Name: Ming-Lun, Lee NUID: 002449184

Course: PHTH 5210 - Biostatistics Date: 2025/04/24

Exploring the Impact of Lifestyle Factors on REM Sleep Percentage: A Multiple Linear Regression Approach

Research Question, Objective, Aim

Student usually sacrifice their sleep due to assignments, stress, or lifestyle habits. However, sleep – especially REM sleep – is an important part of students, because it plays a key role in memory, learning, and problem-solving. Recent studies indicate that alcohol and caffeine intake reduce or delay REM sleep proportion (Gardiner et al., 2024; Weibel et al., 2021). In contrast, exercise shortens the onset time of the REM stage, thereby improving sleep quality (Park et al., 2021). However, the findings on these lifestyle factors remain mixed (Vizentin et al., 2024). Therefore, the project aims to explore how lifestyle habits affect REM sleep duration. These relationships can help students develop strategies to achieve a higher proportion of REM sleep and further enhance academic performance. To better control the confounding effect, the analysis included variables of age and gender. Our **null hypothesis (H_0)** is that there is no significant relationship between lifestyle factors and REM sleep percentage (all $\beta = 0$); the **alternative hypothesis(H_1)** is that at least one lifestyle factor has a significant effect on REM sleep percentage.

Data Description

The dataset "Sleep Efficiency Dataset" was obtained from the Kaggle website (Equilibrium, 2023). It contains 452 rows of observations. The key variables include dependent variables (DV), independent variables (IV), and control variables (CV):

- DV: "REM sleep percentage" refers to the proportion of REM sleep out of the total sleep duration. It is a continuous variable measured in percentage (%).
- IV (Lifestyle Factors): "Caffeine consumption" means the amount of caffeine consumed in the 24 hours before bedtime, measured by milligrams (mg) and treated as a continuous variable. "Alcohol consumption" refers to the amount of alcohol consumed in the 24 hours before bedtime, measured in ounces (oz) and also treated as continuous. "Exercise frequency," which is initially recorded as an integer from 0 to 7, was treated as a categorical variable to better capture non-linear effects. "Smoking status" is whether the test subjects smoke or not, which is treated as a categorical variable.
- CV: "Age" is the age of the test subject, recorded as a continuous variable. "Gender" records whether the participants are male or female and is treated as a categorical variable.

Methods

Before analysis, the dataset was reviewed for missing values and data types. In the dataset, caffeine consumption (25 rows), alcohol consumption (14 rows), and exercise frequency (6 rows) had missing data and were imputed using the median due to skewed distributions and outliers. Outliers in caffeine were kept but flagged as influential points as they were considered valid values.

To examine individual associations between lifestyle factors and REM sleep percentage, we first performed **univariate analysis**. Since REM sleep percentage was not normally distributed, the Mann-Whitney test was used for binary categorical variables (e.g., gender and smoking status). Exercise frequency, converted into a categorical variable with more than two groups, was analyzed using one-way ANOVA. For continuous variables, Spearman's correlation was used. In addition, **univariate linear regression** models were performed on each variable to estimate the association between each independent variable and REM sleep percentage. Results from the regressions informed better subsequent model building in multiple linear regression.

Multiple linear regression was performed to assess the overall combined effect of predictors on REM sleep percentage. Then, we conducted model diagnostics to evaluate the violations of OLS assumption. Influential points were identified and removed using leverage, residuals, Cook's distance, and DFBETAs to examine their impact on

model fit and address assumption violations. Log transformations were subsequently applied to REM sleep percentage (DV), caffeine (IV), and alcohol consumption (IV) to address skewness. Additionally, generalized addictive models (GAM) were used to capture non-linear relationships. A stepwise method was performed to identify the best predictors and simplify the model complexity. Lastly, model selection was compared by adjusted R^2, AIC, and BIC.

Results

- Descriptive Statistics and Univariate Analyses (Table 1)

REM sleep percentage, caffeine, and alcohol consumption were not normally distributed, with the latter two showing right skewness, indicating the potential need of log-transformation. Univariate analyses revealed that gender, caffeine consumption, and exercise frequency were statistically associated with REM sleep percentage. caffeine intake showed a weak but negative correlation with REM sleep (Spearman's, $p = 0.0394$), though scatterplots suggested this was influenced by outliers. Exercise frequency showed a significant group effect (ANOVA, $p < 0.05$). Post-hoc Tukey's HSD test indicated that exercising once or three times per week was more strongly with REM sleep percentage than other frequecied, suggesting a non-linear relationship.

- Univariate Linear Regression and Initial Model

Univariate linear regressions confirmed that gender and moderate exercise frequency were significantly associated with higher REM sleep ($p < 0.05$). The ***initial multiple linear regression (M1)*** included all variables without transformation had an adjusted R^2 of 10.8%, with gender and exercise frequency as significant predictors. However, violations of normality and homoscedasticity were shown, indicating the possible reasons of influential points, data skewness, and non-linear relationship with REM sleep percentage.

- Model Adjustment and Alternative Approaches (Table 2)

To address model assumption violations, our ***second model (M2)*** excluded six influential data points (IDs: 63, 82, 97, 162, 213, 258), based on the approach mentioned in the Method section. The model retained the same set of variables as M1. Gender and exercise frequency showed associations with REM sleep percentage, whereas other factors did not reach statistical significance. The adjusted R^2 slightly decreased to 10.1%, and AIC decreased to 2337.303. In the ***third model (M3)***, we applied a log transformation to the REM sleep percentage to correct the skewness. Age became marginally significant, while the other factors remained non-significant. Although the adjusted R^2 remained stable, the AIC is the lowest (-436.3624). The ***fourth model (M4)*** applied log transformations to caffeine and alcohol consumption instead of the outcome variable. The adjusted R^2 was similar to previous models, but the AIC increased again to 2336.951. To address possible non-linear relationships, a ***fifth model (M5)*** was fitted using a Generalized Additive Model (GAM). This model showed the highest adjusted R^2 (0.147), indicating improved explanatory power. However, its AIC was higher than M3. Additionally, the assumption of normality and homoscedasticity in all models were still violated.

Therefore, in comparison, M3 (log REM) offered the best model fit (the highest AIC) with a simpler structure, while M5 (GAM) provided a better explanation rate (adjusted R^2) and captured more non-linearity patterns. To maintain a simple model, we selected M3 as the final model for stepwise selection. The final model identified through stepwise selection included three variables gender, exercise frequency, and age (Fig. 1). The final formula is *log(REM sleep (%)) = 3.064 + 0.001 * Age + (-0.075) * Gender (Male) + 0.089 * Exercise (freq = 1) + (-0.01) * Exercise (freq = 2) + 0.084 * Exercise (freq = 3) + 0.009 * Exercise (freq = 4) + 0.019 * Exercise (freq = 5)* (Table 3). As the formula shown, exercising once per week increased log REM sleep by 0.089 (95% CI = [0.048, 0.129]). Males had lower REM sleep compared to females. Other lifestyle factors were not statistically significant

Conclusion

In conclusion, the key success of this model is identifying exercise frequency as categorical rather than numerical, since it shows a non-linear relationship with the REM sleep percentage. The log transform of REM

sleep percentage corrected the skewness observed in the original linear regression model and simultaneously acheieved the lowest AIC explaining 10.5% of the variance in REM sleep percentage, indicating improved model fit and interpretability. Even though this model sacrifices the explanatory power of non-linear regression, this model is more stable due to fewer variables and retains the interpretability of a linear regression framework. Moreover, this model was statistically significant ($p < 0.05$), accepting the alternative hypothesis. However, other lifestyle habits (alcohol consumption, caffeine consumption, and smoking status) remained not statistically significant even after applying transformation. Additionally, out final model still violated the assumption of normality and homoscedaciticity, suggesting that some important variables may have been omitted. Therefore, a potential future direction is to include other factors such as stress levels or medication usage, which may better capture the underlying relationship with REM sleep percentage.

Table 1. Univariate Association with REM Sleep Percentage

Univariate Association (DV: REM%) → Not Normal Distribution				
Independent Variable	**Center**	**Spread**	**Test type**	**p-value**
Age	Median = 40	Q1 = 29 Q3 = 52 IQR = 23	Spearman's Correlation	0.67
Alcohol	Median = 0	Q1 = 0 Q3 = 2 IQR = 2	Spearman's Correlation	0.4561
Caffeine	Median = 25	Q1 = 0 Q3 = 50 IQR = 50	Spearman's Correlation	0.039*
Independent Variable		**N/Proportion**	**Test type**	**p-value**
Gender	Female (Ref)	224 (49.56%)	Mann-Whitney	7e-04*
	Male	228 (50.44%)		
Smoke	No (Ref)	298 (65.93%)	Mann-Whitney	0.484
	Yes	154 (34.07%)		
Exercise Frequency	0 (Ref)	116 (25.66%)	Kruskal Test (ANOVA)	3.061e-05*
	1	97 (21.46%)		
	2	60 (13.27%)		
	3	130 (28.76%)		
	4	41 (9.07%)		
	5	8 (1.77%)		

Name: Ming-Lun, Lee NUID: 002449184

Course: PHTH 5210 - Biostatistics Date: 2025/04/24

Table 2. Model Comparison

	Method	R-squared	Adj. R-square	AIC	BIC
M1	Linear Regression	0.128	0.108	2382.944	2432.308
M2	Linear Regression (X Influential Points)	0.121	0.101	2337.303	2386.507
M3	Log Transformation of REM Sleep %	0.121	0.101	**-436.362***	**-387.159***
M4	Log Transformation of Caffeine, Alcohol	0.121	0.1	-436.055	-386.8508
M5	GAM	X	**0.147***	2317.889	2384.16
M6*	Stepwise M3	0.119	0.105	-441.227	404.325

Table 3 Final Model Report

Variable	Estimate	95% CI	p-value
Age	0.001	[0.000, 0.002]	0.079
Gender (Male)	-0.075	[-0.107, -0.043]	4.94e-06
Exercise (freq = 1)	0.089	[0.049, 0.129]	1.61e-05
Exercise (freq = 3)	0.084	[0.044, 0.124]	4.50e-05

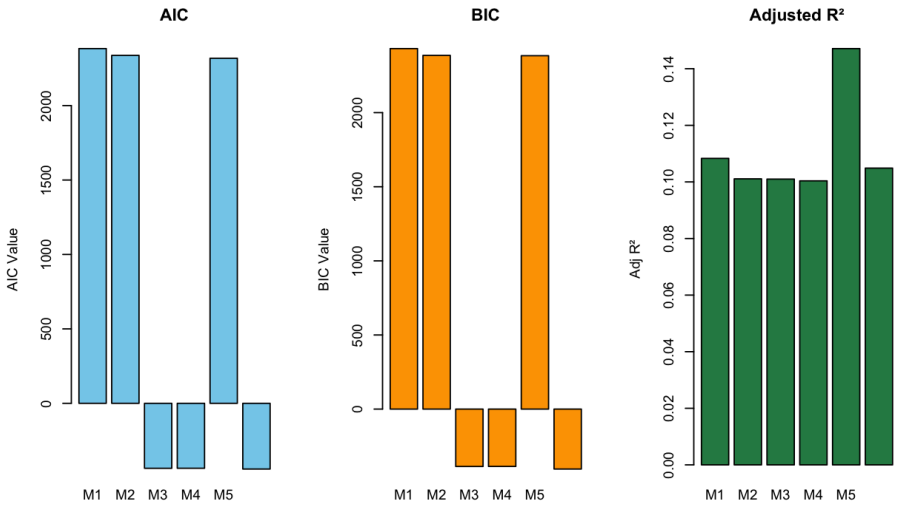

Figure 1. Model Comparison.

Name: Ming-Lun, Lee NUID: 002449184

Course: PHTH 5210 - Biostatistics Date: 2025/04/24

References

Equilibrium. (2019). Sleep Efficiency [Data set]. Kaggle. https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency

Gardiner, C., Weakley, J., Burke, L. M., Roach, G. D., Sargent, C., Maniar, N., ... & Halson, S. L. (2024). The effect of alcohol on subsequent sleep in healthy adults: A systematic review and meta-analysis. *Sleep Medicine Reviews*, 102030. https://doi.org/10.1016/j.smrv.2024.102030

Park, I., Díaz, J., Matsumoto, S., Iwayama, K., Nabekura, Y., Ogata, H., ... & Vogt, K. E. (2021). Exercise improves the quality of slow-wave sleep by increasing slow-wave stability. *Scientific reports, 11*(1), 4410. https://doi.org/10.1038/s41598-021-83817-6

Vizentin, N. P., Giannini, D. T., Takey, M., & Kuschnir, M. C. C. (2024). Caffeine consumption and association with sleep duration and screen time in Brazilian adolescents (ERICA Study). Nutrition, 118, 112233. https://doi.org/10.1016/j.nut.2023.112233

Weibel, J., Lin, Y. S., Landolt, H. P., Berthomier, C., Brandewinder, M., Kistler, J., ... & Reichert, C. F. (2021). Regular caffeine intake delays REM sleep promotion and attenuates sleep quality in healthy men. *Journal of Biological Rhythms, 36*(4), 384-394. https://doi.org/10.1177/07487304211013995

Reflection and Acknowledgment:

First of all, thank you for the helpful suggestions on the final project and throughout the entire semester, which clearly guided me to successfully understand how to conduct a statistical research. I love the way you integrated R with the theory-based course, making the course more practical!

Regarding your suggestion in class, I realized that I initially treated exercise frequency as numerical due to its ordinal characteristics. However, given its non-linear relationship with REM sleep percentage, I transformed exercise frequency into categorical variables. After this change, the model's explanatory power increased significantly.I also tried splines and the GLM model with distinct distributional approaches, but they didn't improve the adjusted R^2 or AIC. Therefore, I chose not to include them in my final report. Lastly, thank you again for all the support throughout this week.