Contributions: Ariane Lin, Jing Jin, Kin Fung Ng(Rex)

Introduction:

Sleep is critical for mental and physical health, with poor sleep quality often linked to mental health disorders like depression. Given the rising awareness of mental health issues, it's important to study how sleep duration influences the likelihood of depression. This research investigates the relationship between the average sleep hour and chances of depression, focusing on demographic and health-related factors, including age, gender, mentally unhealthy days (DaysMentHlthBad), and sleep problems. Understanding these connections will help inform target interventions for improving mental health outcomes by addressing sleep-related issues. Prior studies supported the relevance of this research. "*Personal Sleep Debt and Daytime Sleepiness Mediate the Relationship Between Sleep and Mental Health Outcomes In Young Adults*" highlights how short sleep duration and personal sleep debt increase depression risks. "*Depression and Sleep*" identifies insomnia as both a symptom and a risk factor, with demographic factors like age and gender influencing sleep disturbances. "*Sleep, Insomnia, and Depression,*" discusses chronic insomnia as a transdiagnostic symptom linked to depression. These studies provide a robust framework for examining the complex interplay between sleep duration and depression.

This research uses linear regression to model the relationship between sleep duration and multiple predictors. This approach is particularly effective for analyzing how continuous (e.g., age) and categorical variables (e.g., gender) influence depression. By quantifying the strength and direction, linear regression is useful for understanding how various elements interact. Moreover, it simultaneously accounts for multiple variables, minimizing potential bias and ensuring a comprehensive evaluation of the interplay between sleep and mental health.

Methods:

This study investigates how demographic and health-related factors influence sleep duration using data from NHANES. The response variable is average hours of sleep per night (SleepHrsNight), and predictors include Age, Gender, mental health days (DaysMentHlthBad), sleep trouble, and depression status.

The analysis followed a clear and organized process. The first step was data cleaning. Missing values were removed to ensure a consistent dataset. Categorical variables, such as Gender, SleepTrouble, and Depressed, were converted into factors to allow proper analysis. After cleaning, the data was split into a training set and a testing set to allow for both model building and validation. Next, exploratory data analysis was carried out to understand the dataset. Summary statistics were calculated to describe the main features of the data. Scatterplots were used for continuous predictors like Age and DaysMentHlthBad, while boxplots were used for categorical variables like Gender and SleepTrouble. These visualizations helped identify patterns or differences among groups. Outliers were noted but kept in the dataset to avoid bias. A multiple linear regression model was used to analyze how predictors affected sleep duration. The model's assumptions were carefully tested to ensure valid results. Residuals were checked for normality using Q-Q plots and histograms. Linearity and equal variance (homoscedasticity) were assessed with residual vs. fitted value plots. Multicollinearity between predictors was measured using Variance Inflation Factors (VIF) to confirm that the predictors were not overly correlated. To improve the model, backward selection was applied. Predictors with p-values greater than 0.05 or confidence intervals overlapping zero were removed step by step. Partial F-tests were used to confirm that removing these predictors did not significantly weaken the model. Model fit and complexity was assessed using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), with lower values indicating better models. The model's performance was evaluated using cross-validation. Leave-one-out cross-validation (LOOCV) was applied to the training set, and the Mean Squared Error (MSE) was calculated to measure how well the model could predict new data. This ensured the model was reliable and generalizable. The final model was tested again to confirm that all assumptions were met. If problems such as non-linearity or unequal variance had been detected, transformations would have been applied. Any issues that could not be resolved were noted as limitations.

This process ensured a thorough and reliable analysis of how demographic and health-related factors influence sleep duration. By carefully preparing the data, testing assumptions, refining the model, and validating its performance, this study provides a solid framework for understanding the predictors of sleep duration.

Results:

The following results explore the relationship between predictors including Age, gender, days of mental health issues, sleeping trouble and depression seriousness and the response variable: Hours of sleep per night. First, we start with the analysis of data to understand the trend between gender and sleep duration, in which we found that females reported slightly longer average sleep hours compared to males in general by the boxplot, suggesting gender has an impact on sleep duration. Next, we created the initial model 1 using all the available predictors mentioned above.

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	signif. codes
Intercept	7.207182	0.058463	123.277	< 2e-16	***
Age	0.001764	0.001058	1.668	0.0954	.
Gendermale	-0.262140	0.036897	-7.105	1.38e-12	***
DaysMentHlthBad	-0.019604	0.002851	-6.875	6.95e-12	***
SleepTroubleYes	-0.545839	0.043391	-12.579	< 2e-16	***
DepressedSeveral	-0.007415	0.054739	-0.135	0.8923	
DepressedMost	-0.211289	0.090273	-2.341	0.0193	*

Table 1 This table presents the estimated coefficients, standard errors, t-values, p-values, and significance codes for the multiple linear regression model predicting sleep duration. Significant predictors include Gender (male), DaysMentHlthBad, and Sleep Trouble (Yes), with all showing highly significant p-values ($p < 0.001$). The model intercept represents the baseline sleep duration for a reference individual with average predictors.

The summary of Model 1 revealed that all predictors were significant except for age and depressed (several) determined by their p-values of less than 0.05. From the table, a -0.262 estimate from gender can be explained by being male is linked to a reduction of 0.262 hours in nightly sleep while participants with reported sleep trouble experienced a reduction of 0.546 hours in nightly sleep explained by the estimate of -0.546. Additionally, Mentally unhealthy days also contribute to a reduction of 0.02 hours of sleeping duration. Serious Depression symptom was associated with a reduction of 0.211 hours of sleep. However, Age and several Depression were not statistically significant in this model as their p-values of 0.0907 and 0.7683 are larger than 0.05. Also the adjusted R-squared of model 1 was 0.0668, showing that 6.68% of the variance can be explained by the predictors in sleep duration. Although the variance is low many factors might influence the outcome so we tried to refine the model to find a better fit model.

In the process of model refinement, backward selection is applied to a new model 2 using the Akaike Information Criterion as known as AIC. The full model 2 gives us an AIC value of 16737.56 while a reduced model excluding non-significant predictors had an AIC of 16818.17. The smaller AIC provided us with a result that the full model 2 provided a better fit for our study data. Then we applied Partial F-tests to investigate the contribution of those non-significant predictors and the results supported that all predictors should be kept in the final model and same as model1. Hence the final model equation is

SleepHrsNight = 7.207+ 0.00176(age)-0.262(Gendermale)-0.0196(DaysMentHlthBad)

-0.546(SleepTroubleYes)-0.007(DepressedServeal)-0.211(DepressionMost)

Next, we conducted a diagnostic check to see if any assumption of the linear regression model was violated.

Normality was checked by Q-Q plots (Figure 1) and histograms (Figure) which show a normal distribution with slight deviations.

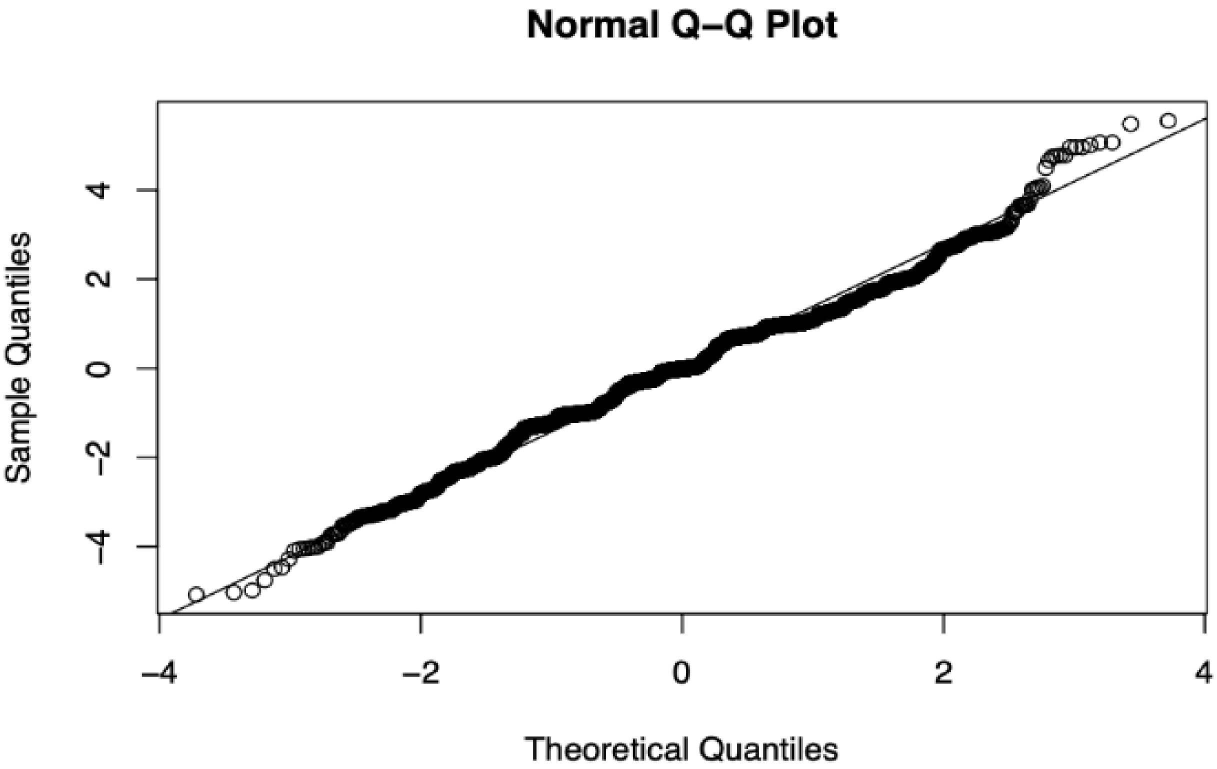

Figure 1. Normal Q-Q Plot for Residuals

The Q-Q plot shows that the residuals from the regression model align closely with the diagonal, indicating that the residuals follow a normal distribution, a key assumption of linear regression.

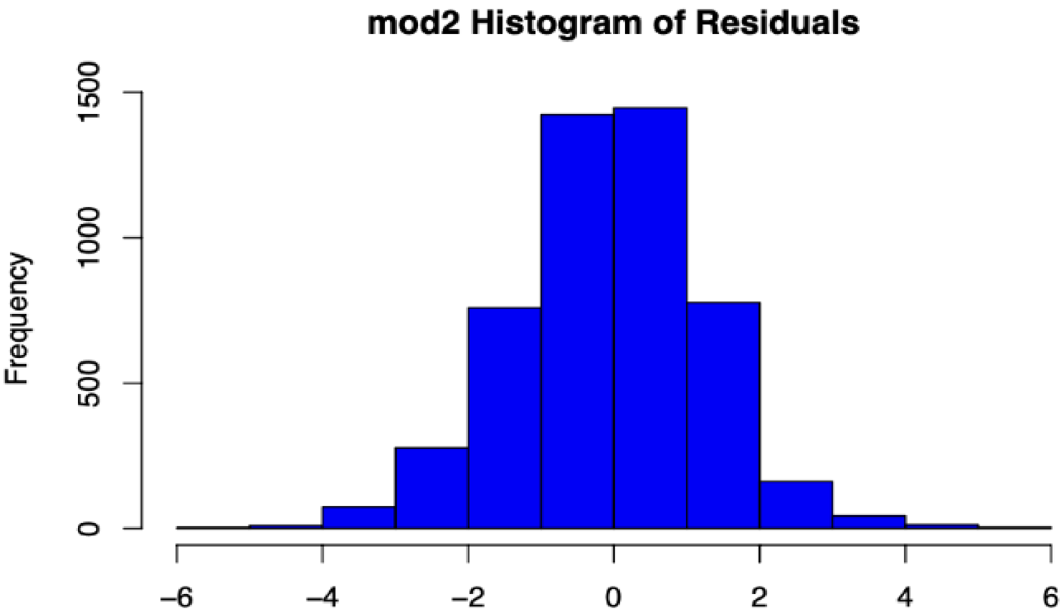

Figure 2. Histogram of Residuals

This histogram displays the distribution of residuals from the regression model. The residuals are approximately symmetric and centered around zero, further supporting the assumption of normality.

Linearity and homoscedasticity were checked by residuals vs. fitted plots (Figure 3), we can see from the plots below that there are no discernible patterns and match with the assumption of constant variance.

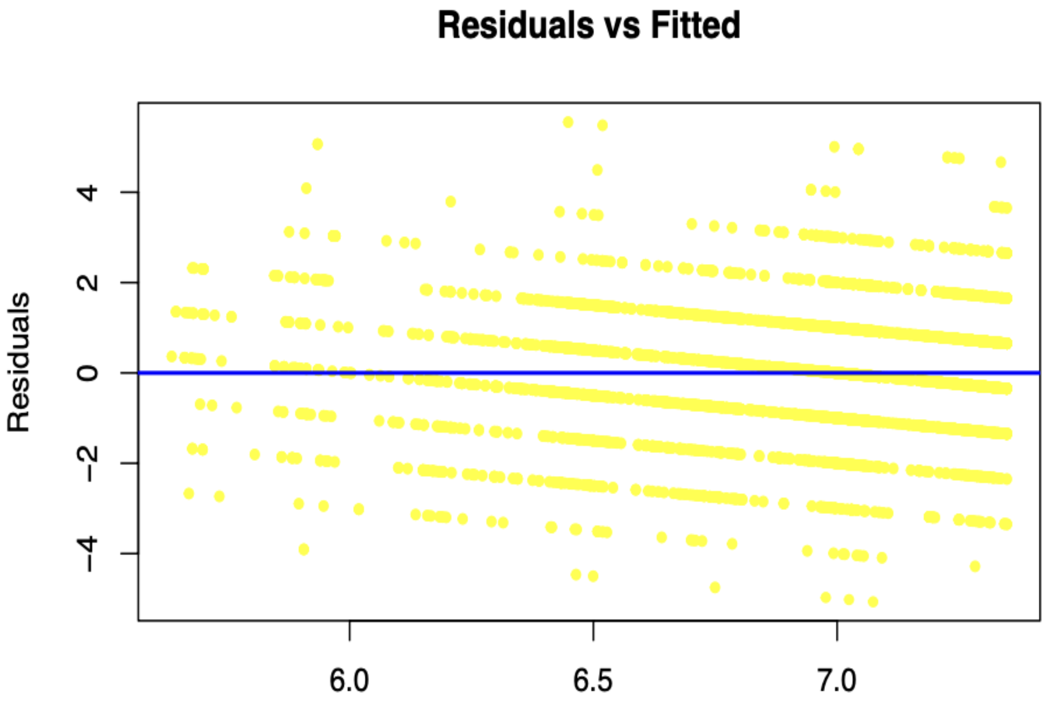

Figure 3. Residuals vs. Fitted Values
The scatterplot illustrates the residuals plotted against the fitted values from the regression model. The lack of discernible patterns suggests that the assumptions of linearity and homoscedasticity are satisfied, indicating that the model fits the data well.

Also, a multicollinearity check is conducted using Variance Inflation Factors (VIF) and pairwise Scatterplots, the VIF values for all predictors were below 2 which rejects multicollinearity. Thus, we have confirmed that the final model fits all the assumptions and is valid.

Last but not least, we have conducted validation through the Leave-One-Out cross-validation (LOOCV), which results in a Mean Squared Error of 1.6766. This number is the average square difference between observed and predicted values and the value of 1.6766 showing the final model was validated. We have conducted an adjusted R-squared of 0.064 for the test dataset, which is close to the value for the training dataset. Hence, they are consistent and support that the final model generalizes well to predict sleep duration from the five predictors.

Conclusion and Limitations:

This study examined how demographic and health-related factors affect sleep duration using a multiple linear regression model. The predictors included Age, Gender, mental health days, sleep trouble, and depression severity. The results showed that Gender and sleep trouble had significant impacts on sleep duration. For example, the coefficient for Gender (male) was -0.262, meaning that males, on average, slept 0.262 fewer hours per night compared to females, holding all other factors constant. Similarly, individuals with sleep trouble slept 0.546 fewer hours per night than those without, on average. These findings are consistent with existing literature, which suggests that males and individuals with sleep difficulties often experience shorter sleep durations. Although the model explained 6.68% of the variance in sleep duration, this result is not surprising. Sleep duration is influenced by numerous factors, including lifestyle, environmental stressors, and physical health, many of which were not included in this analysis. While the findings align with general expectations from prior research, the limited explanatory power highlights the complexity of predicting sleep behaviors.

There are several limitations to this study. First, the model explained only a small proportion of the variance in sleep duration, suggesting that important predictors were not included. Factors such as physical activity, diet, or socioeconomic conditions, which are known to affect sleep, were not available in the dataset. Second, reliance on self-reported sleep duration may introduce reporting biases or inaccuracies. Third, while the assumptions of linear regression were tested and met, minor deviations in residual normality and constant variance could still slightly affect the validity of the results. Additionally, extreme observations (outliers) were retained in the dataset, which might influence the estimates.

Despite these limitations, the study provides valuable insights into how mental health and demographic factors influence sleep. Future research could address these limitations by including a broader set of predictors and employing more precise measures of sleep duration. These improvements could offer a clearer understanding of the factors affecting sleep behavior.

Ethics Discussion:

This study used manual variable selection instead of automated selection methods. The manual selection was chosen because it allows us to use our judgment to carefully evaluate the importance of each variable, ensuring that the model reflects the underlying relationships in the data. This approach is particularly valuable when analyzing complex mental health data, where automated methods might overlook subtle patterns. Although automated methods can save time and provide consistent processes, they often lack the flexibility to account for context-specific factors. For example, automated methods might prioritize statistical significance over practical or theoretical relevance, potentially leading to less meaningful models.

Ethically, manual methods demonstrate a greater sense of responsibility in handling data. They require researchers to think critically about each step in the analysis, ensuring transparency and accountability. This approach also aligns with avoiding negligence by carefully considering all relevant factors before making decisions. While automated methods are not inherently unethical, relying solely on them could result in oversights or misinterpretations. This study prioritized accuracy and relevance by using manual selection, which is essential when dealing with sensitive topics like mental health. This decision reflects a commitment to producing thoughtful and responsible statistical analysis that respects the complexity of the research question.

References:

Centers for Disease Control and Prevention. (2023, May 31). Nhanes - about the National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Dickinson, D. L., Wolkow, A. P., Rajaratnam, S. M. W., & Drummond, S. P. A. (2018). Personal sleep debt and daytime sleepiness mediate the relationship between sleep and mental health outcomes in young adults. Depression and Anxiety, 35(8), 775–783. https://doi.org/10.1002/da.22769

Riemann, D., Krone, L. B., Wulff, K., & others. (2020). Sleep, insomnia, and depression. Neuropsychopharmacology, 45(1), 74–89. https://doi.org/10.1038/s41386-019-0411-y

Steiger, A., & Pawlowski, M. (2019). Depression and Sleep. International Journal of Molecular Sciences, 20(3), 607--. https://doi.org/10.3390/ijms20030607