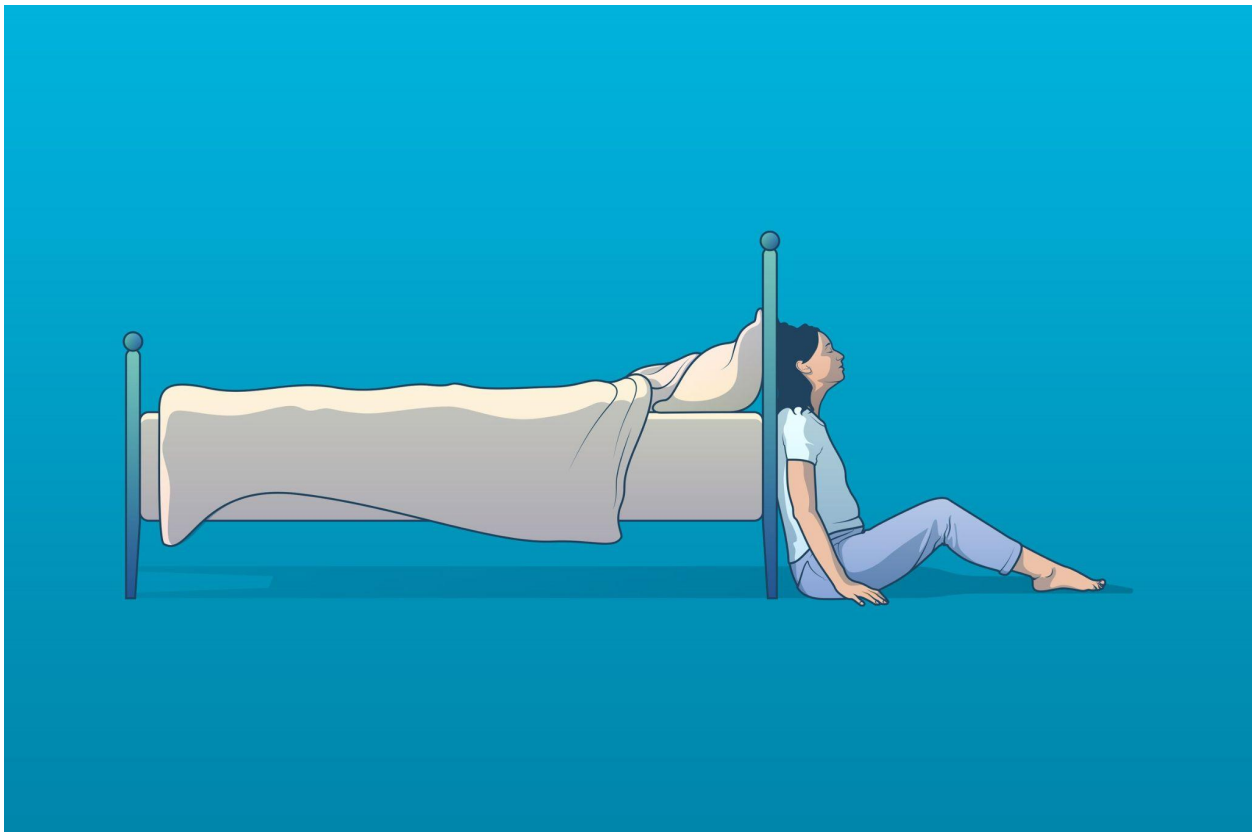
# Exploring the Relationship between Age, Gender, and Other Factors on Sleep Efficiency in Individuals

Bokai Lai

Matthew Ma

Kate Zhang



# Introduction

In the busy and fast-paced society nowadays, sleep quality/efficiency has become a vital issue that people are concerned about. Poor sleep efficiency can cause feelings of daytime sleepiness. More importantly, it can contribute to sleep debt. Missing out on sleep while lying in bed can contribute to rest debt, like staying up too late or waking up too early. In this case study, we use the ***Sleep efficiency dataset*** from the Kaggle website (https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency). We analyze the ***linear relationship*** between several external factors and sleep efficiency by building and comparing different linear models and choosing the best-fitted one. More specifically, our external factors are Age (years), Gender (male/female), sleep duration (hours), sleep efficiency(percentage of time in bed spent asleep), and also smoking status (yes/no).

## Research Question

How age, gender, and other external factors are related to sleep efficiency, and what is the best linear model predicting sleep efficiency based on these factors?

# Analysis

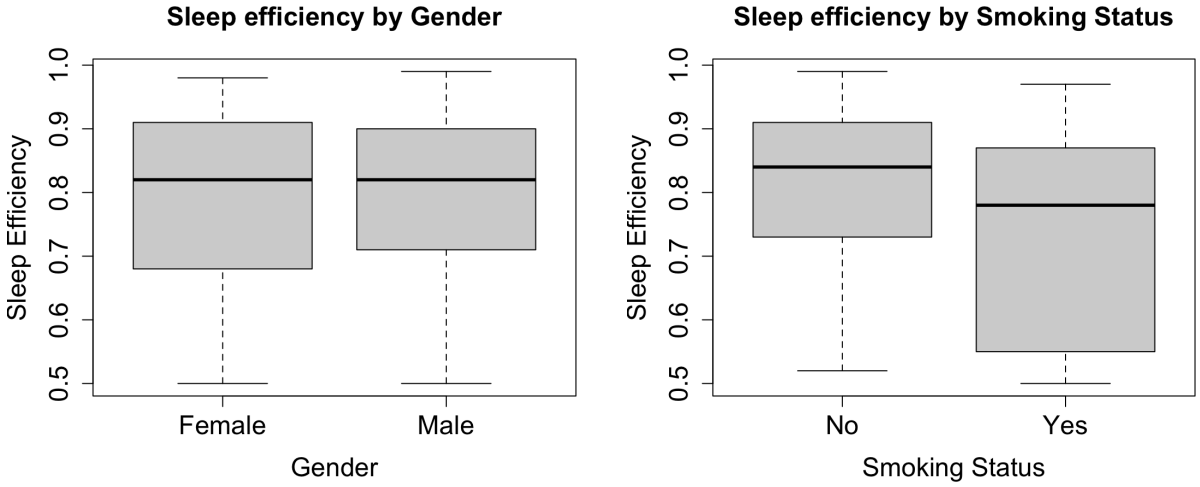
We start by tidying up our data set and using the tidy.csv. To tidy the dataset, we converted the following columns into categorical variables: gender (male/female) and smoking status (smoker/non-smoker). We also removed the light sleep percentage column because it perfectly correlates with deep sleep percentage. Someone who is asleep can only be in light or deep sleep. We also removed the bedtime and wake-up time columns because their datatype is challenging to work with. Finally, we removed rows which have a N/A value. Then to assess which variable has a significant role in affecting sleep efficiency, we first construct a full model - by using ***Sleep.efficiency*** as the response variable and all other terms as explanatory variables.

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.170669 -0.040915  0.007032  0.040159  0.146387

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.3489327  0.0413858   8.431 7.35e-16 ***
Age                     0.0009515  0.0002427   3.920 0.000105 ***
GenderMale              0.0014379  0.0069493   0.207 0.836184
Sleep.duration          0.0017482  0.0035196   0.497 0.619683
REM.sleep.percentage    0.0066733  0.0009386   7.110 5.86e-12 ***
Deep.sleep.percentage   0.0055671  0.0002377  23.425  < 2e-16 ***
Awakenings             -0.0318783  0.0025249 -12.626  < 2e-16 ***
Caffeine.consumption    0.0002412  0.0001131   2.132 0.033674 *
Alcohol.consumption    -0.0061303  0.0021174  -2.895 0.004009 **
Smoking.statusYes      -0.0460092  0.0067858  -6.780 4.65e-11 ***
Exercise.frequency      0.0063932  0.0022973   2.783 0.005658 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06071 on 377 degrees of freedom
Multiple R-squared:  0.8051,    Adjusted R-squared:  0.7999
F-statistic: 155.7 on 10 and 377 DF,  p-value: < 2.2e-16
```

From the summary of the full model, we found P-values for ***Gender*** and ***Sleep. duration*** is huge, so we might want to remove these two terms from our model to make it fit better. However, we should also use other visualizations to decide whether to remove these two terms. All other explanatory variables have negligible enough P-values. The adj $R^2$ is 0.7999, which is pretty nice but can be better.

---

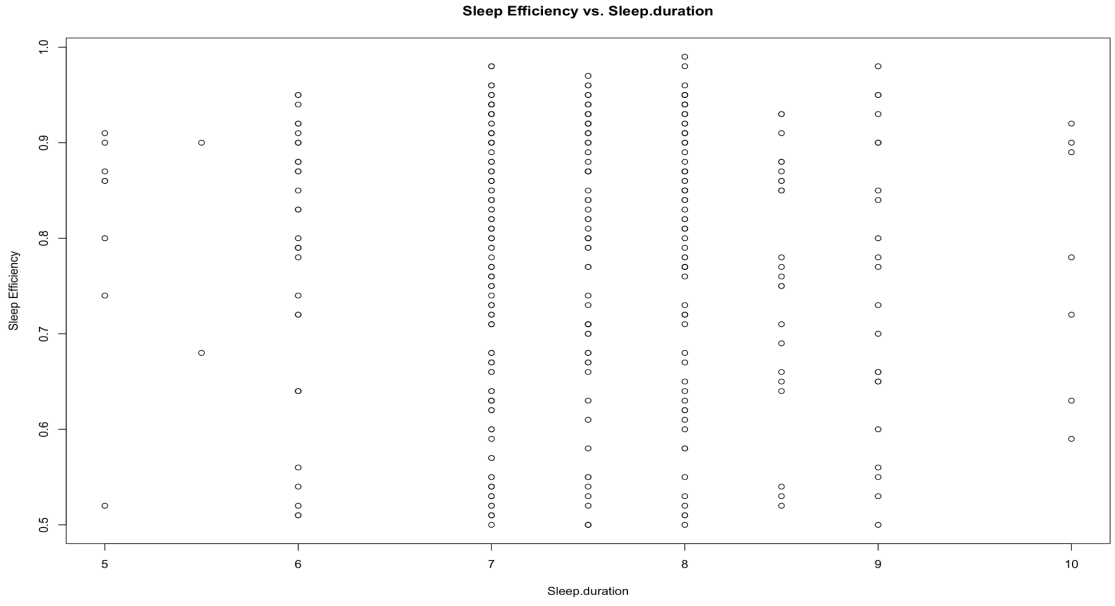
- Plot side-by-side boxplots with the two categorical variables (i.e. ***Gender*** and ***Smoking.status***):




The side-by-side boxplot on the left shows that the female has a broader distribution, whereas the male has a narrower range, but the median is quite close. The boxplot for the Female is more right-skewed, while the Male one is more symmetric. This finding indicates that ***Gender*** might not be a significant explanatory variable that affects ***sleep.efficiency***. Since their median and spreads are roughly the same, We observe that non-smokers have a higher median than smokers on the right. The latter has a wider distribution of sleep efficiency, and the boxplot is skew to the left. In contrast, non-smokers have a more symmetric boxplot. These differences might indicate that ***smoking status*** is a significant explanatory variable which affects sleep efficiency.

---

- Scatterplot between ***Sleep.efficiency*** and ***Sleep.duration***

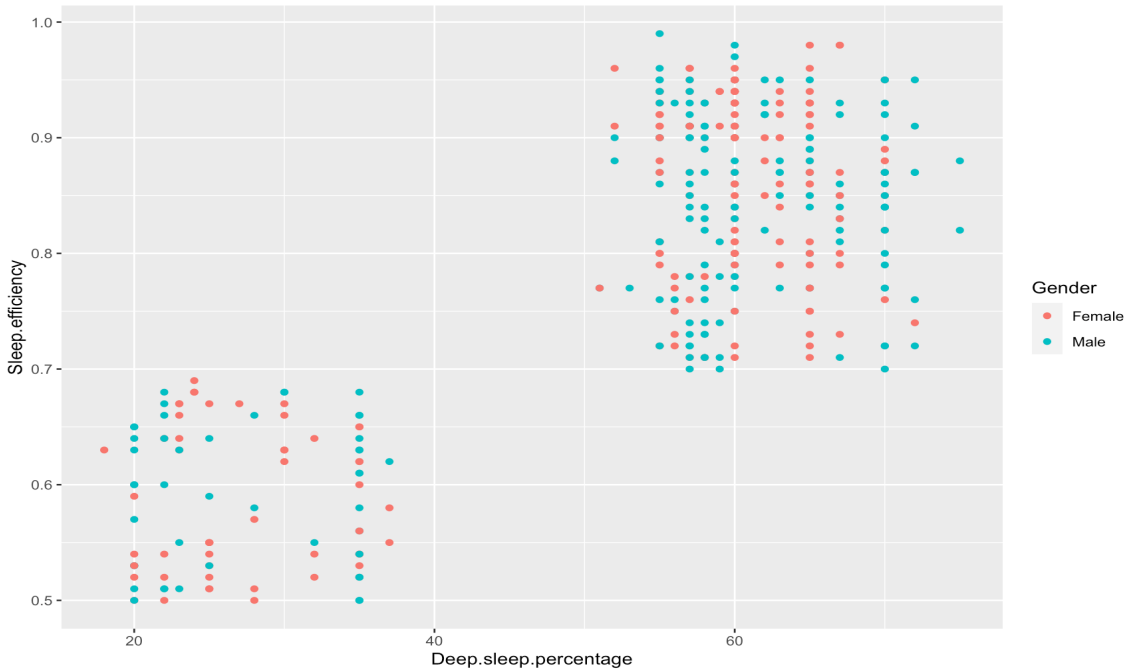


By observing this scatterplot between ***Sleep.efficiency*** and ***Sleep.duration***, we found that there is no apparent linear relationship between them. To ensure that, we computed the ***correlation*** coefficient between Sleep.efficiency and Sleep.duration, which is **-0.0192,** which is small. Therefore, we decided to remove Sleep.duration from our linear model and see if the adj $R^2$ would increase.

After removing the Sleep.duration from our linear model, we found that the ***adj $R^2$ increased*** to **0.8003**. The adj $R^2$ did NOT change much, so it is reasonable to remove Sleep.duration.

---

- Scatterplot between ***Sleep.efficiency*** and ***Deep.sleep.percentage*** using ***Gender*** to group and colour




This scatterplot shows a strong positive relationship between ***Sleep.efficiency*** and ***Deep.sleep.percentage***. Therefore, we can group data points by ***Gender*** and colour Male and Female by different colours to see if ***Gender*** would affect ***Sleep.efficiency.*** However, by doing that, the graph of Male and Female data points is randomly distributed, which does not show any relationship between ***Gender*** and

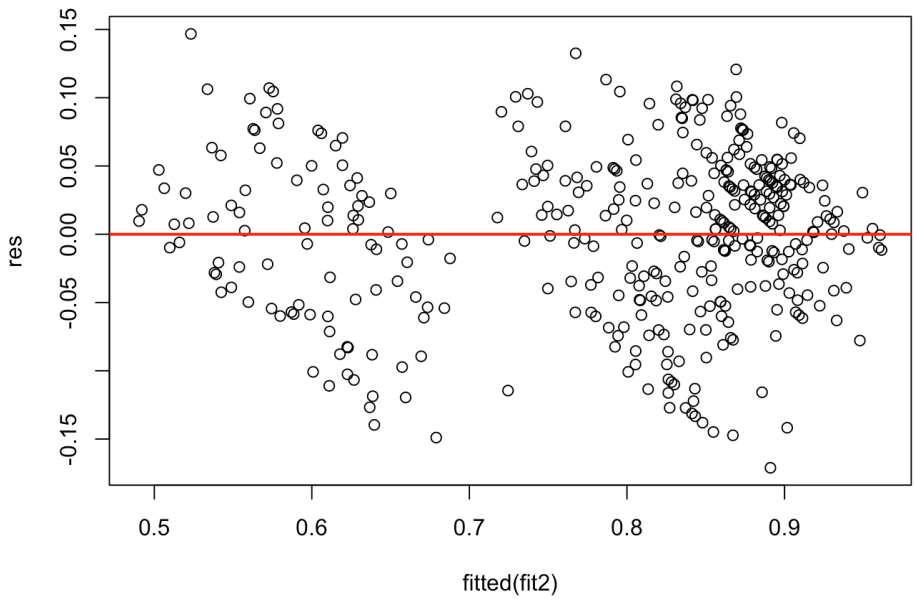
***Sleep.efficiency***. Therefore, we should test if removing ***Gender*** from our linear model would make it fit better.

By removing ***Gender*** from our model, we observed that the ***adj $R^2$*** increased to 0.8008, a little higher than the previous model. The ***adj $R^2$*** did NOT change much, which indicates that ***Gender*** does not have much effect on the linear model, and it is reasonable to remove it.

---

- Residual plot

To determine whether the recently changed model (call it *fit2*) is appropriate to Predict our ***Sleep.efficiency***, we create a residual plot to visualize the variation of residuals.




From the residual plot above, we observe roughly randomly distributed residuals, illustrating that the variation of residuals is small. Therefore, the *fit2* model is suitable for predicting ***Sleep.efficiency*** for now.

---

- Correlation matrix

To further improve our model and determine if there is a need to add interaction terms, we create the correlation matrix for every model-included variable to see the correlation between each other.

| | Age | Efficiency | Duration | REM | Deep | Awakening | Caffeine | Alcohol | Exercise |
|---|---|---|---|---|---|---|---|---|---|
| Age | 1 | 0.12 | -0.07 | 0.02 | 0.06 | 0 | -0.17 | 0.07 | 0.07 |
| Efficiency | 0.12 | 1 | -0.02 | 0.06 | 0.79 | -0.57 | 0.07 | -0.4 | 0.27 |
| Duration | -0.07 | -0.02 | 1 | -0.02 | -0.04 | -0.01 | -0.03 | -0.05 | -0.05 |
| REM | 0.02 | 0.06 | -0.02 | 1 | -0.19 | -0.02 | 0.11 | -0.04 | 0.04 |
| Deep | 0.06 | 0.79 | -0.04 | -0.19 | 1 | -0.33 | -0.02 | -0.37 | 0.17 |
| Awakening | 0 | -0.57 | -0.01 | -0.02 | -0.33 | 1 | -0.11 | 0.21 | -0.23 |
| Caffeine | -0.17 | 0.07 | -0.03 | 0.11 | -0.02 | -0.11 | 1 | -0.1 | -0.08 |
| Alcohol | 0.07 | -0.4 | -0.05 | -0.04 | -0.37 | 0.21 | -0.1 | 1 | 0 |
| Exercise | 0.07 | 0.27 | -0.05 | 0.04 | 0.17 | -0.23 | -0.08 | 0 | 1 |

Based on the correlation matrix, there do not seem to be any explanatory variables that are highly correlated with each other. Therefore, we decided NOT to include any interaction terms.

---

- Forward selection

Below is the result of the forward selection using ***Sleep.efficiency*** as y and all other explanatory variables in the full model as x. This is for double-checking our model.

| | (Intercept) | Age | GenderMale | Sleep.duration | REM.sleep.percentage | Deep.sleep.percentage | Awakenings | Caffeine.consumption | Alcohol.consumption | Smoking.statusYes | Exercise.frequency |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE |
| 2 | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE |
| 3 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE |
| 4 | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| 5 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | FALSE | TRUE | FALSE |
| 6 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE | FALSE |
| 7 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | TRUE | TRUE | TRUE |
| 8 | TRUE | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE |

From this result, we can observe that if we choose eight variables as explanatory variables in our model (since we have eight explanatory variables in our *fit2* model), we should choose ***Age, and REM.sleep.percentage, Deep.sleep.percentage, Awakenings, and Caffeine.consumption, Alcohol.consumption, Smoking.status, and Excercise.frequency***. This choice of forward selection is the same as our *fit2,* which indicates that our model is good and ready for making predictions.

- Train/test cross-validation

To evaluate our fit2, we used cross-validation with training and a test set. To create the training set, we randomly selected 80% of the rows to fit our model. From the training step, we got the following model:

Efficiency = 0.3634105 + 0.0009562 * Age + 0.0066438 * REM.sleep.percentage + 0.0055625 * Deep.sleep.percentage - 0.0318493 * Awakenings + 0.0002333 * Caffeine.consumption - 0.0062024 * Alcohol.consumption - 0.0457656 * Smoking.status.yes + 0.0064591 * Exercise.frequency.

Based on this model, we predicted sleep efficiency based on the remaining 20% of rows. We calculated the root mean square error (RMSE) for the predictions using this formula:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{x_i})^2}$$

Where N is the total number of observations, *i* is the index of a given observation, is the value of the ith observation, and is the fitted value of the ith observation.

The RMSE is 0.07001605. In the dataset, sleep efficiency measures the proportion of time spent asleep in bed. This means that, on average, the model's prediction of a person's sleep efficiency differed from the person's actual sleep efficiency by 0.07001605, or around 7%.

# Conclusion

To conclude, we found that the best linear model to predict sleep efficiency is *fit2*. Based on all analyses, we observe that **Age**, **REM sleep percentage**, **Deep sleep percentage**, **awakening**, **Caffeine consumption**, **Alcohol consumption**, **Smoking status**, and **Exercise frequency** are significant factors affecting Sleep efficiency. We can even predict Sleep efficiency with precision based on the linear model fitted by

these factors. More specifically, increasing Age, REM sleep percentage, Deep sleep percentage, and Exercise frequency would increase Sleep efficiency. Decreasing Alcohol consumption and Smoking behaviour would result in a decrease in Sleep efficiency.

To demonstrate our prediction and measure the accuracy of our model, we used the 76th row of our tidy.csv data set as our demo source. A 24-year-old non-smoker with 26% of REM sleep and 56% of Deep sleep consumes 20 mg of caffeine but no alcohol 24 hours before bedtime. She wakes up four times during sleep and exercises once weekly with a sleep efficiency of 0.77. By our model, we predict this person will have a sleep efficiency of 0.755 which has an error of **1.88%**, which is acceptable in accuracy.

This study obtained critical findings on how sleep efficiency is affected by other common factors in life. Sleep quality is a common problem nowadays, especially for those university students who must study at night. However, based on the findings of this study, we suggest that people who have issues getting quality sleep to sleep early, drink less, stop smoking, and exercise more. Everyone deserves a nice sleep.