80. Scores less than 31 may indicate low or no anxiety, scores between 31 and 49 an average level of anxiety or borderline levels, and scores higher than 50 a high level of anxiety or positive test results;

- *STAI2*: Trait Anxiety value obtained from the State-Trait Anxiety Inventory. The results are range from 20 to 80. Scores less than 31 may indicate low or no anxiety, scores between 31 and 49 an average level of anxiety or borderline levels, and scores higher than 50 a high level of anxiety or positive test results;
- *PSQI*: Pittsburgh Sleep Quality Questionnaire Index. It gives a score rating from 0 to 21, with values lower than 6 indicating good sleep quality;
- BIS/BAS: Behavioural avoidance/inhibition index. BIS/BAS scales are a typical measure of reinforcement sensitivity theory that establish biological roots in personality characteristics, derived from neuropsychological differences;
- *Daily_stress*: Daily Stress Inventory value (DSI) is a 58 items self-reported measures which allows a person to indicate the events they experienced in the last 24 hours. It gives a score between 0 and 406. The higher is this values, the higher is the frequency and degree of the events and the perceived daily stress;
- *PANAS*: Positive and Negative Affect Schedule. It gives a score rating between 5 and 50 for both positive and negative emotions. The higher is the PANAS value, the higher is the perceived emotion.

- Activity.csv - list of the activity categories throughout the day. The categories are (the activities listed below correspond to the numeric ID of each activity in the csv file):

  1. sleeping;
  2. laying down;
  3. sitting, e.g. studying, eating and driving;
  4. light movement, e.g. slow/medium walk, chores and work.;
  5. medium, e.g. fast walk and bike;
  6. heavy, e.g. gym, running;
  7. eating;
  8. small screen usage, e.g. smartphone and computer;
  9. large screen usage, e.g. TV and cinema;
  10. caffeinated drink consumption, e.g. coffee or coke;
  11. smoking;
  12. alcohol assumption.

  'Start' and 'end' columns refer to the time of the day (hours:minutes) when the event happened, while 'day' columns refers to the day when it happened (1 and 2 refer to the first and second day of data recording, respectively).

- Actigraph.csv - accelerometer and inclinometer data recorded throughout the day:
  - *Axis1*: Raw Acceleration data of the X-axis expressed in Newton-meter;
  - *Axis2*: Raw Acceleration data of the Y-axis expressed in Newton-meter;
  - *Axis3*: Raw Acceleration data of the Z-axis expressed in Newton-meter;
  - *Steps*: number of steps per second;
  - *HR*: beats per minutes (bpm);
  - *Inclinometer Off*: values equal to 1 refer to no activation of the inclinometer. The values are reported per second;
  - *Inclinometer Standing*: values equal to 1 refer to the standing position of the user, while 0 refers to other user positions. Values are reported per second;
  - *Inclinometer Sitting*: values equal to 1 refer to the sitting position of the user, while 0 refers to other user positions. Values are reported per second;
  - *Inclinometer Lying*: values equal to 1 refer to the lying position of the user, while 0 refers to other user positions. Values are reported per second;
  - Vector Magnitude: vector movement derived from raw acceleration data expressed in Newton-meter;
  - *day*: 1 and 2 refer to the first and second day of data recording, respectively;

  - *time*: day time when the heartbeat happened (hours:minutes:seconds).

- saliva.csv - clock genes and hormones concentrations in the *saliva* before going to bed and after waking up. Two samples per participant are included, one before sleep and one after waking up, as indicated by the "Sample" data column. Melatonin levels are reported in g of *melatonin* per g of protein, while *cortisol* levels are in g of cortisol per 100 g of protein. No clock genes and hormones concentrations data was provided for User_21 due to problem in the salivary samples that do not permit to analyse it.

** All information provided in the chapter was taken from https://physionet.org/content/mmash/1.0.0/*

3 Tasks that can be solved using MMASH data

There are no mentions of researches that have used the MMASH database as a basis yet. However, the authors of the database indicated some possible options for what this database could be used.

For example:

- Development machine learning algorithms (detecting daily activities, moods, emotions, individual predisposition to react toward aversive or positive events, and stress following cardiovascular responses and/or actigraphy data) to predict people's routine by using accelerometers data and cardiovascular responses.
- Investigation the relationship between several aspects of psychophysiological responses having a complete overview of the users' daily lives.
- Development a variety of methodologies to improve sleep quality by identifying patterns.

4 Project aim and objectives

Sleep quality is essential as it directly affects memory, emotional recovery, and performance.

Thus, the project aim is to investigate the relationship between different factors and sleep quality, measured as PSQI, for further improvement/development of sleep techniques.

Tasks:

1. Review the literature about methods for identifying patterns related to sleep quality.
2. Select approach/methodology, describe it.
3. Describe the formulation of the experiment.
4. From the database, identify and select specific data that will be used.
5. Analyze and describe the data: hypothesis, summary statistics, Gauss-Markov assumptions.
6. Make results: regression models, interpretation.
7. Suggest possible areas for further researches in this or related fields.
8. Make conclusions.

5 Literature review

Today, there is a large number of papers trying to trace the factors that influence sleep quality. Nevertheless, a big part of them is concentrated in identifying patterns between the quality of sleep of sick people, or people of certain specific groups and various factors, but not a random sample of healthy people.

Yin Bai, Bin Xu, Yuanchao Ma, Guodong Sun, and Yu Zhao, in their study "Will You Have a Good Sleep Tonight? Sleep quality Prediction with Mobile Phone" (2012), proposed a novel approach for predicting an individual's sleep quality using mobile data. They used the correlation method, statistical analysis, and model learning algorithm (ML) to get the result. The researchers first defined the problem by designing a questionnaire based on PSQI and extracted several features such as daily activity, living environment, and social activity from mobile phone data. Then, they proposed a method based on a factor graph for predicting sleep quality using these features. For evaluating the proposed model, they developed the SleepMiner system, which uses mobile as the only client and collects context data in real-time. As a result, the researchers proposed a novel framework for sleep quality prediction based on mobile phone data.

Shailesh Bihari, R. Doug McEvoy, Elisha Matheson, B.Nurs., Susan Kim, Richard J. Woodman, Andrew D. Bersten, conducted a study "Factors Affecting Sleep Quality of Patients in Intensive Care Unit" (2012) in which they evaluated sleep quality among patients admitted to ICU and investigated environmental and non-environmental factors that affect sleep quality in ICU. They proved a hypothesis that interruptions of sleep in the ICU are multifactorial. In addition to previously identified environmental factors, several non-environmental factors, such as prior quality of sleep at home, use of regular sleeping tablets before ICU admission, treatment for hypo-/hyperthyroidism, plus the use of benzodiazepines and steroids administration during the ICU stay were associated with self-reported poor quality sleep. The researchers used statistical analysis to get the result; in particular, Mixed-effects linear regression was used to assess whether sleep quality was different at different time points with the patient as a random effect (random intercept). A linear regression model was used to examine the effect of length of stay on sleep quality and daytime sleepiness, and a generalized linear model was used to assess factors affecting the ICU's quality of sleep.

Daniel J. Buysse, Martica L. Hall, Patrick J. Strollo, Thomas W. Kamarck, Jane Owens, Laisze Lee, Steven E. Reis, Karen A. Matthews in the paper "Relationships Between the Pittsburgh Sleep Quality Index (PSQI), Epworth Sleepiness Scale (ESS), and Clinical/Polysomnographic Measures in a Community Sample" (2008) used correlation, cluster analysis, principal components analysis, MANOVA, ANOVA, and regressions to characterize the relationships between the PSQI, ESS, and other study variables. They investigated that PSQI and ESS correlated weakly with each other but segregated from principal components analysis. Groups of participants categorized by either cluster analysis of PSQI and ESS scores or by scores above or below traditional cut-off values differed from each other on psychological/stress symptoms and quantitative and qualitative sleep diary measures, but not on actigraphic or polysomnographic measures. The PSQI and ESS measure sleep-wake symptoms orthogonal dimensions, but neither is related to objective sleep measures. The researchers concluded that the PSQI is more closely related to psychological symptom ratings and sleep diary measures than the ESS.

Y.Liu, T.Li, L.Guo, R.Zhang, X.Feng in "The mediating role of sleep quality on the relationship between perceived stress and depression among the elderly in urban communities: a cross-sectional study" (2017) investigated that reveals that not all dimensions of sleep quality are relevant factors affecting depression in the elderly. It is because there may be partial mediation effects of sleep quality, mainly through sleep duration and daytime dysfunction, within the impact of perceived stress on depression. This signifies that coping with perceived stress can be expected to ameliorate the severity of depression in the elderly by the intermediary role of sleep quality and the direct effect. The researchers conducted a cross-sectional survey among 1050 community residents older than 60 years in China to get the results. To estimate perceived stress, depression and sleep quality, and depression, the Perceived Stress Scale, Epidemiological Studies Depression Scale, and the Pittsburgh Sleep Quality Index were used, respectively. Data from surveys were analyzed with correlation, multiple linear regression, and structural equation modeling.

Karen Caldwell, Mandy Harrison, Marianne Adams, Rebecca H. Quin MA, and Jeffrey Greeson in "Developing Mindfulness in College Students Through Movement-Based Courses: Effects on

Self-Regulatory Self-Efficacy, Mood, Stress, and Sleep Quality" (2020) investigated whether mindfulness increased through participation in movement-based courses and whether mood changes and perceived stress the relationship between increased mindfulness and better sleep. Participants were 166 college students enrolled in the 2007–2008 academic year in 15-week classes. At the beginning, middle, and end of the semester, participants completed measures of mindfulness, self-regulatory self-efficacy, mood, perceived stress, and sleep quality. As a result, total mindfulness scores and mindfulness subscales increased overall.

6 Methodology explanation

6.1 Theoretical information

6.1.1 OLS estimation

Ordinary Least Squares (OLS) is a widely used method for estimating the unknown parameters in a linear regression model.

In OLS models, the parameters are chosen as a linear function of a set of explanatory variables by the principle of least squares, i.e., minimizing the squares of the differences between the observed dependent variable in the dataset those predicted by the linear function.

$$SSR = \sum_{i=1}^{n} u^2 \tag{1}$$

Estimating Single-Independent Variable Model with OLS

Theoretical model

$$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_{1,i} \tag{2}$$

Empirical model

$$y_i = \beta_0 + \beta_1 x_{1,i} + u_i \tag{3}$$

where
y_i - dependent variable
x_i - independent variable
β_0- y-intercept (constant term)
β_1- slope coefficient
u_i - the model's error term (residuals)

6.1.2 Multiple Linear Regression Models

Multiple linear regression (MLR), or multiple regression, is a statistical technique that uses several explanatory variables to predict a response variable's outcome. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

Multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

Formula of Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + u_i \tag{4}$$

where
y_i - dependent variable
x_i - independent variable
β_0- y-intercept (constant term)
β_1- slope coefficient
u_i – the model's error term (residuals)

6.1.3 R-squared

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable.

The Formula for R-Squared is

$$\mathrm{R}^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} \tag{5}$$

R^2 values range from 0 to 1 and are stated as percentages from 0% to 100%.

6.1.4 Classical Assumptions for OLS estimations

OLS estimators would be the best available if data met classical assumptions.

The seven classical assumptions are:

1. The regression model is linear, correctly specified, and has an additive error term.
2. The error term has a zero population mean.
3. All explanatory variables are uncorrelated with the error term.
4. Observations of the error term are uncorrelated with each other (no serial correlation).
5. The error term has a constant variance (no heteroskedasticity).
6. No explanatory variable is a perfect linear function of any other explanatory variable(s) (no perfect multicollinearity).
7. The error term is normally distributed (this assumption is optional but usually is invoked).

6.1.5 Hypothesis testing

Hypothesis testing is used in a variety of settings.
If a hypothesis cannot be proved, the hypothesis can be rejected.
There are two hypothesis types:
- Null hypothesis (H_0): the outcome that the researcher does not expect.
- Alternative hypothesis (H_a): the outcome the researcher does expect.

t-test

$$t = \frac{\beta_j}{\mathrm{SE}(\hat{\beta}_j)} \tag{6}$$

Confidence Interval

The confidence interval (CI) contains a range of values that is likely to include a population value with a certain degree of confidence. CI is often expressed as a %. Thus, a population mean lies between an upper and lower interval.

A $(1-\alpha)\%$ confidence interval is defined as $\hat{\beta}_j \pm c * SE(\hat{\beta}_j)$, where c is the $(1-\alpha/2)$ percentile in a t_{n-k-1} distribution.

P-values

P-value is the probability that the null hypothesis is true.

6.2 Practical information

This exploration aims to determine what factors have a significant impact on sleeping quality using linear models and tests for statistical significance.

For the exploration, Multiple Linear Regression models with OLS estimators were chosen.

The data used in the analysis are cross-sectional and were taken from the MMASH database. The independent variables were chosen based on a review of the literature and intuitively. The dependent variable was selected based on researches about sleep quality metrics.

All variables were analyzed and checked on Gauss-Markov Assumptions. For each variable in the models, the statistical significance of this variable was determined. The conclusion was made using the results of the analysis.

The analysis was performed using Python and R software. All graphs, tables, and code are attached.

7 Data description

Many different factors can affect sleep quality. To build the most accurate models, I decided to include various data.

Dependent variable

PSQI: It is a measure of sleep quality. It is a self-report questionnaire that assesses sleep quality. The measure consists of 19 individual items, creating seven components that produce one global score, and takes 5–10 minutes to complete. A score can be ranged from 0 to 21, where lower scores denote a healthier sleep quality.

Independent variables

Age: I suppose that the lower age is, the higher the sleep quality is. Since the younger a person is, the more time he needs to sleep.

Weight, height: I assume that the lower the variables are, the lower are sleep quality.

BMI: It measures Body Mass Index and is calculated as $weight/(height^2)$. The norm is $18.5 - 24.9$, and I think if a person's BMI value is out of range of the interval, the person can have low sleep quality.

Cortisol before, cortisol after, melatonin before, melatonin after sleep: The variables measure cortisol and melatonin levels before and after sleep. Cortisol is a "happy hormone," and melatonin is responsible for skin health. Generally, melatonin is produced while a person is sleeping. So, I suppose that the higher these parameters are, the higher the sleep quality will be.

Latency_efficiency, total_minutes_in_bed, total_sleep_time: They measure the percentage of sleep time on total sleep in a bed, minutes spent in the bed per night, and length of the sleep per night expressed in minutes, respectively. These three variables are connected and refer to sleep duration time. So, I hypothesize that the higher they are, the higher sleep quality is.

RR-mean: It reflects the mean of time in seconds between two consecutive heartbeats during a day. I assume that the high RR-mean corresponds to high sleep quality.

Daily Stress Inventory value (DSI): It shows 58 items of self-reported measures indicating the events a person experienced in the last 24 hours. It gives a score between 0 and 406. The higher is these values, the higher is the frequency and degree of the events and the perceived daily stress. So, I suppose that the lower is DSI, the higher the sleep quality will be.

Activity: medium, Activity: heavy: The variables reflect a time duration during which a person was a fast walk and bike or having a gym and running. I hypothesize that the higher the value is, the higher the sleep quality is.

Activity: small screen usage, Activity: large screen usage: They measure the time duration during which a person was watching small and large screen devices. I assume that these lower values correspond to higher sleep quality.

Activity: smoking, Activity: alcohol assumption: The variables show a time duration when a person was smoking or drinking alcohol. I think that the higher the variables are, the higher sleep quality will be.

Actigraph:X-mean, Actigraph:Y-mean, Actigraph:Z-mean: The variables refer to a mean of raw acceleration data of the X-, Y-, Z-axis, respectively, during a day. I hypothesis that the higher these values correspond to high sleep quality.

Steps-mean: It measures a mean of a step number per second during a day. I assume that the higher the value is, the higher the sleep quality be.

Other **data characteristics** are listed in the table below.

Variable	Abbreviation	Units
PSQI	PSQI	index
Age	Age	Years
Gender	Gender	M/F
Weight	Weight	Kg
Height	Height	cm
Body mass index	BMI	index
Cortisol before sleep	Cortisol_before	μg of cortisol per 100 μg of protein
Cortisol after sleep	Cortisol_after	μg of cortisol per 100 μg of protein
Melatonin before sleep	Melatonin_before	μg of melatonin per μg of protein
Melatonin after sleep	Melatonin_after	μg of melatonin per μg of protein
Latency efficiency	Latency_Efficiency	%
Total time in bed	Total_minutes_in_bed	minutes
Total sleep time	Total_sleep_time	minutes
Beat-to-beat interval	RR	seconds
Daily Stress Inventory value	Daily_stress	index
Activity: medium	Activity_medium	minutes
Activity: heavy	Activity_heavy	minutes
Activity: small screen usage	Activity_small_screen_usage	minutes
Activity: large screen usage	Activity_large_screen_usage	minutes
Activity: smoking	Activity_smoking	minutes
Activity: alcohol assumption	Activity_alcohol_assumption	minutes
Actigraph X-mean	Actigraph_X_mean	Newton-meter
Actigraph Y-mean	Actigraph_Y_mean	Newton-meter
Actigraph Z-mean	Actigraph_Z_mean	Newton-meter
Steps-mean	Steps_mean	steps per second

Summary statistics for the described variables are shown in the table below.

Abbreviation	Observations	Mean	St.Dev.	Min	Max
PSQI	22	5.318182	1.985336	2	9
Age	22	27.272727	4.107853	20	40
Weight	22	75.045455	12.789420	60	115
Height	22	179.909091	8.216760	169	205
BMI	22	23.123541	3.093798	19.409358	33.240837
Cortisol_before	22	0.028053	0.029735	0.012017	0.155777
Cortisol_after	22	0.069860	0.051937	0.015572	0.261252
Melatonin_before	22	8.330234e-09	6.545962e-09	1.629907e-09	2.396239e-08
Melatonin_after	22	7.281474e-09	6.042238e-09	8.283802e-10	2.853905e-08
Latency_Efficiency	22	83.520476	6.498276	73.490000	94.230000
Total_minutes_in_bed	22	382	89.180395	165	630
Total_sleep_time	22	318.571429	80.157328	144	578
RR	22	0.813643	0.072447	0.670949	0.995826
Daily_stress	22	32.181818	16.296761	10	74
Activity_medium	22	20.909091	33.863745	0	129
Activity_heavy	22	93.590909	32.854164	40	180
Activity_small_screen_usage	22	57.363636	54.862204	0	174
Activity_large_screen_usage	22	17.500000	17.065002	0	51
Activity_smoking	22	7.272727	16.292510	0	60
Activity_alcohol_assumption	22	10.318182	11.047027	0	45
Actigraph_X_mean	22	16.434876	16.434876	3.131175	24.693821
Actigraph_Y_mean	22	16.780694	4.245410	10.515468	27.243074
Actigraph_Z_mean	22	19.271554	4.202566	13.269682	29.023668
Steps_mean	22	0.184510	0.047359	0.118963	0.287923

Gauss-Markov Assumptions

1.*Linear in Parameters*

The dependent variable can be written as a linear combination of the independent variables plus an error term. The scatter graphs between each of the independent variables and the dependent variable *(included in Appendix)*. All models in this paper satisfy this assumption.

2.*Random sampling*

The data has random sampling since it reflects variables for 22 adult males.

3.*Zero Conditional Mean*

This assumption means that the independent variables contain no information about the size of the error.

4. *No serial errors correlations*

The error terms are uncorrelated with each other.

5.*Homoskedasticity*

The independent variables contain no information about the variance of the error. It is shown by scatter plots between each of the independent variables and the dependent variable.

6.*No perfect collinearity*

Correlations between independent variables are not high *(Correlation Matrix included in Appendix)*. There is no perfect collinearity between the independent variables.

So, all models below meet Gause-Markov Assumptions.

8 Results

Model 1: First Multiple Regression

Equation:

PSQI = β_0+$\beta_1 Weight$+$\beta_2 Height$+$\beta_3 Melatonin\_before$+$\beta_4 Melatonin\_after$+$\beta_5 Latency\_Efficiency$+ $\beta_6 Activity\_medium$+$\beta_7 Activity\_large\_screen\_usage$+$\beta_8 Actigraph\_X\_mean$+$\beta_9 Actigraph\_Z\_mean$

After Regression:

PSQI = -3.47 - 1.33 Weight + 2.57 Height + 1.6 Melatonin_before + 6.77 Melatonin_after - 9.92 Latency_Efficiency - 4 Activity_medium - 4.74 Activity_large_screen_usage + 2.01 Actigraph_X_mean - 1.18 Actigraph_Z_mean

N = 22

R^2 = 0.557

Table 1: Estimation Results - Model 1

Variable	Coefficient	Std. Error	T-value	P-value	H0:b=0 H1:b/=0
(Intercept)	-3.471e+01	1.787e+01	-1.942	0.0759	Fail to reject at 10%
Weight	-1.326e-01	7.110e-02	-1.865	0.0868	Fail to reject at 10%
Height*	2.570e-01	1.080e-01	2.380	0.0348	Reject at 10%
Melatonin_before	1.595e+08	8.386e+07	1.901	0.0815	Fail to reject at 10%
Melatonin_after	6.769e+07	7.733e+07	0.875	0.3986	Fail to reject at 10%
Latency_Efficiency	-9.920e-02	8.032e-02	-1.235	0.2404	Fail to reject at 10%
Activity_medium	-4.003e-02	2.646e-02	-1.513	0.1563	Fail to reject at 10%
Activity_large_screen_usage	4.737e-02	4.512e-02	1.050	0.3145	Fail to reject at 10%
Actigraph_X_mean*	2.006e+00	7.670e-01	2.616	0.0226	Reject at 10%
Actigraph_Z_mean*	-1.182e+00	4.762e-01	-2.481	0.0289	Reject at 10%

(*Statistically Significant at 10%, **Statistically Significant at 5%, ***Statistically Significant at 1%)

Interpretation of Model 1

For the first model, independent variables were chosen from a model with all independent variables removing ones with the highest p-values.

In the model, only Height, Actigraph_X_mean and Actigraph_Z_mean are statistically significant at 10% confidence interval (or 90% confidence level). So, the Null hypotheses that β_2, β_8 and β_9 are equal to 0 at confidence level 90% can be rejected.

However, $R^2 = 0.557$. It means that the model explains only 55.7% of the PSQI.

Model 2: Second Multiple Regression

Equation:

PSQI = $\beta_0 + \beta_1 Height + \beta_2 Melatonin\_before + \beta_3 Total\_minutes\_in\_bed + \beta_4 RR + \beta_5 Activity\_medium + \beta_6 Activity\_smoking + \beta_7 Actigraph\_X\_mean$

After Regression:

PSQI = - 5.99 + 7.64 Height + 1.63 Melatonin_before + 1.87 Total_minutes_in_bed - 1.99 RR - 1.33 Activity_medium + 1.990 Activity_smoking + 3.26 Actigraph_X_mean

N = 22

$R^2 = 0.761$

Table 2: Estimation Results - Model 2

Variable	Coefficient	Std. Error	T-value	P-value	H0:b=0 H1:b/=0
(Intercept)	-5.988e+00	6.327e+00	-0.946	0.36001	Fail to reject at 10%
Height*	7.643e-02	3.389e-02	2.255	0.04065	Reject at 10%
Melatonin_before**	1.628e+08	4.470e+07	3.641	0.00267	Reject at 5%
Total_minutes_in_bed***	1.871e-02	3.450e-03	5.422	8.99e-05	Reject at 1%
RR**	-1.988e+01	4.899e+00	-4.057	0.00118	Reject at 5%
Activity_medium	-1.334e-02	9.953e-03	-1.341	0.20141	Fail to reject at 10%
Activity_smoking	1.990e-02	1.702e-02	1.169	0.26186	Fail to reject at 10%
Actigraph_X_mean**	3.262e-01	9.091e-02	3.587	0.00297	Reject at 5%

(*Statistically Significant at 10%, **Statistically Significant at 5%, ***Statistically Significant at 1%)

Interpretation of Model 2

For the second model, independent variables were selected in a next way: firstly, were created all 245157 combinations with 7 from 23 variables, then for every combination was created a regression model and from all models a one with the highest R-squared was chosen.

In the model, Height is are statistically significant at 10% confidence interval, Melatonin_before, RR and Actigraph_X_mean at 5% confidence interval and Total_minutes_in_bed at 1% confidence interval. So, the Null hypothesis that β_1 is equal to 0 at confidence level 90% can be rejected as well as β_2, β_4 and β_6 are equal to 0 at confidence level 95% and β_2 is equal 0 at confidence level 99%.

R-squared is equal to 0.761. That means that the model explains 76.1% of the PSQI. That is higher than the previous one, and the model has more degrees of freedom.

Model 3: Second Multiple Regression

Equation:

PSQI = $\beta_0 + \beta_1 Age + \beta_2 Weight + \beta_3 Melatonin\_before + \beta_4 Total\_minutes\_in\_bed + \beta_5 RR + \beta_6 Activity\_heavy + \beta_7 Actigraph\_X\_mean + \beta_8 Steps\_mean$

After Regression:

PSQI = -6.13 + 3.15 Age + 1.04 Weight + 1.55 Melatonin_before + 2.3 Total_minutes_in_bed -2.32 RR + 1.83 Activity_heavy + 6.1 Actigraph_X_mean - 4.27 Steps_mean

N = 22

$R^2 = 0.813$

(*Statistically Significant at 10%, **Statistically Significant at 5%, ***Statistically Significant at 1%)

Interpretation of Model 3

For the third model, independent variables were selected in a next way: firstly, were created all 490314 combinations with 8 from 23 variables, then for every combination was created a regression model and from all models a one with the highest R-squared was chosen.

Table 3: Estimation Results - Model 3

Variable	Coefficient	Std. Error	T-value	P-value	H0:b=0 H1:b/=0
(Intercept)	-6.128e+00	5.017e+00	-1.221	0.243653	Fail to reject at 10%
Age*	3.145e-01	1.269e-01	2.478	0.027711	Reject at 10%
Weight**	1.042e-01	2.960e-02	3.522	0.003754	Reject at 5%
Melatonin_before**	1.551e+08	4.132e+07	3.755	0.002405	Reject at 5%
Total_minutes_in_bed***	2.297e-02	3.709e-03	6.192	3.26e-05	Reject at 1%
RR***	-2.322e+01	4.803e+00	-4.836	0.000325	Reject at 1%
Activity_heavy	1.835e-02	9.223e-03	1.990	0.068034	Fail to reject at 10%
Actigraph_X_mean*	6.097e-01	2.238e-01	2.724	0.017375	Reject at 10%
Steps_mean*	-4.263e+01	1.795e+01	-2.374	0.033657	Reject at 10%

In the model, Age, Actigraph_X_mean and Steps_mean are statistically significant at 10% confidence interval, Weight and Melatonin_before at 5% confidence interval and RR and Total_minutes_in_bed at 1% confidence interval. So, the Null hypothesis that $\beta_6 = 0$ fails to reject at 10% confidence interval.

The model explains 81.3% of the PSQI. That is better than in previous models, and the model has more degrees of freedom.

Model 4: Second Multiple Regression

Equation:

PSQI = β_0+$\beta_1 Age$+$\beta_2 Weight$+$\beta_3 Melatonin\_before$+$\beta_4 Latency\_Efficiency$+$\beta_5 Total\_minutes\_in\_bed$+ $\beta_6 RR + \beta_7 Activity\_heavy + \beta_8 Actigraph\_X\_mean + \beta_9 Steps\_mean$

After Regression:

PSQI = -2.97 + 3.67 Age + 1.01 Weight+1.49 Melatonin_before -4.96Latency_Efficiency+ 2.35 Total_minutes_in_bed -2.4 RR+ Activity_heavy+ 7.51 Actigraph_X_mean -5.39 Steps_mean

N = 22

$R^2 = 0.83$

Table 4: Estimation Results - Model 4

Variable	Coefficient	Std. Error	T-value	P-value	H0:b=0 H1:b/=0
(Intercept)	-2.972e+00	5.751e+00	-0.517	0.61462	Fail to reject at 10%
Age*	3.670e-01	1.347e-01	2.724	0.01848	Reject at 10%
Weight**	1.013e-01	2.949e-02	3.436	0.00493	Reject at 5%
Melatonin_before**	1.483e+08	4.147e+07	3.576	0.00381	Reject at 5%
Latency_Efficiency	-4.961e-02	4.524e-02	-1.096	0.29441	Fail to reject at 10%
Total_minutes_in_bed***	2.354e-02	3.718e-03	6.332	3.76e-05	Reject at 1%
RR***	-2.399e+01	4.816e+00	-4.980	0.00032	Reject at 1%
Activity_heavy	1.814e-02	9.154e-03	1.981	0.07095	Fail to reject at 10%
Actigraph_X_mean*	7.515e-01	2.570e-01	2.924	0.01274	Reject at 10%
Steps_mean*	-5.388e+01	2.056e+01	-2.621	0.02236	Reject at 10%

(*Statistically Significant at 10%, **Statistically Significant at 5%, ***Statistically Significant at 1%)

Interpretation of Model 4

For the fourth model, independent variables were selected in the next way: firstly, were created all 817190combinations with 9 from 23 variables, then for every combination was created a regression model and from all models a one with the highest R-squared was chosen.

In the model, Age, Actigraph_X_mean and Steps_mean are statistically significant at 10% confidence interval, Weight and Melatonin_before at 5% confidence interval and RR and Total_minutes_in_bed at 1% confidence interval. So, only the Null hypotheses that $\beta_4 = 0$ and $\beta_7 = 0$ fail to reject at 10% confidence interval.

$R^2 = 0.83$. So, the model explains 83% of the PSQI. The R-squared is the highest compared to the previous models, but the model has smaller degrees of freedom than the previous two.

9 Possible improvements

The models can be improved in the following ways:

- addin variables (lighting, body position, sensory monotony, temperature, atmospheric pressure...);
- taking into account time factors (for example, time of X activity before Y minutes/ours to sleep);
- sample clustering (division into categories with specific features).

Possible areas for further researches:

- increasing the sample of the database;
- using time-series and other models.

10 Conclusion

This project explores the factors contributing to sleep quality, which is measured as a PSQI. In order to take into different dimensions, I used such variables as age, weight, height, BMI, cortisol and melatonin before sleep, percentage of sleep time on total sleep in bed, total time in bed, total sleep time, heart beat-to-beat interval, Daily Stress Inventory value, Activity: medium, heavy, small screen usage, large screen usage, smoking, alcohol assumption, Actigraph X-, Y-, Z-mean and steps-mean.

Many models were created, and four of them with the highest R-squares for specific degrees of freedom were included in the paper.

Model 1, in which independent variables were chosen from a model with all independent variables removing ones with the highest p-values (12 degrees of freedom), has not the best R-squared (only 0.557). The model includes such variables as Weight, Height, Melatonin_before, Melatonin_after, Latency_Efficiency, Activity_medium, Activity_large_screen_usage, Actigraph_X_mean and Actigraph_Z_mean. Only 3 of them (Height, Actigraph_X_mean and Actigraph_Z_mean) are statistically significant at a confidence level of 90%.

Models 2, 3 and 4 were constructed in the following way: firstly, were created all combinations with 7, 8 and 9 (respectively) from 23 variables, then for every combination was created a regression model and from all models, a one with the highest R-squared was chosen. These three models have degrees of freedom not lower the first model, but their R-squares are higher.

So, Model 4 has the highest R-squared among the other models in this paper (83%). It includes Age, Weight, Melatonin_before, Latency_Efficiency, Total_minutes_in_bed, RR, Activity_heavy, Actigraph_X_mean and Steps_mean variables. In this model, Age, Actigraph_X_mean and Steps_mean are statistically significant at 10% confidence interval, Weight and Melatonin_before at 5% confidence interval and RR and Total_minutes_in_bed at 1% confidence interval. Contrary to my assumption, increasing age correlates with an increase in PSQI; increases in RR and Steps-mean correlated with decreasing PSQI in the model. My hypotheses that increasing in Weight, Melatonin_before, Total_minutes_in_bed and Actigraph_X_mean correspond to increases in PSQI were not rejected.

In the models above, there are omitted many factors that could have an impact on sleep quality. Reducing the missing factors may make the models more relevant. Additionally, more observations would strengthen the models.

References

- Rossi, A., Pozzo, E., Menicagli, D., Tremolanti, C., Priami, C., Sirbu, A., . . . Morelli, D. (2020, June 19). Multilevel Monitoring of Activity and Sleep in Healthy People. From https:/physionet.org/content/mmash/1.0.0/
- Why Do We Need Sleep? (2020, December 11). From https://www.sleepfoundation.org/how-sleep-works/why-do-we-need-sleep
- Root, D. (2020, November 10). Getting Started with Data Science in Python. From https://towardsdatascience.com/getting-started-with-data-science-in-python-92326c171622
- Fernando, J. (2020, November 18). R-Squared. From https://www.investopedia.com/terms/r/r-squared.asp
- Pittsburgh Sleep Quality Index. (2020, November 16). From https://en.wikipedia.org/wiki/Pittsburgh_Sleep_Quality_Index
- Yin Bai, Bin Xu, Yuanchao Ma, Guodong Sun, Yu Zhao, Will You Have a Good Sleep Tonight? Sleep Quality Prediction with Mobile Phone (2020). From https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.837.648
- Bihari, S., Bihari, A., McEvoy, R., Health, A., Matheson, E., Department of Intensive and Critical Care Medicine, . . . PL, W. (2012, June 15). Factors Affecting Sleep Quality of Patients in Intensive Care Unit. From https://jcsm.aasm.org/doi/full/10.5664/jcsm.1920
- Buysse, Hall, Strollo, Kamarck (2008, December 15). Relationships Between the Pittsburgh Sleep Quality Index (PSQI), Epworth Sleepiness Scale (ESS), and Clinical/Polysomnographic Measures in a Community Sample. From https://jcsm.aasm.org/doi/full/10.5664/jcsm.27351
- Liu, Y., Li, T., Guo, L., Zhang, R., Feng, X., amp; Liu, K. (2017, May 18). The mediating role of sleep quality on the relationship between perceived stress and depression among the elderly in urban communities: A cross-sectional study. From https://www.sciencedirect.com/science/article/abs/pii/S0033350617301439
- Developing Mindfulness in College Students Through Movement-Based Courses: Effects on Self-Regulatory Self-Efficacy, Mood, Stress, and Sleep Quality. (n.d.). From https://www.tandfonline.com/doi/abs/10.1080/07448480903540481