Investigating Sleep Efficiency with Linear Regression

Ali Afkhami (30271805), Daniela Mañozca Cruz (30262558), Evan Losier (30022571),
Luisa Alejandra Sierra Guerra (30261956), Ruby Nouri Kermani (30261323)

1. INTRODUCTION

1.1. MOTIVATION

1.1.1. Context

Sleep is a fundamental biological need for humans, as highlighted in Maslow's hierarchy of needs. However, sleep alone is not sufficient; it must be efficient. Sleep efficiency refers to the proportion of time spent actually sleeping while in bed. Understanding sleep efficiency is critical because poor sleep quality can negatively impact individuals across multiple dimensions, including cognitive, emotional, and physical health. For example:

- Cognitive decline: Poor sleep efficiency has been linked to neuronal cell loss and impaired brain health (Wang & Aton, 2022).
- Mental health: Inefficient sleep is associated with symptoms of depression (Pan et al., 2022).
- Brain health: Short sleep duration increases the risk of reduced cognitive function and brain health (Fjell et al., 2023).
- Gut health: Sleep quality also affects the gut microbiota, which plays a role in overall health (Deng et al., 2024).

These issues not only affect an individual's body, but also their social and professional lives. Therefore, understanding the factors that influence sleep efficiency is essential for improving overall well-being.

1.1.2. Problem

The primary problem this project aims to address is identifying the key factors that influence sleep efficiency and understanding how these factors interact to affect sleep quality. We will investigate which factors provided in the dataset are most significant for predicting sleep efficiency and will visualize the construction of the model along the way.

1.1.3. Challenges

This problem is challenging because everyone's sleep schedule and habits are different, so we might encounter barriers when trying to create a model that accurately predicts sleep efficiency for everyone. For example, none of us have backgrounds in sleep or health, so we had to build our understanding of the various predictors and how they might relate to sleep efficiency.

1.2. OBJECTIVES

1.2.1. Overview

By identifying the factors that influence sleep efficiency, we can provide evidence-based recommendations to help individuals improve their sleep quality and, consequently, their overall quality of life.

1.2.2. Goals & Research Questions

Clearly state each of your data/visual analytics goals and research questions.

- Analyze how specific factors (e.g., bedtime, wake-up time, sleep duration, REM sleep, deep sleep, light sleep, awakenings, caffeine/alcohol consumption, smoking status, and exercise frequency) influence sleep efficiency.
- Provide actionable recommendations for improving sleep quality based on the findings.

2. METHODOLOGY

2.1 Data

The dataset was obtained from Kaggle which is free to use for project purposes. The dataset contains various factors affecting the sleep efficiency listed in different columns. See *Table 1* in the Appendix for a sample of the dataset.

1. Age: The participant's age. Sleep habits and quality might change as a person gets older.
2. Gender: A person's gender. We will investigate if different genders have different patterns in sleep efficiency.

3. Bedtime: The time at which a person goes to bed. It is a key part of the body's natural circadian rhythm.
4. Wake-up time: The time a person wakes up. This can disrupt sleep cycles if it is inconsistent or too early.
5. Sleep duration: The amount of time spent sleeping. Both insufficient and excessive sleep can harm sleep quality.
6. REM sleep: The percentage of total sleep that is REM sleep. It is crucial for cognitive restoration and emotional regulation.
7. Deep sleep: The percentage of total sleep that is deep sleep. This phase is essential for physical recovery and immune function.
8. Light sleep: The percentage of total sleep that is light sleep. While less restorative, light sleep still plays a role in transitioning between sleep stages.
9. Awakenings: The number of times a participant awoke during the night. Frequent awakenings during the night can fragment the sleep stages.
10. Caffeine consumption: Amount of caffeine consumed (mg) 24-hour before bedtime. Caffeine could keep someone awake longer than necessary and disrupt their circadian rhythm.
11. Alcohol consumption: Amount of alcohol consumed (oz) in the 24-hour before bedtime. Alcohol could keep someone awake longer than necessary and disrupt their circadian rhythm.
12. Smoking status: Whether the participant smokes or not. Nicotine is a stimulant that can interfere with falling asleep and staying asleep.
13. Exercise frequency: The number of times a participant exercises in a week. Regular physical activity has been shown to reduce stress and improve sleep quality.
14. Sleep Efficiency: This will be the response variable, is quantitative, and as stated in the project proposal is measured in percentage. Given there are no values collected as 0 or 100, and the values close to 100 represent less than 3% of the records within the dataset , we see it is viable to try to model the sleep efficiency in this project.

2.2 **Approach**

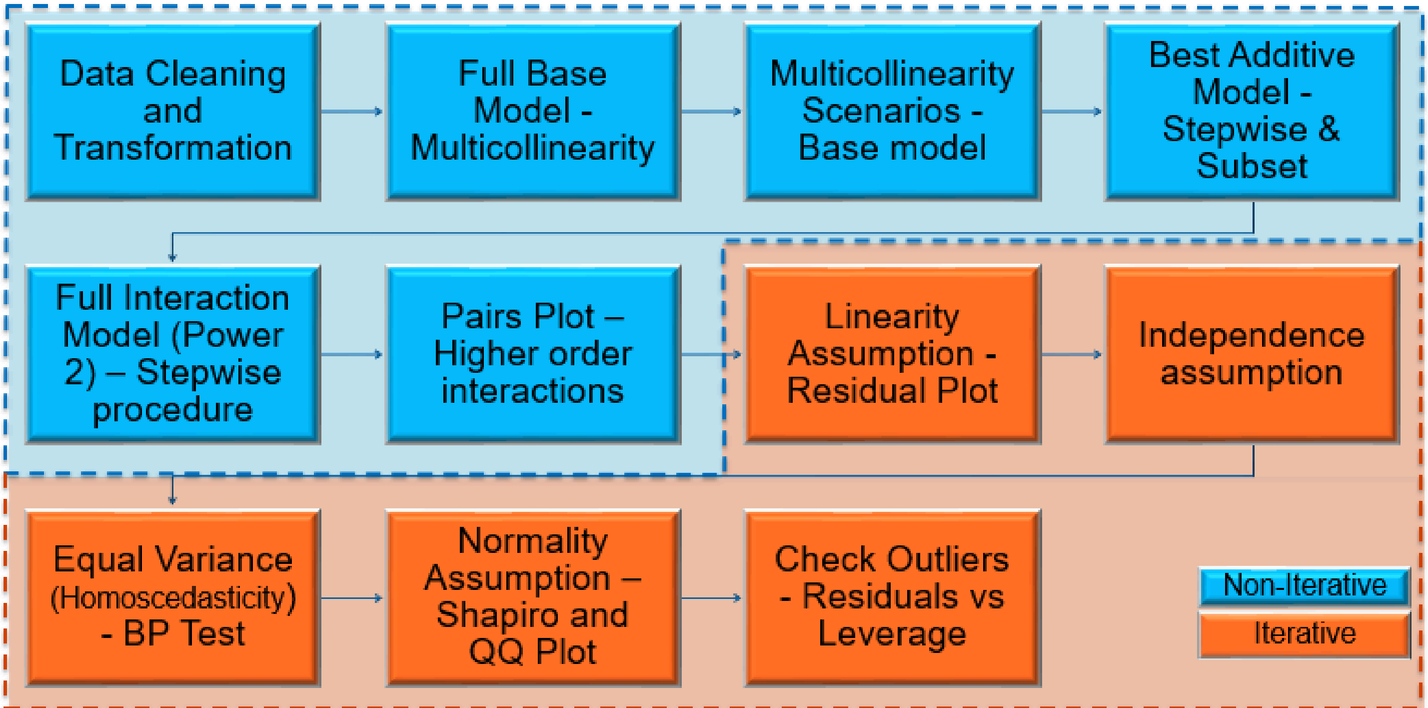
We used R to investigate various assumptions for the multilinear regression model and found the best additive, interaction, and higher order model to better predict the target variable.

2.3 **Workflow**

- Step 1. Data Cleaning and Data Transformation: First, all columns were checked for missing values and they were replaced by zero to avoid all the errors related to null values present in the dataset. Afterwards, the "Bedtime" and "Wakeup time" columns were converted to datetime format and shifted to a linear scale of hours past earliest bedtime/wake-up time to be better suited as inputs for the model.
- Step 2. Initial models were used to verify assumptions such as linearity and multicollinearity. The Variance Inflation Factor (VIF) test was applied to identify multicollinearity. If more than two variables exhibited high VIF values, alternative models were constructed to retain the best most appropriate predictors.
- Step 3. The best additive model was created following the Subset procedure, it was possible given the dataset only had 452 records and there were no computer performance issues.
- Step 4. A two-way interaction model was constructed using a stepwise selection procedure to evaluate interaction effects between variables.
- Step 5. The potential of a higher order model was evaluated through the inspection of a pairs plot and residual plots.
- Step 6. Based on the previous assessments, a higher order model was constructed to capture more complex relationships.
- Step 7. Model assumptions were evaluated : Linearity was assessed via a residuals plot. Independence of errors, was determined to not fully apply in this context given the data is not time series. Equal variance (Homoscedasticity) was assessed through the Breusch Pagan (BP) test. Normality of residuals was evaluated through the Shapiro-Wilk test and Q-Q plots. Finally, outliers were investigated using a residual vs leverage plot.

The process is illustrated in the following schema, which is mainly divided between Non Iterative and Iterative components. Blue sections represent one-time steps, while orange sections represent tasks or steps that were repeated multiple times to optimize the model performance for the dataset analysed in this project.

Figure 1. Methodological Workflow



Source: Author's own work

The test or techniques applied during this project are presented in the following summary table 2:

Table 2. Summary of Analytical Techniques Applied

Test or Technique	Use case	Example code
Fitted vs Residual plot	This plot was mainly used to assess the assumptions of linearity and equal variance. It allows us to visually evaluate whether the residuals are randomly distributed around zero. Any visible pattern may indicate violation of the assumption.	ggplot(full_model, aes(x=.fitted, y=.resid)) + geom_point() + geom_smooth() + geom_hline(yintercept = 0)
VIF Test	Used to detect multicollinearity in the base full model. This test identifies if two or more predictors present redundant information. It was applied multiple times to keep the best predictors in the base model. H_0: Predictor variables have no multicollinearity VIF=0, H_a: Predictor variables have multicollinearity VIF=1.	imcdiag(full_model, method="VIF")

Test or Technique	Use case	Example code
Step -Wise Procedure (Both)	Applied to construct the best additive model based on the adjusted R^2. This method sequentially adds or removes predictors based on the p-values provided.	full_model_rem_deep_wise = ols_step_both_p(full_model_rem_deep, p_enter = 0.05, p_remove = 0.3, details=FALSE) summary(full_model_rem_deep_wise$model)
All Possible-Regression selection procedure.	Used to create the best additive model, based on adjusted R^2 and AIC. The selected model coincided with the one obtained through step wise procedure.	ExecSubsets=ols_step_best_subset(full_model_rem_deep, details=TRUE)
Pairs plot	Used to identify correlations between the response variable and the predictors. This plot guided the decision if higher order terms had to be included in the model.	pairs(~Sleep.efficiency + Awakenings + Alcohol.consumption + factor(Smoking.status) + Exercise.frequency, data=sleep_data)
Two ways Interaction Model	Created to improve adjustment in the model by accounting for interaction effects between pairs of predictors.	full_interaction_model = lm(Sleep.efficiency ~ (Age + REM.sleep.percentage + Deep.sleep.percentage + Awakenings + Alcohol.consumption + Smoking.status + Exercise.frequency)^2, data=sleep_data) summary(full_interaction_model)
Higher order Model	Created to improve adjustment of the model and to address violations of Multilinear regression model assumptions.	model1 = lm(Sleep.efficiency ~ REM.sleep.percentage + Age + Awakenings + Exercise.frequency + Smoking.status + Alcohol.consumption + Deep.sleep.percentage + Smoking.status*Deep.sleep.percentage + Awakenings*Deep.sleep.percentage + Age*Deep.sleep.percentage + REM.sleep.percentage*Awakenings + I(Deep.sleep.percentage^2) + I(Deep.sleep.percentage^3) + I(Awakenings^2) + I(Awakenings^3) + I(Awakenings^4), data=sleep_data) summary(model1)
Breusch-Pagan Test (BP Test)	It was used to identify heteroscedasticity. H_0: Heteroscedasticity is not present, H_a: Heteroscedasticity is present.	bptest(model1)
Shapiro Test	Used to test normality of residuals. This test compares observed residuals to a normal distribution. H_0: The data are normally distributed, H_a: The data are not normally distributed.	shapiro.test(residuals(model1))

Test or Technique	Use case	Example code
Q-Q plot	Used to validate the Normality assumption. In this graph when residuals are normally distributed they will align to the 45° reference line. Deviations such as an S-shape could mean there is skewness or non-normality in the data.	ggplot(sleep_data, aes(sample=model1$residuals)) + stat_qq() + stat_qq_line()+labs(title="Model 1")
Residual vs. Leverage Plot	It was used to identify influential outliers. Identifying them can help to improve performance and ensure the residuals meet the normality assumption.	plot(model1 ,which=5)
Individual t-test	We employed t-tests to evaluate the statistical significance of both individual predictor terms and higher-order polynomial terms in our regression model. H_0: $\beta_i = 0$, H_a: $\beta_i \neq 0$ (i = 1,2,...,p)	summary(model)
Anova Test	This test was used to identify whether the reduced model was a better fit or the full model. H_0: The reduced model is preferred, H_a: The full model is preferred	anova(fulladditivemodel, full_model)

Source: Author's own work

All the tests and different statistical techniques applied were run at a significance level of **0.05**.

2.4 **Contributions**

This group's tasks were divided equitably based on individual strengths, ensuring balanced effort. **Ali** handled critical coding tasks, including data cleaning, preprocessing and assumption validation, along with interpreting results. **Aleja** supported coding (testing, plots, model refinement), contributed significantly to the report (introduction, dataset summary, discussion), and prepared key slides (analysis, results, conclusions). **Ruby** focused on the RMD report, testing interpretation, and finalizing models, while also contributing to the report's result section. **Daniela** assisted in coding by testing multicollinearity to define base predictors, co-developed the best additive model, and helped create the final model. For the report, she explained the response variable and detailed the workflow and statistical techniques. She also prepared presentation slides covering the introduction, background, objectives, and workflow**. Evan** worked on the written portions and editing of the RMD report, tested outlier and linearity assumptions, verified data cleaning, aided in model discovery, and contributed heavily to the report's results and introduction.

In addition to these tasks, all group members collaborated in-person and helped each other with a lot more variety of tasks. All group members agree that total contribution to the project was fair and equal.

3 MAIN RESULTS OF THE ANALYSIS

3.1 **Results**

Overall, our results indicate that we were able to create a multiple linear regression model to predict sleep efficiency with a variety of significant predictors from the dataset. First, when investigating the multicollinearity assumption, we discovered that our predictors included a set of three variables that had multicollinearity with each other and had to exclude the least significant of the three from our initial additive model. When assessing the linearity of the initial additive model, we discovered that interaction and higher-order terms were required *(fig.2)*. In our initial additive model we noted that of the seven selected predictors, deep sleep percentage was most significant with a **p-value less than $2*10^{-16}$** and exercise frequency was least significant with a **p-value of 0.02267**. To obtain an interaction model that would improve our linearity and adjusted R^2 values, we used a stepwise procedure to include significant interaction terms. We discovered that smoking status, age, and awakenings all interact with deep sleep percentage, while awakenings also interact with REM sleep percentage. These interaction effects only slightly improved the linearity of the model *(fig. 3)*, suggesting that the model requires higher-order terms to meet the assumption. We tested a variety of higher-order models that included higher-order terms for deep sleep percentage, awakenings, and age. The linearity assumption for our model was improved in all cases where we added higher-order predictors, the results of one of which can be found in the appendix *(fig. 4)*. The process of removing a predictor to avoid multicollinearity and creating a higher-order model to improve linearity was an expected part of reaching our final model. With a baseline set of higher-order models established, we proceeded with investigating the other assumptions. While investigating outliers, we found that although no points had a large Cook's distance, the leverage of some points in what ended up being our final selected model *(fig. 5)* were more pronounced than the leverage for any points in our other investigated models *(fig. 6)*. This result was slightly surprising, but could likely be explained by the presence of higher-order terms with larger exponents, which could exponentially inflate the leverage of predicted points having large values for those specific higher-order terms. For the independence assumption, we know that since each row in the data is associated with a unique test subject and are not related to each other in a time-series, we can safely assume that the measurements are independent. If we suspected the measurements might not be independent, we could plot error terms in the order in which they occurred in the dataset and try to observe any pattern in the plot. The equal variance assumption was investigated using residual plots *(fig. 7)* and scale-location plots *(fig. 8)*, accompanied by Breusch-Pagan tests to determine if the models had homoscedasticity or heteroscedasticity. The null hypothesis of the Breush-Pagan test is that the model has homoscedasticity (which is what we desire for the model) and the alternate hypothesis is that the model has heteroscedasticity (undesired). One of our unused models had a **p-value of 0.03796**, meaning we rejected the null hypothesis and concluded **the model had heteroscedasticity**. However, the model we ultimately selected as our final model had a **p-value of 0.2247**, meaning we fail to reject the null hypothesis and conclude **the model has homoscedasticity**. Finally, when investigating the normality assumption, we discovered that despite our other assumptions being met and the model not having any clear outliers, the residuals were not normally distributed. Our Q-Q-plot *(fig. 9)* had a distinct bow shape and only followed the ideal diagonal line with the middle few results, suggesting that the error terms were not normally distributed. We confirmed this with a Shapiro-Wilk normality test which returned a **p-value of $2.285*10^{-5}$**, meaning we reject the null hypothesis that the residuals are normally distributed, and accept the alternate hypothesis that the residuals are not normally distributed. Although we were not able to verify all the assumptions for our selected model, we were able to solve the problems that arose with all the assumptions except for normality. Our best model that had all significant predictors, a relatively high adjusted R-squared value, and met the most assumptions can be written as:

$$
\begin{aligned}
Sleep.\widehat{efficiency} = &-2.308 + 0.0058X_{REM.sleep.percentage} + 0.1067X_{Age} + 0.0779X_{Awakenings} + 0.0071X_{Exercise.frequency} \\
&- 0.1514X_{Smoking.statusYes} - 0.0062X_{Alcohol.consumption} + 0.2858X_{Deep.sleep.percentage} \\
&- 0.0061X_{Age}^2 + 0.00017X_{Age}^3 - 0.00000216X_{Age}^4 + 0.0000000105X_{Age}^5 \\
&- 0.0148X_{Deep.sleep.percentage}^2 + 0.00036X_{Deep.sleep.percentage}^3 - 0.00000412X_{Deep.sleep.percentage}^4 + 0.0000000177X_{Deep.sleep.percentage}^5 \\
&+ 0.0022X_{Smoking.statusYes} * X_{Deep.sleep.percentage} \\
&- 0.00099X_{Awakenings} * X_{Deep.sleep.percentage} \\
&- 0.0000261X_{Age} * X_{Deep.sleep.percentage} \\
&- 0.0024X_{REM.sleep.percentage} * X_{Awakenings}
\end{aligned}
$$

Since this model contains a categorical predictor of smoking status, we can more easily interpret the model by analyzing two submodels instead; one for smokers and one for non-smokers. The model for smokers can be written as:

$$
\begin{aligned}
Sleep.\widehat{efficiency} = &-2.4594 + 0.0058X_{REM.sleep.percentage} + 0.1067X_{Age} + 0.0779X_{Awakenings} + 0.0071X_{Exercise.frequency} \\
&- 0.0062X_{Alcohol.consumption} + 0.2880X_{Deep.sleep.percentage} \\
&- 0.0061X_{Age}^2 + 0.00017X_{Age}^3 - 0.00000216X_{Age}^4 + 0.0000000105X_{Age}^5 \\
&- 0.0148X_{Deep.sleep.percentage}^2 + 0.00036X_{Deep.sleep.percentage}^3 - 0.00000412X_{Deep.sleep.percentage}^4 + 0.0000000177X_{Deep.sleep.percentage}^5 \\
&- 0.00099X_{Awakenings} * X_{Deep.sleep.percentage} \\
&- 0.0000261X_{Age} * X_{Deep.sleep.percentage} \\
&- 0.0024X_{REM.sleep.percentage} * X_{Awakenings}
\end{aligned}
$$

The model for non-smokers can be written as:

$$
\begin{aligned}
\widehat{Sleep.efficiency} = & -2.308 + 0.0058X_{REM.sleep.percentage} + 0.1067X_{Age} + 0.0779X_{Awakenings} + 0.0071X_{Exercise.frequency} \\
& - 0.0062X_{Alcohol.consumption} + 0.2858X_{Deep.sleep.percentage} \\
& - 0.0061X^2_{Age} + 0.00017X^3_{Age} - 0.00000216X^4_{Age} + 0.0000000105X^5_{Age} \\
& - 0.0148X^2_{Deep.sleep.percentage} + 0.00036X^3_{Deep.sleep.percentage} - 0.00000412X^4_{Deep.sleep.percentage} + 0.0000000177X^5_{Deep.sleep.percentage} \\
& - 0.00099X_{Awakenings} * X_{Deep.sleep.percentage} \\
& - 0.0000261X_{Age} * X_{Deep.sleep.percentage} \\
& - 0.0024X_{REM.sleep.percentage} * X_{Awakenings}
\end{aligned}
$$

The interpretation of models with interaction and higher-order predictors is difficult to clearly state, but the interpretation for the additive variables of the full model is that sleep efficiency (represented as a proportion of time spent asleep while in bed between 0 and 1) will have a baseline, or intercept, of **-2.308**. An increase in REM sleep (as a percentage of total time spent asleep) of 1 percent will increase the proportion of sleep efficiency by **0.0058**. An increase in age of 1 year will increase the proportion of sleep efficiency by **0.1067**. The age variable demonstrates a complex, non-linear relationship with sleep efficiency, as evidenced by its polynomial expansion up to the 5th degree in our model. An increase in awakenings of 1 awakening per night will increase the proportion of sleep efficiency by **0.0779**, which is counter-intuitive, but serves as a good reminder that interpreting only the additive terms does not reflect the total effect of awakenings due to the presence of interaction and higher-order terms. An increase in exercise frequency of 1 time per week will increase the proportion of sleep efficiency by **0.0071**. However, this relationship is nuanced - we observed a compensatory interaction with deep sleep percentage, where smokers show a slight increase in deep sleep's positive effect on efficiency. Importantly, this mitigating effect (**+0.0022** through the interaction term) remains substantially smaller than smoking's overall detrimental impact, resulting in a net negative influence on sleep efficiency. An increase in alcohol consumption of 1 ounce in the previous 24 hours will decrease the proportion of sleep efficiency by **0.0054**. Finally, an increase in deep sleep (as a percentage of total time spent asleep) in its additive term increases sleep efficiency by **+0.288**. However, this variable includes a polynomial term up to the 5th degree, with two of these terms being negative, which complicates the relationship with sleep efficiency. Nevertheless, it can be observed that both too little and too much sleep can be detrimental to sleep efficiency. Overall, this model has an adjusted R^2 value of **0.8544**, meaning that **85.44%** of the variation of the response variable can be explained by the model.

4 CONCLUSION AND DISCUSSION

4.1 Approach

The analysis conducted provides promising insights into how various lifestyle and physiological factors influence sleep efficiency, even if the model is not perfectly suited for individual-level predictions. The final model explains 85.44% of the variance in sleep efficiency, highlighting key influences such as exercise, alcohol, smoking, age, awakenings, and deep sleep percentage.

Alcohol consumption and smoking both demonstrate negative effects on sleep efficiency, with smoking showing a particularly strong detrimental impact. These findings align with existing medical research about substance use and sleep quality. Interestingly, while smoking generally reduces sleep efficiency, the model suggests deep sleep may slightly mitigate this effect, possibly due to nicotine's temporary relaxing properties.

The relationship between age and sleep efficiency proves complex, following a nonlinear pattern that changes across different life stages. This likely reflects how various life circumstances and health factors influence sleep differently at various ages, rather than being solely caused by biological aging itself.

Sleep architecture plays a crucial role in sleep efficiency. REM sleep shows a clear positive association with better sleep quality, supporting its importance for cognitive restoration. Deep sleep presents a more nuanced relationship, where moderate amounts are beneficial but excessive duration may become counterproductive, indicating balance is key.

The unexpected positive association between nighttime awakenings and sleep efficiency warrants further investigation. While initially counterintuitive, this effect is likely explained by the negative offset provided by the interaction between awakenings and deep sleep percentage. Individuals who are able to achieve a large percentage of deep sleep seem to be less affected by awakenings since they can still achieve enough deep sleep. The positive interpretation of awakenings might also suggest measurement limitations in the study design.

4.2 **Future Work**

Future studies should conduct a longitudinal, life-course investigation of how social and developmental factors interact with biological aging to affect sleep efficiency. This research should specifically examine how transitional life stages (e.g., adolescence to adulthood, midlife career changes, retirement) and their associated psychosocial stressors influence sleep patterns. The study design should incorporate both quantitative sleep metrics and qualitative assessments of life circumstances to better understand the nonlinear relationship captured in the current model. Second, the paradoxical awakening findings require resolution through rigorous measurement. This would clarify whether the occasionally positive association reflects measurement error or potentially some other significant pattern.

In summary, alcohol and smoking negatively impact sleep efficiency, while exercise improves it. Age has a non-linear effect, meaning it does not change sleep efficiency directly but follows a curve. This may be influenced not only by biological aging but also by lifestyle factors at different ages. Importantly, sleep duration requires balance - both insufficient and excessive sleep harm efficiency. These findings highlight modifiable factors for better sleep, though individuals should consult healthcare professionals before making significant lifestyle changes.

5 REFERENCES

List all references (e.g., books, journal, conferences, reports, magazines) you are citing in this proposal. Format:

1. Deng, Z., Liu, L., Liu, W., Liu, R., Ma, T., Xin, Y., Xie, Y., Zhang, Y., Zhou, Y., & Tang, Y. (2024). Alterations in the fecal microbiota of methamphetamine users with bad sleep quality during abstinence. BMC Psychiatry, 24(1), 324-12. https://doi.org/10.1186/s12888-024-05773-5
2. ENSIAS. (2021). Sleep Efficiency Dataset. Kaggle. Retrieved [March 11, 2025], from https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/data
3. Fjell, A. M., Sørensen, Ø., Wang, Y., Amlien, I. K., Baaré, W. F. C., Bartrés-Faz, D., Boraxbekk, C., Brandmaier, A. M., Demuth, I., Drevon, C. A., Ebmeier, K. P., Ghisletta, P., Kievit, R., Kühn, S., Madsen, K. S., Nyberg, L., Solé-Padullés, C., Vidal-Piñeiro, D., Wagner, G., . . . Walhovd, K. B. (2023). Is short sleep bad for the brain? brain structure and cognitive function in short sleepers. The Journal of Neuroscience, 43(28), 5241-5250. https://doi.org/10.1523/JNEUROSCI.2330-22.2023
4. Maslow, A. H. (1943). A theory of human motivation. Psychological Review, 50(4), 370-396.
5. Pan, L., Li, L., Peng, H., Fan, L., Liao, J., Wang, M., Tan, A., & Zhang, Y. (2022). Association of depressive symptoms with marital status among the middle-aged and elderly in rural china: Serial mediating effects of sleep time, pain and life satisfaction. Journal of Affective Disorders, 303, 52-57. https://doi.org/10.1016/j.jad.2022.01.111
6. Wang, L., & Aton, S. J. (2022). Perspective – ultrastructural analyses reflect the effects of sleep and sleep loss on neuronal cell biology. Sleep (New York, N.Y.), 45(5), 1. https://doi.org/10.1093/sleep/zsac047

APPENDIX

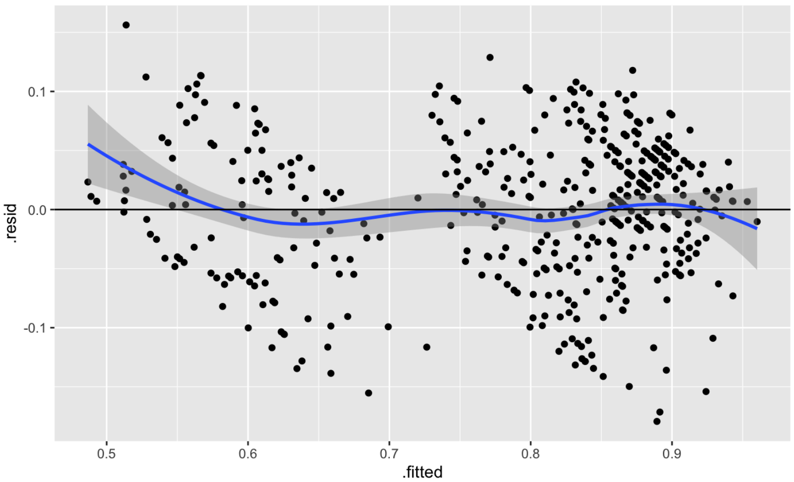
Optional part of the proposal. Include any support material you may have, e.g., figures, tables, samples of your data set(s). Make sure to cite any of the Appendix material in the main proposal text.

Table 1. Sample of the dataset

ID	Age	Gender	Bedtime	Wakeup time	Sleep duration	REM sleep %	Deep sleep %	Light sleep %	Awakenings	Caffeine consumption	Alcohol consumption	Smoking status	Exercise frequency	Sleep efficiency
1	65	Female	3/6/2021 1:00	3/6/2021 7:00	6	18	70	12	0	0	0	Yes	3	0.88
2	69	Male	12/5/2021 2:00	12/5/2021 9:00	7	19	28	53	3	0	3	Yes	3	0.66
3	40	Female	5/25/2021 21:30	5/25/2021 5:30	8	20	70	10	1	0	0	No	3	0.89
4	40	Female	11/3/2021 2:30	11/3/2021 8:30	6	23	25	52	3	50	5	Yes	1	0.51
5	57	Male	3/13/2021 1:00	3/13/2021 9:00	8	27	55	18	3	0	3	No	3	0.76
6	36	Female	7/1/2021 21:00	7/1/2021 4:30	7.5	23	60	17	0		0	No	1	0.9

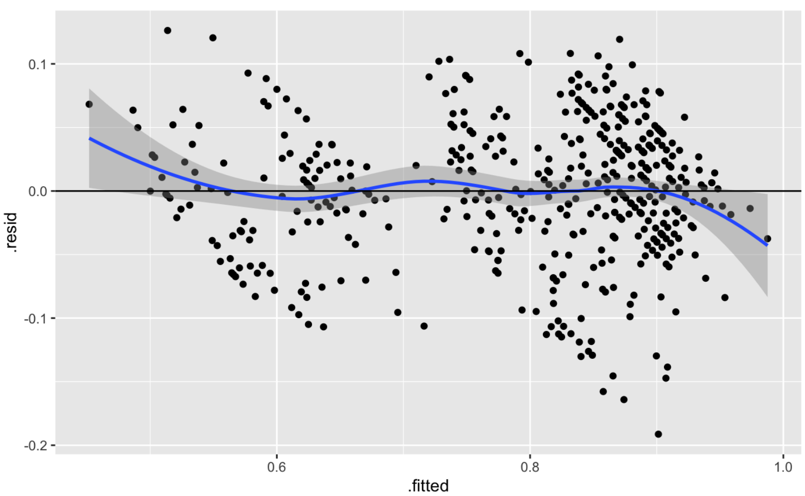
Source: ENSIAS (2021). Sleep Efficiency Dataset. Retrieved from Kaggle.

Fig 2. Initial linearity of the additive model



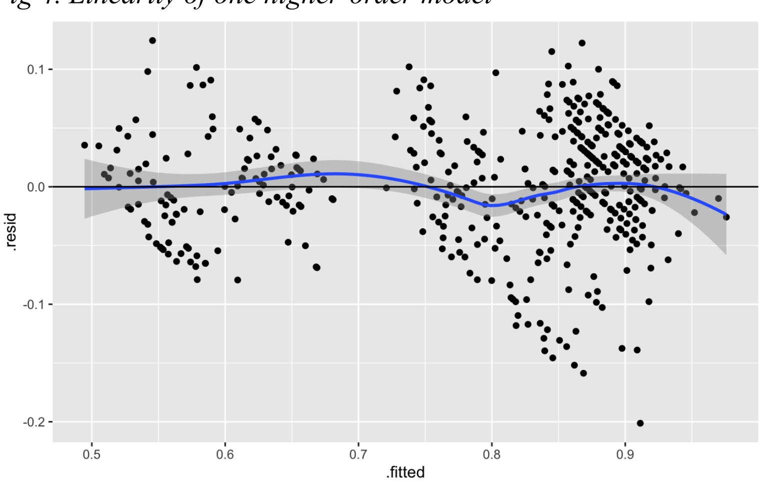
Source: Author's own work

Fig 3. Linearity of the interaction model



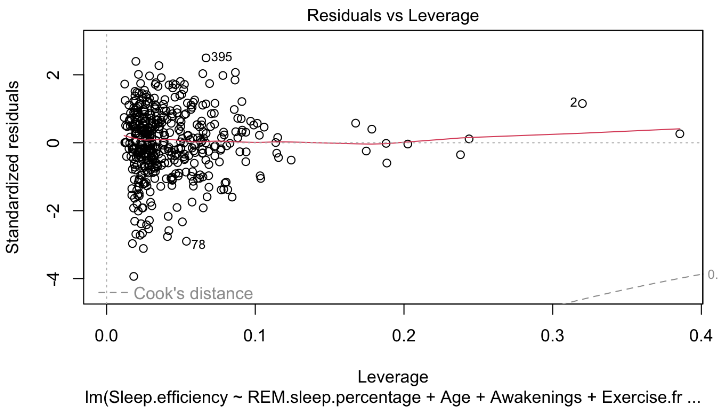
Source: Author's own work

Fig 4. Linearity of one higher-order model



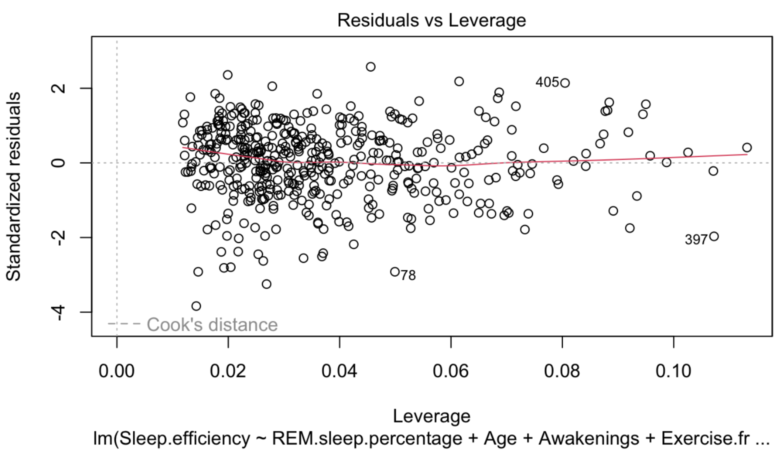
Source: Author's own work

Fig 5. Cook's distance/leverage of our selected higher-order model



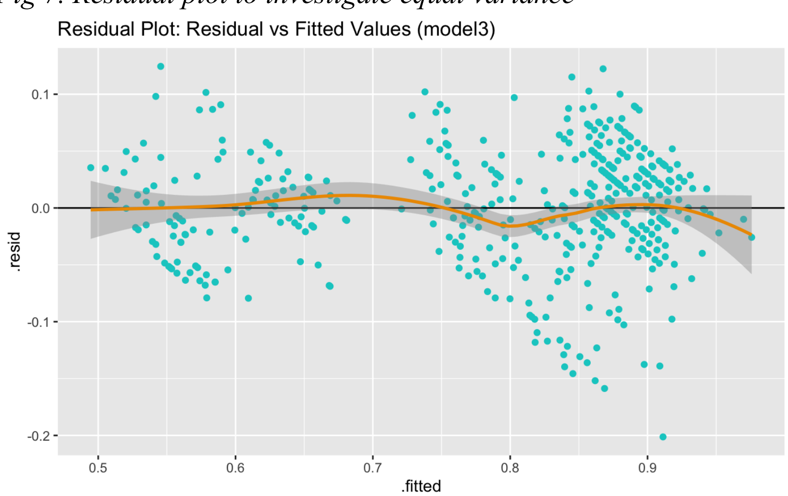
Source: Author's own work

Fig 6. Cook's distance/leverage of one of our unused higher-order models



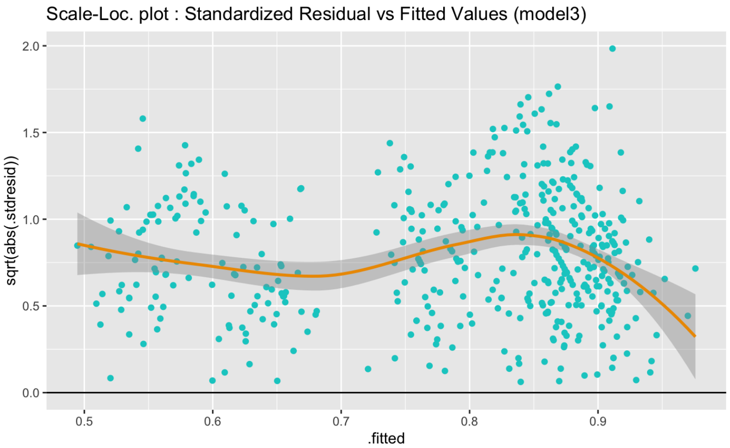
Source: Author's own work

Fig 7. Residual plot to investigate equal variance



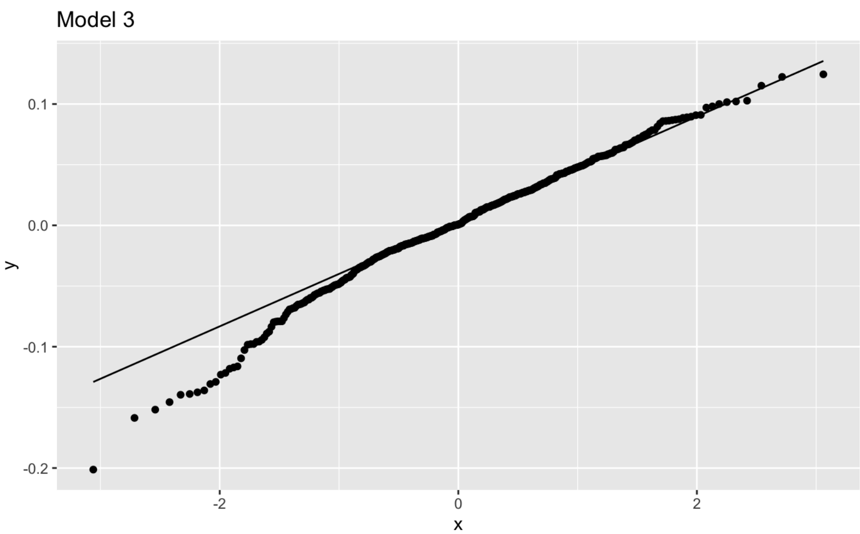
Source: Author's own work

Fig 8. Scale-location plot to investigate equal variance



Source: Author's own work

Fig 9. Q-Q-Plot for normality of residuals



Source: Author's own work

End of Project Report