Lab 2: Describing Generations' Sleep

Neeha Kotte, Tazrian Ahmed, Deva Empranthiri

Introduction:
Sleep plays a vital role in maintaining healthy physical and mental well-being. According to the School of Public Health at the University of Michigan, "sleep is essential to every process in the body, affecting our physical and mental functioning the next day, our ability to fight disease and develop immunity, and our metabolism and chronic disease risk" [1] Our research focuses on exploring the relationship between age and sleep duration, while also considering additional variables such as sex and general health status. Therefore, our overall research question is: how does age correlate with an individual's sleep duration, and what role do sex and health status play in this association?

Data Understanding:
Our dataset, 'sleep75', is sourced from a textbook entitled: *Introductory Econometrics: A Modern Approach, 5th Edition* which originally received the data from an article published in 1990, "Sleep and the Allocation of Time", in the *Journal of Political Economy*. The article explains: "The 1975-76 Time Use Study obtained data from four days of time diaries kept by members of 1519 households. The days were at three-month intervals, with two being weekdays, one a Saturday and the fourth a Sunday. The data on time use are combined into"synthetic weeks"... of the 1519 respondents, 421 were excluded ... [and] others were disqualified if data on one of the control variables of interest was missing, or if there were severe inconsistencies in the data. This left a usable sample of 706 individuals..." [2] Therefore, this dataset includes 706 observations and 34 different variables and more specifically, contains *cross-sectional individual data on sleep patterns* [3] Variables in this dataset include demographic characteristics such as sex, race, employment, health status, marriage status, etc. and important individual behaviors, such as sleep duration and nap duration. Despite the article being published 1990, regression analysis is still relevant, because it provides valuable and historical insights into the relationships and patterns that can help inform current understanding of sleep duration.

How Key Concepts are Operationalized:
To represent our most important X feature, we will use the variable 'age' from the 'sleep75' dataset. In this dataset, 'age' is a metric variable ranging from 23 years to 65 years old. When interpreting our linear models below, we will have to account for our baseline starting from 23 years old, rather than 0 years old. To measure the quantity of sleep, or our Y (i.e. dependent variable), we will use the metric variable 'slpnaps' in the dataset, which represents the minutes of sleep at night (and during naps) an individual receives per week. We note here that this variable only measures the *quantity* of sleep, and not necessarily the *quality* of sleep. While there is another variable called 'sleep' (which simply measures the minutes of sleep at night per week an individual receives), we opted to proceed with 'slpnaps' as we thought this variable provided a more holistic picture of individuals' sleep durations. As mentioned above, we are also interested in concepts such as sex and health. We use the following indicator variables to operationalize these concepts: 'male' (1 for males), 'gdhlth' (1 if in good or excellent health). We acknowledge that the 'male' variable is not gender inclusive, as it relies on a binary classification of male or female - if the dataset had included additional gender categories, we may have been able to draw more nuanced conclusions. Similarly, if the 'gdhlth' variable included specific health conditions (whether or not an individual is on medication, heart conditions, etc.) we may have also been able to understand specific medical root causes, rather than only the general health status.

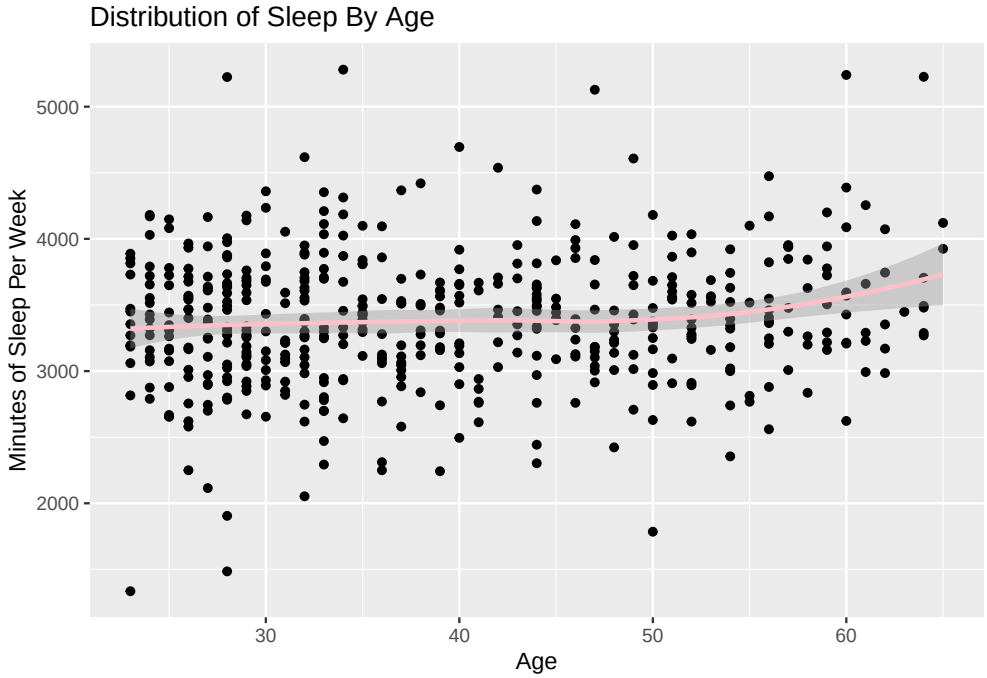
[1] University of Michigan School of Public Health. "Sleep 101: Why Sleep Is So Important to Your Health." (2020).
[2] Journal of Political Economy. "Sleep and the Allocation of Time." (1990).
[3] Research Gate. "Sleep75." (2000).

Data Wrangling:
During the data wrangling process, the analysis revealed distinct characteristics within the confirmation dataset. 'Age' exhibited a robust distribution, ranging from 23 to 65, without any discernible outliers. Similarly, the variable representing sleep duration ('slpnaps') demonstrated a consistent distribution, spanning from 1335 to 5280, displaying no outliers and following a normal distribution pattern. Binary features such as gender displayed binary values (0 or 1) with no missing data. The mean value for gender was calculated as 0.5556, indicating a balanced distribution. Regarding general health status, the binary feature "good health" also showed binary values (0 or 1) with no missing data. The mean value for good health status was calculated as 0.897, indicating a prevalence of individuals reporting good health within the dataset. Despite attempts to enhance the linear regression model's performance through transformations such as logarithmic or squared adjustments to age and sleep values, no significant improvements were observed.

Model Specification:
Because of our interest in observing the associations between health status, gender/sex and age on sleep, we opted to include covariates to partial out effects of age vs. health status vs. gender/sex on sleep (especially as these may also vary based on each other). Therefore, we studied the following models:

1. slpnaps ~ age
2. slpnaps ~ age + gdhlth
3. slpnaps ~ age + gdhlth + male

Model Assumptions/Limitations:
The two main assumptions for OLS regression in a large sample are 1) identically and independently distributed (IID) data and 2) a unique best linear predictor (BLP). The data is likely IID since each respondent is from a separate household. As this was collected for an econometrics study, we assume these households are nationally representative or at least, randomly sampled. Additionally, the distribution of the variable 'sleep' is fairly symmetric, while the distribution of the variable 'age' is slightly right skewed. The other two explanatory variables we employed, 'gdhlth' and 'male', are both indicator variables. In this case, we do not see evidence of any *heavy* tails (although there are slight skews) or multicollinearity, so we can conclude that there is likely a unique BLP.



```
##
## ================================================================================
##                                        Dependent variable:
##                    ----------------------------------------------------------
##                                              slpnaps
##                            (1)                 (2)                  (3)
## --------------------------------------------------------------------------------
## age                     4.722* (1.966)       4.288* (1.961)       4.446* (1.960)
## gdhlth                                     -195.675** (74.042)  -191.446** (73.967)
## male                                                             -73.072 (45.143)
## Constant            3,202.399*** (80.271) 3,394.931*** (108.045) 3,425.526*** (109.511)
## --------------------------------------------------------------------------------
## Observations                495                  495                  495
## R2                         0.012                0.025                0.031
## Adjusted R2                0.010                0.021                0.025
## Residual Std. Error  502.045 (df = 493)   499.026 (df = 492)   498.206 (df = 491)
## F Statistic         5.768* (df = 1; 493) 6.411** (df = 2; 492) 5.162** (df = 3; 491)
## ================================================================================
## Note:                                              *p<0.05; **p<0.01; ***p<0.001
```

Model Results & Interpretation: An exploration dataset was used to gain understanding of the data and variables. Additionally, a cross-validation procedure was performed to ensure consistency in the significance of predictors. By removing the seed and generating multiple random samples, consistent significant results were found across iterations, reinforcing the reliability of the models' coefficients. The analysis progressed by sequentially building and comparing three linear regression models ('model_one', 'model_two', and 'model_three') to understand the factors influencing variations in 'slpnaps'. Initially, 'model_one' was constructed with only the 'age' variable as a predictor which showed that for every one year increase in age, there was an expected increase of approximately 4.722 minutes in 'slpnaps', which is not very practically significant. The intercept for 'model_one' is 3202.399 mins in 'slpnaps'; a 23 year old individual, however, on average, receives about 3311.005 mins or about 55 hours of sleep. However, the model's explanatory power was limited, as indicated by a low adjusted R-squared of 0.00956. Expanding upon this, Model 2 incorporated both age and good health as predictors. The interpretation showed a surprising result: individuals with good or excellent health tend to have lower minutes of 'slpnaps', controlling for age, with a coefficient of -195.675 (i.e. about 3 hours less sleep per week, a practically significant result). This model exhibited a slightly improved adjusted R-squared of 0.02144. The comparison between 'model_one' and 'model_two' through an ANOVA test demonstrated a statistically significant improvement in model fit, suggesting that the inclusion of 'gdhlth' enhanced the model's ability to explain variations in 'slpnaps' beyond age alone. Model 3 included gender ('male') as an additional predictor, but the coefficient for gender was not statistically significant, suggesting it may not be a strong predictor of 'slpnaps'. In addition, the ANOVA test comparing 'model_two' and 'model_three' indicated no significant difference in model fit, implying that the addition of the 'male' variable did not contribute meaningfully to explain variations in 'slpnaps' beyond the variables included in 'model_two'. Nevertheless, further exploration or refinement may be necessary to capture additional variance in sleep duration and improve accuracy. In conclusion, 'model_two', which includes 'age' and 'gdhlth', provides the most concise explanation for the variability in 'slpnaps'. While 'gdhlth' and 'age' were found to be statistically significant predictors in 'model_three', the inclusion of the 'male' variable did not yield a significant improvement in model fit. Therefore, for predicting 'slpnaps', 'model_two' is recommended as it strikes a balance between model complexity and explanatory power.

Conclusion:
Our findings suggest that age and good health are more robust predictors of sleep duration ('slpnaps') compared to gender. With a more robust dataset, further research could explore additional predictors or interactions to better capture the complexity of sleep and enhance our understanding of demographic, health and behavioral factors which are associated with it.

Appendix: Please see github repo here: https://github.com/mids-w203/lab-2-lab2_chalcedony

Per feedback in the live session, we also explored adding the variable 'totwrk' (mins worked per week) in another model which did significantly increase our r-squared value, as it accounted for more of the variation in sleep duration. However, the significance of our age coefficient dropped - its significance was likely absorbed by the highly significant 'totwrk' variable.

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                slpnaps
## ------------------------------------------------
## age                          3.688 (1.884)
## gdhlth                     -161.283* (71.117)
## male                         50.931 (47.230)
## totwrk                      -0.166*** (0.025)
## Constant                 3,711.643*** (113.703)
## ------------------------------------------------
## Observations                    495
## R2                             0.109
## Adjusted R2                    0.102
## Residual Std. Error       478.014 (df = 490)
## F Statistic             15.045*** (df = 4; 490)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```