Caffeine Consumption and Its Association with Key Health Biomarkers

Danielle, Beckett, and Alexia

Use of AI tools

We used ChatGPT to help conduct a test on the linearity of the logit assumption for logistic regressions and to create cleaner tables for our logistic regression models.

Introduction

Caffeine is the most commonly consumed drug worldwide. Despite its effectiveness as a stimulant, caffeine can cause both psychological and physiological effects on the body (Samoggia & Rezzaghi, 2021). The drug is associated with antagonistic effects on adenosine receptors, which regulate stress, arousal, and fatigue (Reichert et al., 2022). As a result, scientific inquiry has increasingly focused on the potential health impacts of caffeine.

Previous research has consistently linked caffeine use to measurable changes in key health indicators. Studies have shown increases in heart rate and blood pressure following frequent daily consumption (Green & Suls, 1996), and others have associated higher caffeine intake with reduced sleep quality (Akova et al., 2023). Findings on caffeine's relationship to stress, however, remain mixed and at times contradictory (AlAteeq et al., 2021; Lane & Williams Jr., 1987). Additionally, high caffeine consumption has been positively associated with higher BMI (Tucker & Beltran, 2025).

Despite extensive work on individual health outcomes, relatively few studies examine caffeine's effects on multiple biomarkers—such as BMI, heart rate, stress, and sleep quality—within a single analytic framework. The goal of this research is to further examine the effects of high caffeine intake on key biomarkers: BMI, heart rate, sleep quality, and stress levels. A clearer understanding of caffeine's impact on health may inform public health practices and guidelines. Identifying which groups are most at risk could help target interventions and encourage individuals to monitor their consumption before negative effects occur.

The present study uses the "Global Coffee Health Dataset," which contains 10,000 synthetic records. The dataset includes "patients" from 20 countries and provides information on daily

coffee intake, caffeine levels, sleep duration and quality, BMI, heart rate, stress, physical activity, health issues, occupation, smoking, and alcohol consumption. Its aim is to simulate real-world data on caffeine consumption, sleep, and health outcomes to support wellness research using logistic analysis and predictive modeling.

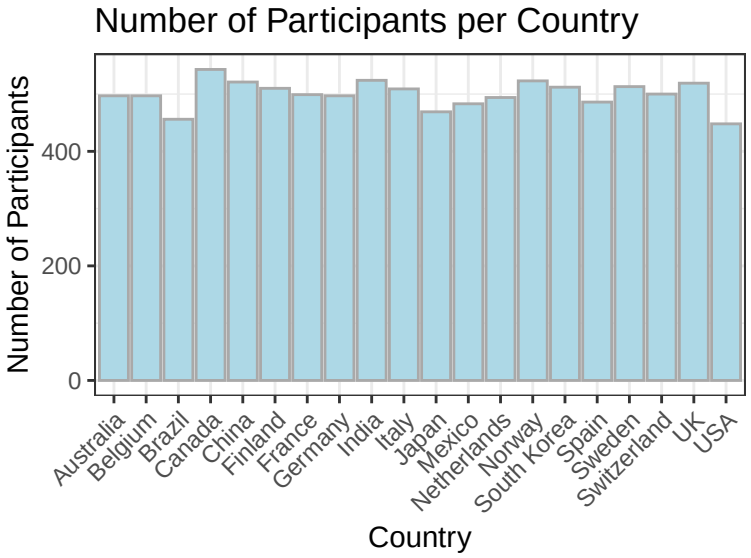

Figure 1: Distribution of Participants by Country

Overall, Figure 1 shows that the distribution is fairly balanced, with most countries contributing a similar number of participants, approximately between the mid-400s to low-500s. A few countries show slightly higher participation, while others are slightly below the average, but no extreme outliers are visible. This suggests that the sample is relatively evenly distributed across the included countries.

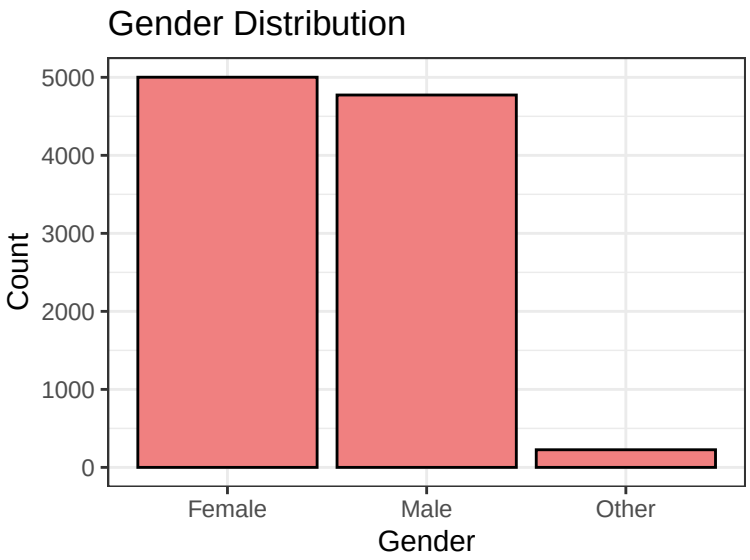

Figure 2: Distribution of Participants' Gender

Figure 2 presents the number of participants identifying as Female, Male, or Other. The majority of participants identify as either Female or Male, with both groups contributing roughly similar counts (approximately 5,000 each). A much smaller portion of the sample identifies as Other (~250), indicating limited representation in that category. Overall, the figure highlights a largely binary gender distribution within the participant pool.

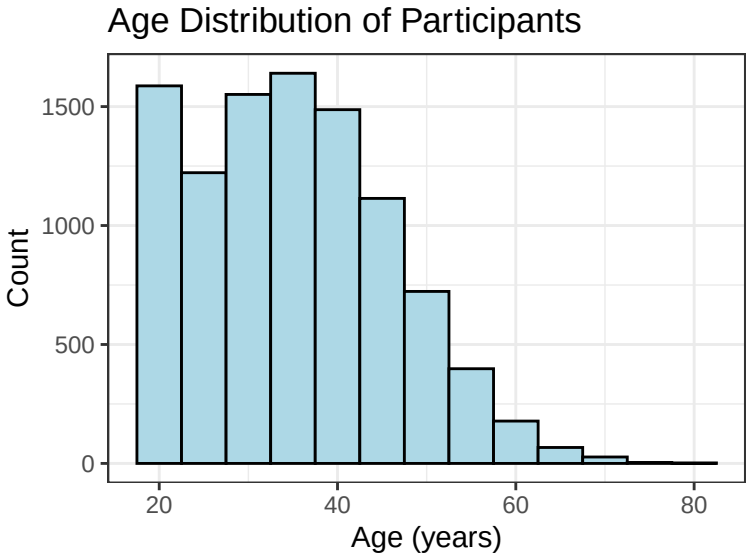

Figure 3: Distribution of Participants' Age

Figure 3 displays the distribution of participants' ages, showing that most individuals fall between their early 20s and mid-40s. The largest concentrations appear around the late 20s to late 30s, while the participation declines steadily after about age 45, with very few individuals above age 60. Overall, the age distribution is right-skewed, with its tail extending into older ages.

Methodology

This study fit logistic regression models to examine the association between high caffeine consumption and four predictors: BMI, heart rate, sleep quality, and stress level. Logistic regression was appropriate because the outcome variable—high caffeine intake—was binary, defined using the FDA threshold of 400 mg/day. All variables used in the models contained no missing values. The baseline for outcome was considered "low caffeine" intake <400 mg/day. The baseline for stress level was "high," and the baseline for sleep quality was "Excellent".

When an initial model that included both sleep quality and stress level was fitted, two of the stress-level coefficients became non-estimable due to perfect or near-perfect collinearity between the these categorical predictors (Appendix A). Because this violates the logistic regression assumption of no perfect multicollinearity, we refit the analyis using two separate logistic regression models. Model 1 included BMI, heart rate, and sleep quality, while Model 2 included BMI, heart rate, and stress level. This approach allowed each predictor to be evaluated independently without violating assumptions of the logistic regression model.

1. $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{BMI}_i + \beta_2 \text{HeartRate}_i + \beta_3(\text{Sleep}_i = \text{Good}) + \beta_4(\text{Sleep}_i = \text{Fair}) + \beta_5(\text{Sleep}_i = \text{Poor})$
2. $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{BMI}_i + \beta_2 \text{HeartRate}_i + \beta_3(\text{Stress}_i = \text{Low}) + \beta_4(\text{Stress}_i = \text{Medium})$

Our null hypothesis for each logistic regression model stated that, after holding the non-target variables (BMI, sleep quality, stress level, or heart rate) constant, the slope parameter for the predictor of interest is equal to 0; therefore, the predictor has no statistically significant association with high caffeine consumption. The alternative hypothesis stated that, after holding the non-target variables constant, the slope parameter is not equal to 0; therefore, the predictor is significantly associated with the outcome. Statistical significance was evaluated at the $\alpha = 0.05$ level.

Predictor	**Null Hypothesis (H_0)**	**Alternative Hypothesis (H_1)**
BMI	$\beta_1 = 0$	$\beta_1 \neq 0$
Heart Rate	$\beta_2 = 0$	$\beta_2 \neq 0$
Sleep: Good vs Excellent	$\beta_3 = 0$	$\beta_3 \neq 0$
Sleep: Fair vs Excellent	$\beta_4 = 0$	$\beta_4 \neq 0$
Sleep: Poor vs Excellent	$\beta_5 = 0$	$\beta_5 \neq 0$
Stress: Low vs High	$\beta_3 = 0$	$\beta_3 \neq 0$
Stress: Medium vs High	$\beta_4 = 0$	$\beta_4 \neq 0$

Table 1: Null and Alternative Hypotheses for Logistic Regression Coefficients

It is also important to evaluate the assumptions that accompany our logistic regression models: linearity of log odds and independence of observations. Although the data is synthetic, this doesn't guarantee that either assumptions are satisfied For independence to be met, each observation must provide unique data that is not influenced by another observation. In the real world of clinical trials, this kind of data would come from taking biomarkers of two patients who have no overlapping lifestyles, where neither of these patients has the opportunity to affect the data of the other. However, an example of non-independent data would be if the data was taken from siblings. In this study, we proceed under the assumption that independence holds. Linearity of the log-odds for continuous predictors was assessed using the Box–Tidwell procedure, which provided no evidence of violation (Appendix B). Overall, the modeling approach is replicable and appropriately aligned with the requirement of a binary outcome variable; however, conclusions remain subject to the inherent uncertainty surrounding assumption verification in synthetic datasets. With these considerations, our logistic regression framework offers a reasonable method for evaluating how levels of caffeine consumption relate to various health markers.

Results

Table 2 presents a logistic regression examining the association between high caffeine intake and sleep quality, including BMI and heart rate as additional predictors.

Variable	**Estimate**	**P-value**	**Odds Ratio**	**95% CI for OR**
(Intercept)	-3.31	$< 2 \times 10^{-16}$	0.036	(0.020, 0.067)
BMI	-0.000289	0.971	0.9997	(0.985, 1.015)
Heart Rate	0.00910	0.00333	1.009	(1.002, 1.016)
Sleep Quality: Fair	0.981	8.75×10^{-15}	2.67	(2.08, 3.42)
Sleep Quality: Good	0.612	2.47×10^{-7}	1.84	(1.46, 2.33)
Sleep Quality: Poor	1.33	$< 2 \times 10^{-16}$	3.79	(2.90, 4.95)

Table 2: Logistic Regression of High Caffeine Intake (with Sleep Quality Predictor)

Table 3 shows a logistic regression examining the association between high caffeine intake and stress level, with BMI and heart rate as additional predictors.

Variable	**Estimate**	**P-value**	**Odds Ratio**	**95% CI for OR**
(Intercept)	-1.98	5.22×10^{-11}	0.138	(0.077, 0.249)
BMI	-0.00124	0.873	0.999	(0.984, 1.014)
Heart Rate	0.00944	0.00230	1.009	(1.003, 1.016)
Stress Level: Low	-0.816	$< 2 \times 10^{-16}$	0.442	(0.372, 0.526)
Stress Level: Medium	-0.351	0.000459	0.704	(0.579, 0.857)

Table 3: Logistic Regression of High Caffeine Intake (with Stress Level Predictor)

Table 2 and Table 3 indicate that there is sufficient evidence that several physiological and behavioral factors have statistically significant relationships with high caffeine consumption. In both tables, heart rate emerged as a significant positive predictor (p=0.002), with each one-unit increase in heart rate multiples the participant's predicted odds of high caffeine intake by 1.0095, holding all other variables constant. Sleep quality in Table 2 also demonstrated strong associations: individuals reporting fair, good, or poor sleep quality had statistically significantly higher odds of high caffeine consumption compared to the reference group, as indicated by the positive log odds coefficients and p-values below 0.05, again holding all other variables constant. Stress level showed a similar pattern in Table 3; when all other variables are held constant, both low- and medium-stress categories significantly predicted the odds of high caffeine intake relative to the high-stress reference group (both $p < 0.001$), with patients in these categories showing odds multiplied by 0.442 and 0.704, respectively. In contrast, when we hold all other variables constant, BMI was not a significant predictor ($p = 0.97$) in either Table 2 or Table 3, indicating there was insufficient evidence to support a statistically significant association between BMI and high caffeine consumption in this dataset.

Discussion

The primary goal of this study was to evaluate how caffeine consumption relates to key health markers—BMI, heart rate, sleep quality, and stress level. These findings partially support our hypothesis that caffeine intake would be associated with heart rate, sleep, and stress but not BMI, leading us to successfully identifying several physiological and behavioral variables that predict high caffeine intake.

Our logistic regression models effectively identify the health factors associated with high caffeine consumption. Heart rate, sleep quality, and stress level consistently emerged as significant predictors of consuming more than 400 mg of caffeine per day. These results suggest that caffeine intake is related to physiological and psychological states, particularly stress and sleep patterns. In contrast, BMI did not appear to be a significant predictor.

Logistic regression was an appropriate starting point for predicting a binary threshold of high caffeine intake. However, several limitations emerged. Category imbalance and redundancy between variables prevented certain levels of sleep quality and stress from being included simultaneously. Alternative approaches, such as regularized logistic regression (LASSO or ridge), decision trees, or random forest models, could better handle correlated predictors and complex interactions (Wohlwend, 2023). Additionally, assessing multicollinearity formally using variance inflation factors would further strengthen the reliability of model selection.

Data quality also introduces constraints on validity. Because the dataset consists of synthetic health records, the associations identified may not fully reflect real-world variability in caffeine use or health outcomes. Synthetic data often smooths over noise, reducing natural variance between observations and producing artificial linear relationships (Melton, 2025). In addition, due to being synthetic data, contextual information about the caffeine intake can not be found though they may influence caffeine's physiological effects.

If this study were conducted again in the future, several improvements would strengthen the analysis. First, incorporating real-world observational data would significantly improve external validity and correlations. Second, including additional variables, such as socioeconomic status, occupation, genetics, and chronic health conditions, would offer a more comprehensive understanding of caffeine's impact as well as allowing us to conclude stronger relationships.

As caffeine intake continues to rise, especially in younger populations, its crucial to understand the relationship between its use and health markers (Branum et al., 2014). Further research into this field may inform discussion about current initiatives to ban beverages with large caffeine concentrations. Similarly, a study conducted using a population of only adolescents may provide a scientific base for evaluating laws that prevent the purchase of caffeinated beverages that exceed a caffeine threshold.

Overall, our findings suggest that high caffeine consumption is strongly associated with elevated heart rate, poorer sleep quality, and heightened stress, while BMI shows no significant relationship with high caffeine intake. Although the synthetic nature of the dataset limits external validity, the patterns identified here highlight the potential value of monitoring stress and sleep behaviors when assessing caffeine-related health risks. Future studies using real clinical or observational data—and incorporating richer demographic and lifestyle variables—will be essential for clarifying the mechanisms linking high caffeine intake to cardiometabolic and psychological outcomes. Ultimately, this work underscores that caffeine consumption reflects measurable physiological and behavioral states that may help guide health assessment and intervention.

References

[1] Akova, İ., Duman, E. N., Sahar, A. E., & Sümer, E. H. (2023). The Relationship Between Caffeine Consumption and Depression, Anxiety, Stress Level and Sleep Quality in Medical

Students. *Journal of Turkish Sleep Medicine*, 10(1), 65–70. https://doi.org/10.4274/jtsm.galenos.2022.06078

[2] AlAteeq, D. A., Alotaibi, R., Saqer, R. A., Alharbi, N., Alotaibi, M., Musllet, R., & Alraqibah, R. (2021). Caffeine consumption, intoxication, and stress among female university students: A cross-sectional study. *Middle East Current Psychiatry*, 28(1). https://doi.org/10.1186/s43045-021-00109-5

[3] Branum, A. M., Rossen, L. M., & Schoendorf, K. C. (2014). Trends in Caffeine Intake Among US Children and Adolescents. *Pediatrics*, 133(3), 386–393. https://doi.org/10.1542/peds.2013-2877

[4] Caffeine in Tea vs. Coffee: How Do They Compare? (n.d.). Retrieved November 20, 2025, from https://www.healthline.com/nutrition/caffeine-in-tea-vs-coffee#caffeine-concerns

[5] Commissioner, O. of the. (2024). *Spilling the Beans: How Much Caffeine is Too Much?* FDA. https://www.fda.gov/consumers/consumer-updates/spilling-beans-how-much-caffeine-too-much

[6] Green, P. J., & Suls, J. (1996). The effects of caffeine on ambulatory blood pressure, heart rate, and mood in coffee drinkers. *Journal of Behavioral Medicine*, 19(2), 111–128. https://doi.org/10.1007/BF01857602

[7] Lane, J. D., & Williams Jr., R. B. (1987). Cardiovascular Effects of Caffeine and Stress in Regular Coffee Drinkers. *Psychophysiology*, 24(2), 157–164. https://doi.org/10.1111/j.1469-8986.1987.tb00271.x

[8] Melton, A. (2025). *Synthetic Data in Healthcare: When It Works & When It Fails*. Invene.com. https://www.invene.com/blog/synthetic-data-healthcare

[9] Reichert, C. F., Deboer, T., & Landolt, H. (2022). Adenosine, caffeine, and sleep–wake regulation: State of the science and perspectives. *Journal of Sleep Research*, 31(4), e13597. https://doi.org/10.1111/jsr.13597

[10] Reyes, C. M., & Cornelis, M. C. (2018). Caffeine in the Diet: Country-Level Consumption and Guidelines. *Nutrients*, 10(11), 1772. https://doi.org/10.3390/nu10111772

[11] Samoggia, A., & Rezzaghi, T. (2021). The Consumption of Caffeine-Containing Products to Enhance Sports Performance: An Application of an Extended Model of the Theory of Planned Behavior. *Nutrients*, 13(2), 344. https://doi.org/10.3390/nu13020344

[12] Supplements, P. C. for a W. on P. H. H. A. with C. of C. in F. and D., Board, F. and N., Policy, B. on H. S., & Medicine, I. of. (2014). Intake and Exposure to Caffeine. In *Caffeine in Food and Dietary Supplements: Examining Safety: Workshop Summary*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK202226/

[13] Temple, J. L., & Ziegler, A. M. (2011). Gender Differences in Subjective and Physiological Responses to Caffeine and the Role of Steroid Hormones. *Journal of Caffeine Research*, 1(1), 41–48. https://doi.org/10.1089/jcr.2011.0005

[14] Tucker, L. A., & Beltran, F. (2025). Use of caffeine in 19,660 randomly selected U.S. adults: The role of overweight and obesity. *Frontiers in Nutrition*, 12, 1588447. https://doi.org/10.3389/fnut.2025.1588447

[15] Verweij, K. J. H., Treur, J. L., & Vink, J. M. (2018). Investigating causal associations between use of nicotine, alcohol, caffeine and cannabis: A two-sample bidirectional Mendelian randomization study. *Addiction*, 113(7), 1333–1338. https://doi.org/10.1111/add.14154

[16] Wohlwend, B. (2023, July 14). Three Regression Models for Data Science: Linear Regression, Lasso Regression, and Ridge Regression. *Medium*. https://medium.com/@brandon93.w/three-regression-models-for-data-science-linear-regression-lasso-regression-and-ridge-regression-6aac73c0d7a5

Appendix A

This study fit two logistic regression models examining the associations between high caffeine consumption and four predictors: BMI, sleep quality, stress level, and heart rate. All variables had no missing values and therefore required no imputation. An initial model including all four predictors was attempted:

Variable	**Estimate**	**P-value**
(Intercept)	-3.31	$< 2 \times 10^{-16}$
BMI	-0.0003	0.971
Heart Rate	0.009	0.00333
Sleep Quality: Fair	0.981	8.75×10^{-15}
Sleep Quality: Good	0.612	2.47×10^{-7}
Sleep Quality: Poor	1.332	$< 2 \times 10^{-16}$
Stress Level: Low	NA	NA
Stress Level: Medium	NA	NA

Table 4: Logistic Regression of High Caffeine Intake (with All Predictors)

However, Table 4 indicates that two coefficient estimates were not defined due to singularities, specifically for the Low and Medium categories of stress level. This suggests perfect or near-perfect collinearity between the categorical predictors (sleep quality and stress level), preventing the model from estimating all parameters. Because the full model could not be reliably estimated, the analysis was separated into two logistic regression models: one including sleep quality (but excluding stress level), and one including stress level (but excluding sleep quality). This approach resolved the singularity issue and allowed each predictor to be evaluated without collinearity constraints.

Appendix B

The test for the linearity of log odds can only be conducted on continuous variables, (e.g., BMI, Heart Rate). Categorical variables (Sleep_Quality, Stress_Level) can't be tested. Categorical variables are converted to dummy variables automatically in R if they are factors, which we confirmed they were using the factor function. The model estimates log-odds differences relative to the reference, so there is no linearity assumption because each category is treated separately. To test the linearity of the logit assumption in logistic regression in R, we ran a Box–Tidwell test (logit linearity test).

Variable	**Estimate**	**P-value**
(Intercept)	-4.27	0.148
BMI	0.061	0.837
BMI_log	-0.015	0.836
Heart Rate	0.055	0.759
HR_log	-0.009	0.798
Sleep Quality: Fair	0.981	9.0×10^{-15}
Sleep Quality: Good	0.612	2.5×10^{-7}
Sleep Quality: Poor	1.33	$< 2 \times 10^{-16}$

Table 5: Box–Tidwell Test for Logit Linearity Test

Table 5 shows that the p-values for the BMI_log and HR_log terms (0.836 and 0.798, respectively) both exceed 0.05. This indicates insufficient evidence of non-linearity, and therefore the linearity assumption is considered to be satisfied.