**Group 26**

**Academic Grasshoppers**

**Yinjie Wu**

**Yuan Tian**

**Xinran Yu**

# Academic Grasshoppers

*Only interested in the snacks at colloquiums, leaving no food behind!*

***Before*** ***After***

1. **Background knowledge, feedback, and eyes on the field:**
    a. **Literature review**

    We have conducted a literature review on sleep disorders and the potential risk factors associated with them. We reviewed research articles that focus on similar questions in the field of sleep health or employ approaches similar to those in our project. Some of these articles provide background information on sleep disorders and potential risk factors, which are helpful for us to learn more about the possible mechanisms of sleep disorders and how they might be related to many biological, behavioral, and social factors in our lives. Also, from some of these articles, we can gain insights from the approaches the investigators adopted for data analysis and interpretation. Below are some of the insightful papers we have found.

1. Postuma, R.B., et al. "Environmental risk factors for REM sleep behavior disorder." *Neurology*, vol. 79, no. 5, 31 July 2012, pp. 428–434, https://doi.org/10.1212/wnl.0b013e31825dd383.

This paper examines the association between environmental and behavioral factors (e.g. sex, smoking, education, pesticide exposure, etc) on REM sleep behavior disorder. The investigators conducted a case-control study and used logistic regression to analyze their data. This article provides insights into the statistical method for observational study data analysis within the field of sleep disorders. We also drew inspiration regarding the variables that may be related to overall sleep health and sleep duration, which we are investigating in our project.

2. Chennaoui, Mounir, et al. "Sleep and exercise: A reciprocal issue?" *Sleep Medicine Reviews*, vol. 20, Apr. 2015, pp. 59–72, https://doi.org/10.1016/j.smrv.2014.06.008.

This paper discusses the potential relationship between exercise and sleep quality. It is a review article that talks about the potential mechanisms behind this relationship and encourages future research on this topic in-depth. Given the discussion in this article, we would like to add covariates related to exercise and sedentary lifestyle in our analysis.

3. Singareddy, Ravi, et al. "Risk factors for incident chronic insomnia: A general population prospective study." *Sleep Medicine*, vol. 13, no. 4, Apr. 2012, pp. 346–353, https://doi.org/10.1016/j.sleep.2011.10.033.

This paper explores the risk factors of incident chronic insomnia, a common sleep disorder that significantly reduces the quality of life today. This article analyzes many covariates, such as BMI, caffeine and alcohol consumption, mental health, etc. The investigators adopted multivariate logistic regression to model the relationship between covariates and incident chronic insomnia.

4. Whinnery, Julia, et al. "Short and long sleep duration associated with race/ethnicity, Sociodemographics, and socioeconomic position." *Sleep*, vol. 37, no. 3, 1 Mar. 2014, pp. 601–611, https://doi.org/10.5665/sleep.3508.

This paper examines the socioeconomic factors that might be related to sleep duration. The authors used similar NHANES datasets, so the methods they adopted for data analysis can serve as a reference for us. The authors used multinomial logistic regression for data analysis.

5. Ji, Xiaopeng, et al. "The relationship between micronutrient status and sleep patterns: A systematic review." *Public Health Nutrition*, vol. 20, no. 4, 5 Oct. 2016, pp. 687–701, https://doi.org/10.1017/s1368980016002603.

This paper is a meta-analysis of the relationship between dietary micronutrients and sleep patterns. They combined the results from multiple studies to investigate how certain micronutrients such as Fe, Mg, and Zn could have impacts on sleep duration and quality.

**b. Reviewed peers' reviews**

**i) Insights:** We have carefully read all 4 reviews from both teaching staff and classmates.

Feature selection was mentioned in multiple reviews, where we initially included too many potential covariates without determining an association question between a particular covariate and the outcome. Indeed, we did not narrow down our risk factors to a particular association between one feature and one outcome, we were hoping to retrieve the information from data. Some peers found it was a bit confusing on what specific primary question we are asking, some advised that we should determine the covariate through research instead of feature selection methods, and some concerned that we could end up with a model that is too complex to interpret.

Other than covariate selection, dealing with missing data was mentioned in multiple reviews, other than suggesting reading topic note 13, one review proposed to try least squares optimization or MLE in our model.

Another concern was on the outcome of interest, since there is no binary outcome of diagnosed sleep disorder, an alternative outcome of "self-reported ever told doctors had trouble sleeping" was used as a binary categorical outcome. There could be a time difference between actual outcome and examination time. In addition, variables like physical activity can change in weeks, we will need to be careful on reverse causation and measurement errors.

**ii)** For covariate selections, narrowing down the association question to one specific covariate can be done in combination of our interest and some literature review. One fundamental mistake we made was on how we formulated our primary association question, exploring a variety of risk factors and their interaction was pointed out as a prediction task. We had discussed and reviewed literatures, and then narrowed down our primary interest to the association between depression and sleep trouble. Based on literature evidence and data explorations, we adjusted our covariates, specifically, including risk factors meeting classical definition of confounding, such as screen times, vitamins, and blood measurements. In contrast, we removed potential risk factors such as ion concentrations from our primary question, as they do not meet the classical definition of confounding in the question. However, for model complexity, we do not want to remove covariates without carefully examining them. As many phenotypic traits, sleep is not a simple trait that can be determined by a very few factors. We will adjust our covariates based on both literature review (adding and testing potential covariates) and statistical testing (removing covariates). "Don't blame the player, blame the game", we cannot address a complex problem with an over simplified model. Lastly, exploring unexpected risk factors is also more interesting than reviewing already well-known risk factors.

Addressing missing data, we have 3100 individuals with complete information on all 9 potential covariates and outcomes in the 2020 NHANES dataset, we will start with complete case analysis as we have enough statistical power for our results. We will also determine which type of missingness present in our dataset. Lastly, we will perform sensitivity tests to compare complete case analysis and analysis with imputations. For imputation, we will treat it as a prediction problem for the best result.

Unfortunately there is no doctor diagnosed sleep disorder in the 2020 NHANES dataset. For reverse causation, it is an extremely difficult question for us to address based on the data, therefore, we will only model for associations.

I would like to address that we may have confused some reviewers in a secondary question of comparing coefficients of covariates to the dataset from 10 years ago. We do not intend to perform a longitudinal analysis for this matter. We are simply comparing, for example, if alcohol use has a similar or differential effect on sleep nowadays compared to a decade ago. We are applying the same or a similar model

to a similar dataset 10 years ago, to see if there is a significant difference for coefficients.

**3. Analysis Plan:**

Our project is divided into the following steps for data cleaning and subsequent analysis and interpretation.

1. Data Cleaning and Preprocessing

   NHNAES data are classified into several categories based on the type of data collected from study participants, including demographics data, dietary data, examination data, laboratory data, and questionnaire data. Each category is further divided into multiple sections focusing on different measurements.

- We imported the NHANES raw data relevant to sleep disorders and potentially associated risk factors, including sleep disorders and health measurements, standard biochemistry profile, physical activity data, total nutrient intake profile, demographic data, and other laboratory data.
- NHANES raw data from the selected sections will be combined based on the individual-specific ID to generate a profile containing all the information needed for subsequent analysis.
- Generate summaries of raw data to detect possible missing values, wrong data types. Then we remove all rows with NAs and change date types according to the summaries. We will analyze both on complete cases only and imputation data separately, and perform sensitivity analysis afterwards.

2. Exploratory Data Analysis

- Generate summary statistics again to gain an initial understanding of the distribution of outcomes and predictor variables.
- Visualization of outcome variables
  Create a barplot to visualize the distribution of the categorical outcome variable (sleep disorder) and a histogram to visualize the distribution of the continuous outcome variable (sleep duration).
- Visualization of predictor variables
  Create histograms to examine the distribution of continuous predictor variables (BMI, blood pressure, concentration of electrolytes in blood, sedentary activity, etc) and boxplots to examine the distribution of categorical predictor variables (alcohol use, etc). Identify possible skewness.
- Create boxplots to examine the initial patterns in sleep quality measurements for each covariate.

Create a matrix of scatterplots and scatter plots with a loess curve for each covariates to check multicollinearity and possible important covariates.

3. Model Development and Selection

- Assess multicollinearity among covariates using VIF
- Identify regression models
    (1) Logistic regression model: We will model the diagnosis of sleep disorders against depression with identified covariates as potential confounders
    (2) Linear regression model: We will model sleep quality metrics against identified covariates
- Examine potential confounding and effect modification
    (1) Confounders: we will check the classical and operational definition for several potential confounders, such as age, sex, etc.
    (2) Effect modifiers: we will test possible interactions that exist by comparing the statistical results produced with and without interaction terms and the significance of interaction terms.
- Model selection
  For association tasks:
    (1) We will use Elastic Net for variable selection coefficient estimation for our linear regression model. This process allows us to determine which variables to be kept and which statistically insignificant variables can be removed from our models. We will experiment on hyperparameter alpha to search for lower AIC values.
    (2) We will also search for lower AIC values with stepwise, forward, backward selection for model comparison.
    (3) Automatic methods like GAM and automatic spline function will be performed as a reference of AIC/BIC values for comparison.
    (4) Model complexity in terms of interpretability will also be determined in a subjective manner.

- Model diagnostics
    (1) We will conduct analyses (e.g. residual analysis for linear models) to check whether the underlying assumptions of logistic and linear models are satisfied.
    (2) We will also perform sensitivity analysis (e.g. Cook's distance, DFFITS, DFBETAS) to closely examine the leverage and outlying-ness of highly influential data points. Comparison of models with and without these influential points enable us to decide whether or not we can remove these data points.

4. Model Interpretation

- Using the statistically significant covariates identified through the methods discussed above, we will explain how these variables are associated with sleep disorders based on our results.

5. Investigation of Secondary Questions

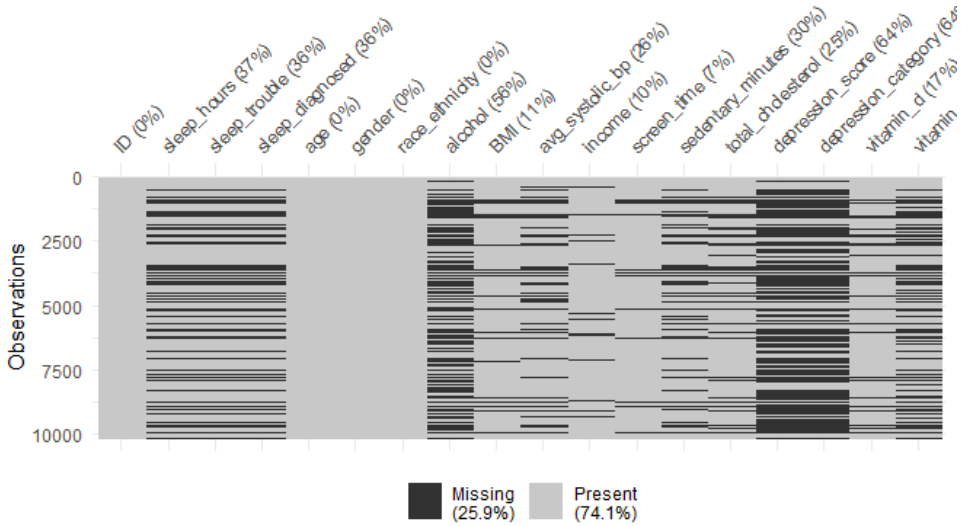
- We will apply the final form of our regression models, using the same covariates, to analyze datasets from previous years (NHANES data from 13-14 years ago). We will compare the coefficients of the two models and interpret the differences. Then we may also explore factors that are less significant now but could be more impactful in 2013-2014.
- Based on the full model, we may be able to explain the effect of subset of covariates such as concentrations of ions on sleep duration and sleep disorder. We may consider doing further explanation using other methods.
- We will also assess how effective our interpretable model can predict whether a person has sleep disorders based on the set of measurements included as covariates, by comparing to full, complicated regression models, and compared to non regression models. Performance metrics may include accuracy, R^2, AUC, and/or etc.. for predictive tasks.

6. Discussion of Model Results and Key Findings

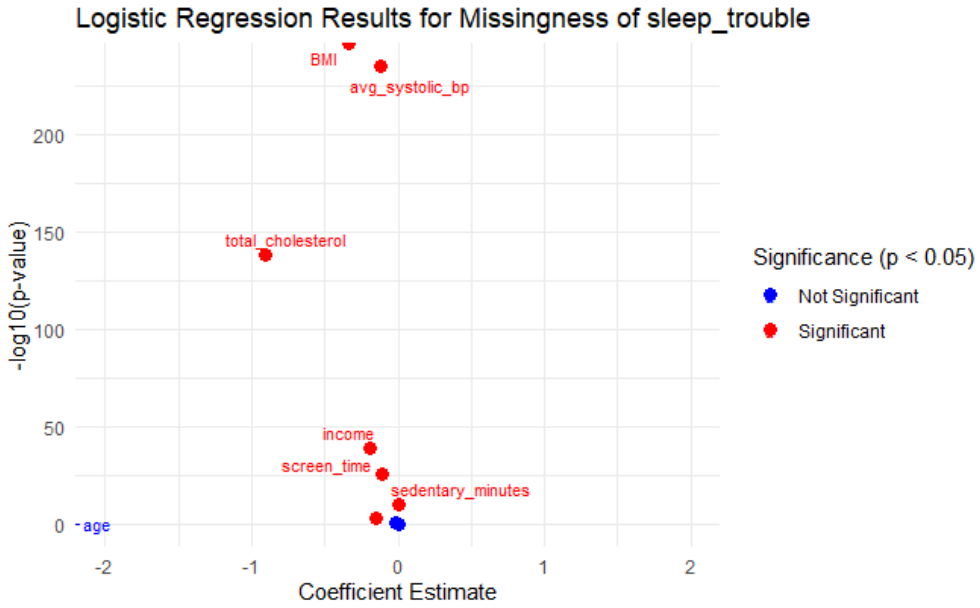
- We will incorporate scientific knowledge (based on advice from experts in epidemiology and biomedical science) to assess our final fitted models within the context of sleep health.
- We will also discuss the limitations to our current models and directions for future works

4. **Missing Data:**

   a. **Type of missing data:**
   First we visualize missingness across all columns including both covariates and outcomes of interest, we observed that we have more than 5% missingness on our outcome of interest, and less than 20% of individuals with complete data.




We created missingness indicators for each column and performed logistic regressions on the missingness indicator $R$ for each covariate $X$ and observed statistical significance between R and many covariates, including gender, BMI, average systolic blood pressure, income, screen time, cholesterol, and sedentary time. We can deny missing completely at random (MCAR).




We cannot really assess MAR or MNAR with formal statistical tests, however, from simple logistic regressions on $R$ and covariates $X$, we believe we have $Y$ missing at random. We do not exclude that our data could be MNAR as well, we will explore MNAR with sensitivity analysis.

**b. Steps might take:**

We performed multiple imputation using the mice package, generated 5 imputations and took their average values. We will then perform sensitivity analysis and compare to complete case analysis to assess how different assumptions about missingness may affect our results.

**5. Modeling:**

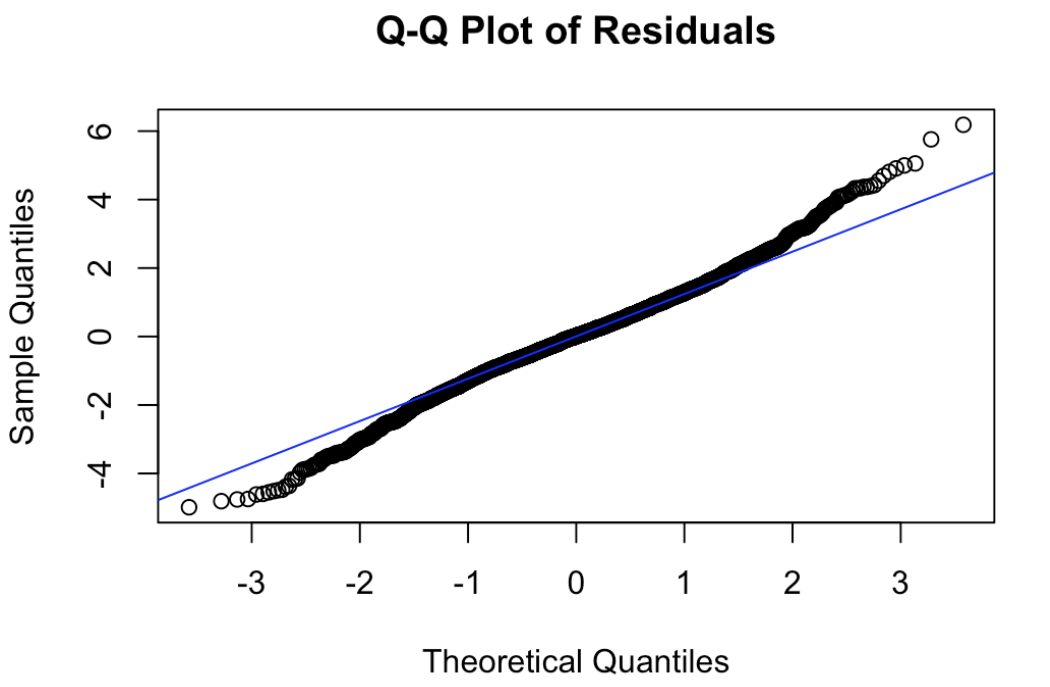
**a. Linear, flexible/additive, or other methods (LASSO, ridge) from this topic:**

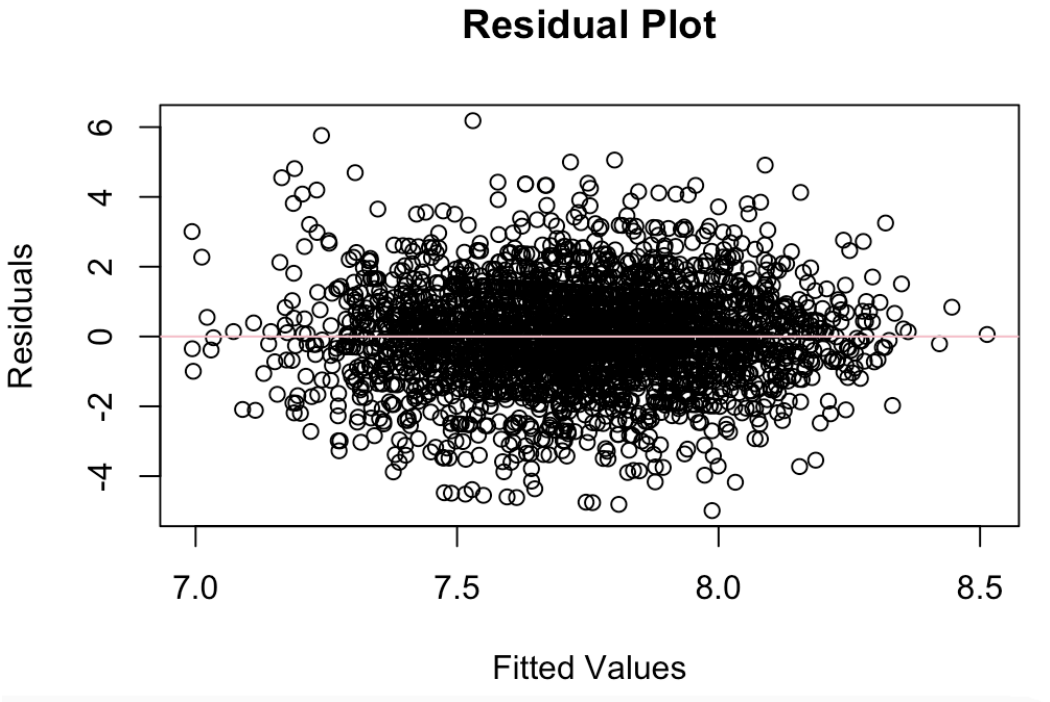
- Linear model is used in one of the secondary questions of our project. Our goal is to explore a number of electrolytes and biochemicals and examine how they might be related to sleep duration.
- In this section, sleep duration is a continuous outcome variable. It is measured independently for each study participant.
- We performed an elastic net to obtain the covariates (biochemicals) that are most strongly associated with sleep duration.
- Our model is:

$$\begin{aligned}\text{Sleep duration}_i = \beta_0 + \beta_1 \text{BMI}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Race}_i + \beta_4 \text{Income}_i + \beta_5 \text{Age}_i + \beta_6 \text{Phosphorus}_i + \beta_7 \text{Potassium}_i \\ + \beta_8 \text{Chloride}_i + \beta_9 \text{Calcium}_i + \beta_{10} \text{Iron}_i + \beta_{11} \text{Glucose}_i + \beta_{12} \text{Cholesterol}_i + \beta_{13} \text{Depression Score}_i + e_i\end{aligned}$$

- Based on the result, we can interpret several important covariates associated with sleep duration:
  Calcium: For every 1 mmol/L increase in calcium in blood, sleep duration tends to increase by 0.433 hours, holding other covariates constant.
  Potassium: For every 1 mmol/L increase in potassium in blood, sleep duration tends to increase by 0.152 hours, holding other covariates constant.
  Total cholesterol: For every 1 mmol/L increase in calcium in blood, sleep duration tends to decrease by 0.0207 hours, holding other covariates constant.
  Phosphorous: For every 1 mmol/L increase in potassium in blood, sleep duration tends to increase by 0.0652 hours, holding other covariates constant.

- By drawing the residual plot and QQ-plot, we checked the assumptions of the linear regression model. The plots show that the relationship between Y and X covariates is linear, Y is normally distributed around its mean, and Y has equal variance for all X.


Q-Q Plot of Residuals
Sample Quantiles
Theoretical Quantiles
-4
-2
0
2
4
6
-3
-2
-1
0
1
2
3

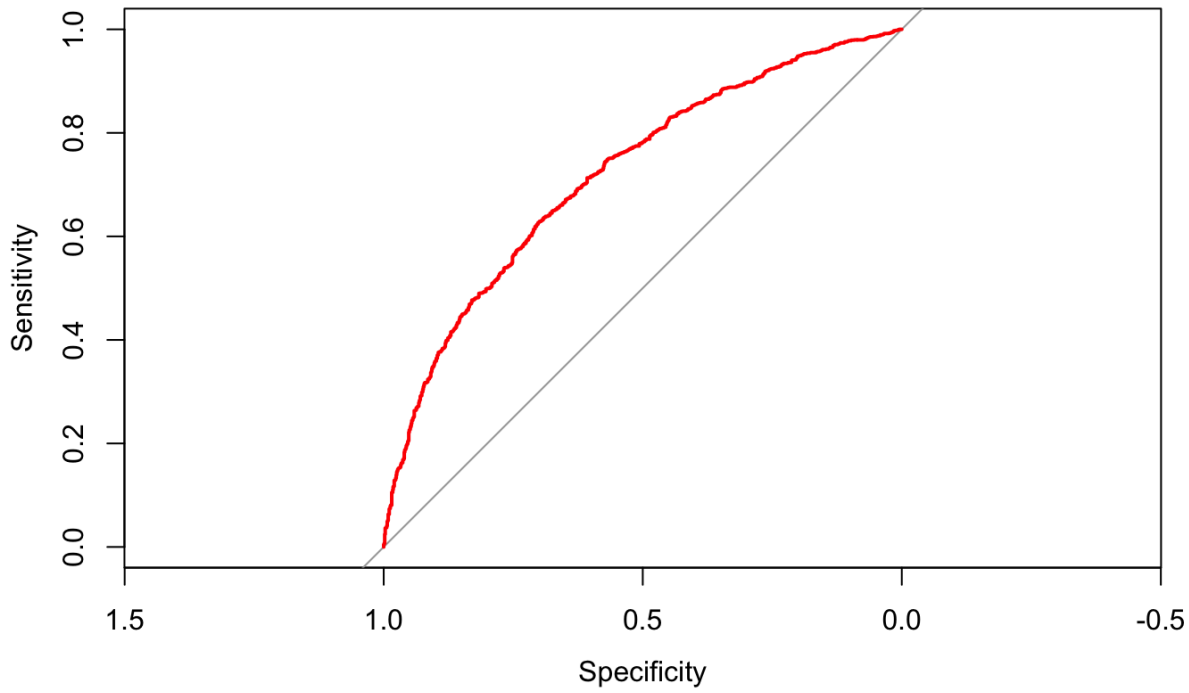

Residual Plot
Residuals
Fitted Values
-4
-2
0
2
4
6
7.0
7.5
8.0
8.5

**b. Logistic, multinomial, ordinal, generalized ordinal:**

- In our primary question, we use multiple logistic regression to model the association between sleep trouble and depression score, adjusting for potential confounders and effect modifiers
- Currently, the final model is:

$$\log\left(\frac{p_{sleep\ disorder}}{1-p_{sleep\ disorder}}\right) = \beta_{k,0} + \beta_{k,1}depression_i + \beta_{k,2}BMI_i + \beta_{k,3}alcohol_i + \beta_{k,4}Race_i + \beta_{k,5}age_i + \beta_{k,6}age_i^2 + \beta_{k,7}income_i + \beta_{k,8}\text{avg\_MAP}_i + \beta_{k,9}\text{total\_cholesterol}_i$$

- Y is an independent and identically distributed (iid) Bernoulli random variable. E(Y) = P(Y=1) = np. No parameter under the logit scale appears as an exponent or is multiplied or divided by another parameter. We have more than 20 samples in each covariate predictor.
- The odds of ever having sleep trouble is estimated to be exp(0.1377799) = 1.147723 times the odds of ever having sleep trouble among people with a one-unit higher depression score on average, holding other covariates constant, according to these sample data.
- ROC:




- Area under the curve: 0.7198. The higher the AUC is, the better the model is at predicting 0's correctly and 1's correctly so in this case the model is doing quite well in predicting individuals who have sleep trouble versus those who don't.

c. **Poisson and Extensions:**
   No, our data and outcome of interest doesn't have any count data. The main objective is a binary outcome. Although the number of sleep hours can be rounded to integers, this is not a rare event and the distribution is approximately normal with an average of 7.7 hours. Thus, using poisson is not a good idea for our project.
d. **Survival analysis:**
   We are not sure yet and most likely will not consider using it after the upcoming week. Our datasets are cross-sectional, with no time variable. However, we have no prior knowledge of survival analysis models, we do not exclude any possibility that survival models can be used in our datasets.

6. Writing:


Abstract

Sleep is an essential process for mental health and cognitive functioning, and it has been shown to be involved in numerous brain functions, such as memory consolidation, brain waste removal, regulation of immune responses, and energy conservation. Insufficient sleep can lead to increased risks of mental illness and physiological diseases, thereby affecting the overall quality of life. Hence, it is crucial to understand the biopsychosocial factors potentially associated with sleep quality and what roles they play in the development of sleep disorders. In this research, we explored the possible relationship between depression and sleep disorders, given the fact that depression has become a common mental health condition affecting large numbers of people today. We utilized publicly available national health and nutrition examination survey (NHANES) datasets and applied logistic regression models to illustrate the association between depression and self reported sleep trouble, while adjusting for any potential confounding and effect modifications. According to our model, the odds of ever having sleep trouble is estimated to be 1.13 times the odds of ever having sleep trouble among people with a one-unit higher depression score. In addition, we found the association between depressive score and sleep trouble remained the same using the 2014 NHANES dataset. Lastly, we investigated sleep duration by examining the electrolytes and biomolecules that might be correlated with sleep disorders and sleep duration. We find that calcium, potassium, total cholesterol, and phosphorus have significant effects on sleep duration. Our results suggest that depression severity is strongly associated with the incidence of sleep disorders, and a number of biochemicals in the human body are associated with sleep disorders and sleep duration. Our findings shed light on future research concerning the possible mechanisms underlying the effects of depression and biochemical factors on sleep disorders.

# Introduction

Sleep is a universal behavior among many vertebrates including humans, and it plays an essential role in the physiological functioning of the human body. Despite the large amount of research conducted over the past few decades, sleep remains to be a scientific enigma waiting to be demystified. Rather than merely a state of reduced responsiveness to the surrounding environment, a number of physiological functions of sleep have gradually been proposed and examined. Among the numerous hypotheses that have been proposed, some important functional roles of sleep have been widely accepted by scientists. Current evidence has suggested that during the stages of sleep, our brain is able to perform a number of tasks that help to restore energy and prepare us for the next day, such as the repair of cells in the brain, the clearance of metabolic wastes and toxins, consolidation of memory, and the regulation of the immune system[1,2]. All these tasks are crucial for the normal functioning of the brain and the entire body. Thus, sleep is an indispensable part of our lives that should not be underestimated.

Despite the importance of sleep, the fast pace of modern life and high levels of work stress have caused a large number of people worldwide to experience problems with sleep, such as sleep deficiency and inadequate patterns of sleep. In severe cases, these will eventually lead to sleep disorders that have devastating impacts on one's life. Some common sleep disorders that people experience today include insomnia disorder, narcolepsy, sleep-related breathing disorders, and circadian rhythm sleep-wake disorder[3]. All of these sleep disorders can cause severe impairment in one's physiological and cognitive functioning. Therefore, it is important to understand the possible mechanisms of sleep disorders and which factors might increase the risk of developing them.

In our society today, as sleep problems become more common, depression is also becoming increasingly prevalent. A large proportion of people nowadays are suffering from depression and mood disorders. Depression can severely affect normal biological and behavioral functioning and significantly reduce life quality[3]. Given the fact that comorbidity is common among different types of mental disorders, we are interested in how depression might be related to sleep health. Although previous research has attempted to elucidate the correlation between mental health and sleep, currently there are only a few scientific studies focusing closely on depression and directly relating it to sleep.

In this study, we narrow the focus to depression and explore its relationship to sleep disorders. In our analysis, the depression score of each study participant is determined based on Patient Health Questionnaire-9 (PHQ-9), and sleep disorders were recorded based on self-report of doctor's

diagnosis. We build a multiple logistic regression model to assess how depression is associated with sleep disorders, adjusting for confounders such as gender, age, race, BMI, and income. We also compare the statistical results from the 2017-2020 dataset with the dataset generated in 2013-2014 to examine how the association between depression and sleep disorders might change in two different time periods.

In addition to this analysis, we also explore the biological risk factors that might be correlated with sleep, including electrolytes involved in neuronal activity and biochemicals closely related to human health. By leveraging the linear regression model, we evaluate how these biochemical molecules are associated with sleep duration, a determinant of sleep health.

## Method (Outline)

### Data from NHANES

We used the data from NHANES 2017-March 2020 Pre-pandemic and NHANES 2013-2014 (The National Health and Nutrition Examination Survey program from CDC), which assesses the health and nutritional status of adults and children in the United States from 2017 to March 2020 and 2013-2014.

This project has taken a number of health and lifestyle measurements on a specific group of people. Each set of measurements are included in separate files, but all files contain the subject's ID. The ID of each person in this project is unique and constant across all measurements. Therefore, we merged the data in the measurements we are interested in based on these IDs. The health and lifestyle measurements vary year to year. 2013-2014 contains more measurements than that of 2020.

For 2017-2020, we includes Respondent_sequence_number, Usual_sleep_time_weekdays, Usual_wake_time_weekdays, Sleep_hours_weekdays, Usual_sleep_time_weekends, Usual_wake_time_weekends, Sleep_hours_weekends, How_often_snore, How_often_snort_or_stop_breathing, Ever_told_doctor_sleep_trouble, Feel_overly_sleepy_day, Body_Mass_Index, How_often_drink_alcohol, Gender, Race_Hispanic_origin, Age_at_screening, Family_income_to_poverty_ratio, Minutes_sedentary_activity, Depression score, and categorical depression score in our raw data set. Due to the different purpose of each question, we use different subset of the raw data set.

### Primary Question Dataset

To explore the association between sleep trouble and depression, the primary question dataset contains the outcome, primary covariates, potential confounders and potential effect modifiers, including "sleep_trouble", "BMI", "alcohol", "gender", "Race_Hispanic_origin", "age", "Family_income_to_poverty_ratio", "dp_score", and "dp_cate"

Secondary Question Dataset

To explore the association between biochemical molecules and sleep duration, the secondary question dataset contains the outcome (sleep duration), standard biochemistry profile dataset, cholesterol dataset, insulin dataset, Perchlorate, Nitrate & Thiocyanate - Urine dataset, Iodine - Urine dataset, and Cholesterol - Low-Density Lipoproteins (LDL) & Triglycerides dataset.

Analysis of association between sleep disorder and depression

We developed a multiple logistic regression model to assess the association between sleep trouble and depression. Starting from the simplest model with only depression score, we firstly tested modeling depression score as continuous and categorical. Then we assess the potential confounders and effect modifiers. To catch the possible nonlinearity, the quadratic terms are involved in the model. After deciding the final model, we run model diagnostics to detect the potential outliers, high leverage points, and high influential points. Then we use the GOF test (Hosmer-Lemeshow test) and ROC to assess the fitness of the model.

Analysis of association between biochemical molecules

We developed a multiple linear regression model to assess the association between various biochemical molecules and sleep duration. In this case, the outcome variable (sleep duration) is a continuous variable, so we used a linear model for this analysis. We performed an elastic net to filter the covariates that are associated with sleep duration. After variable selection, we conducted model diagnostics to evaluate the residuals and variance. We graphed the residual plot and QQ plot to assess the assumptions of the linear model.

Reference


1. Bear, Mark F., et al. *Neuroscience: Exploring the Brain*. Jones & Bartlett Learning, 2020.
2. Zielinski MR, McKenna JT, McCarley RW. Functions and Mechanisms of Sleep. AIMS Neurosci. 2016;3(1):67-104. doi: 10.3934/Neuroscience.2016.1.67. Epub 2016 Apr 21. PMID: 28413828; PMCID: PMC5390528.
3. Nolen-Hoeksema, Susan. *Abnormal Psychology*. McGrawHill Education, 2020.

**7. Intentions:**
This group is dying, we just want to finish this assignment, and hopefully grab some ***FREE*** food during upcoming presentations.