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, a drug consumed daily by 80% of the world population, produces both psychological and physiological effects on the body (Samoggia & Rezzaghi, 2021). The drug creates antagonistic effects on adenosine receptors in the body, which are responsible for the regulation of stress, arousal, and a decrease in fatigue (Reichert et al., 2022). As a result of this, numerous health markers such as BMI, heart rate, cortisol level, and sleep quality are affected by the consumption of caffeine, where research conducted has shown an increase in heart rate and blood pressure levels after frequent daily consumption of the drug (Green & Suls, 1996).

In addition, other research showed an association with chronic caffeine use and other addictions such as nicotine and alcohol. However, others argue that no causal relationship exists between caffeine consumption and the use of other addictive substances like alcohol, nicotine, cannabis, etc. Research has shown that causal relationships that may exist between caffeine and other addictive substances are not phenotypical associations but are due to other disadvantages/liabilities (Verweij et al., 2018). These disadvantages suggest that a causal relationship between caffeine, alcohol, and smoking may exist beyond purely genetic factors.

While caffeine consumption is nearly universal, differences in consumption and effect have been noted in literature across demographic differences. Consumption habits are impacted by geographical origins, culture, and socioeconomic status (Reyes & Cornelis, 2018). This further affects preferred brewing methods and beverage choice across groups, creating a wide range of amounts of caffeine per standard drink in a nation. A prime example of this is that one would need to consume about 5 cups of black tea to get an equal caffeine amount to a standard American coffee (Caffeine in Tea vs. Coffee: How Do They Compare?, n.d.). This

trend seems to continue for most Western countries, as they tend to consume more caffeine than their Eastern counterparts. Further discrepancies in consumption have been observed in between age ranges and sex, with 50-64 year old women consuming the most amount of caffeine (Supplements et al., 2014; Temple & Ziegler, 2011).

The goal of this research is to further pinpoint the effect of caffeine on key biomarkers: BMI, Heart Rate, Sleep Quality, and Stress levels. In addition, it aims to note differences in consumption and health outcomes based on demographic differences: gender, age, and nationality. Further analysis of caffeine's impact on health could inform health practices and guidelines. A specific breakdown of what groups are most at risk could help to warn groups to be mindful of their consumption before negative impacts set in.

The data explored is from the "Global Coffee Health Dataset," which contains 10,000 synthetic records. The data's "patients" span 20 countries, each with their own health outcome values for daily coffee intake, caffeine levels, sleep duration and quality, BMI, heart rate, stress, physical activity, health issues, occupation, smoking, and alcohol consumption. The goal of the data set is to reflect real-world data on caffeine consumption, sleep, and health outcomes, to allow for wellness studies using logistical analysis and predictive modeling.

Methodology

To show the differences in patient demographics across caffeine consumption habits, distributions of participants by country, gender, and age were considered. For the categorical variables, country and gender, a bar chart was used to visualize counts. For age, a continuous variable, a histogram was used to plot the overall numeric values of different variables, showing a frequency distribution of values, which groups data into continuous intervals. On the other hand, a bar chart was used for the categorical variables, as their length/height represents the quantity of each variable, rather than a type of distribution. Therefore, each respective visualization represents both the continuous and categorical patient demographics in the dataset.

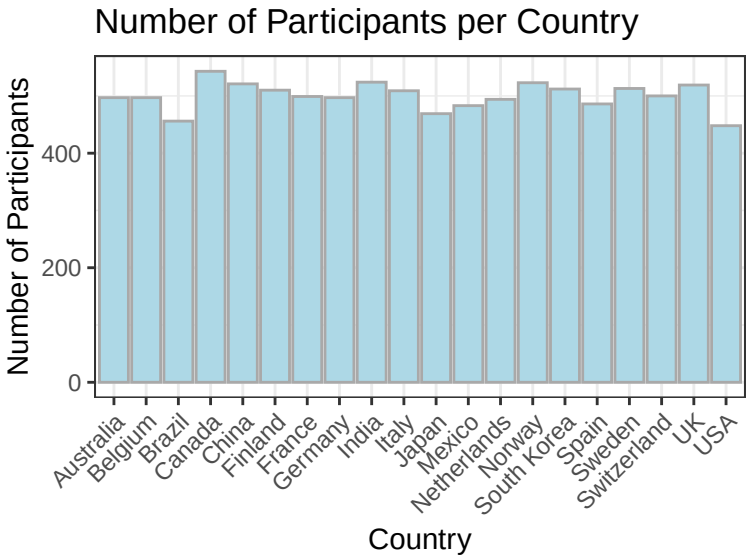

Figure 1: Distribution of Participants by Country

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.

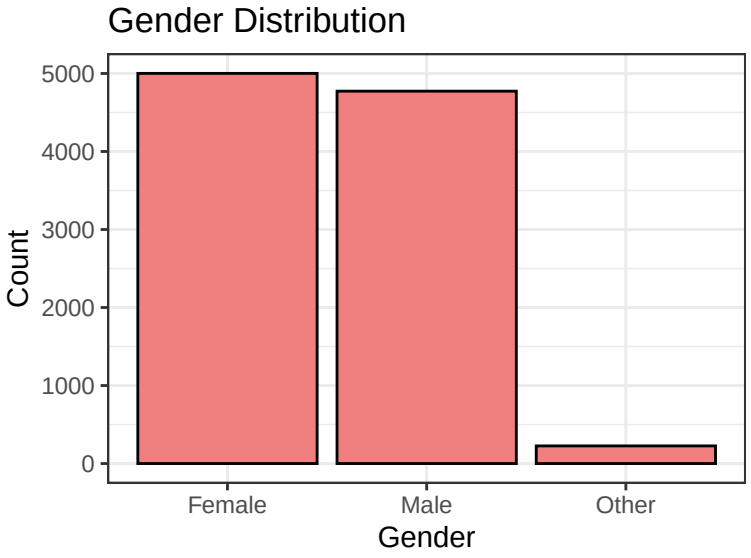

Figure 2: Distribution of Participants' Gender

Figure 2 shows 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 (around 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.

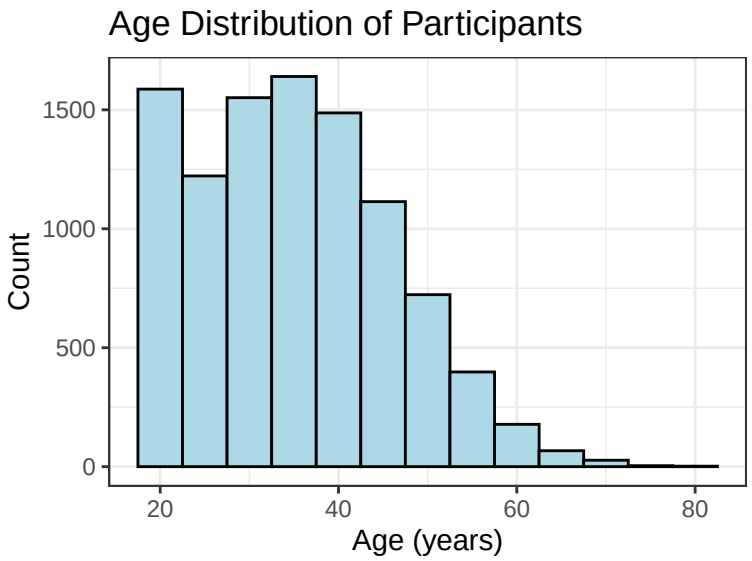

Figure 3: Distribution of Participants' Age

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.

Logistic regression was chosen because the outcome variable is binary (high vs. low caffeine intake). To build our logistic regression model, we found our binary outcome variable based on FDA guidance indicating that daily intake up to 400 mg is associated with minimal negative effects (Commissioner, 2024). This study defined high caffeine intake as above 400 mg and low caffeine intake as below 400 mg, which then set the threshold for the binary variable accordingly.

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, where there existed no missing data values (therefore not considered/included). Because including both sleep quality and stress level in the same logistic regression model produced collinearity issues, the predictors could not be simultaneously estimated. To address this, we fit two separate

logistic regression models: one including sleep quality and one including stress level. This approach allowed each variable to be evaluated without violating model assumptions (refer to Appendix B).

It is also important to analyze the assumptions that accompany these tests: linearity of log odds and independence of observations. Although the data is synthetic, this doesn't prove that either assumptions are valid. 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 assume the former. There is insufficient evidence of non-linearity after conducting a Box–Tidwell test, so the linearity assumption holds (see Appendix A).The methodology provided is replicable and assumes that the requirement of a binary outcome variable, followed by a logistic regression model, can predict (with limitations due to uncertainty of assumptions verification) how levels of caffeine consumption relate to various health markers.

Results

Table 1 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**
(Intercept)	-3.31	1.85×10^{-26}
BMI	-0.000289	0.971
Heart Rate	0.00910	0.00333
Sleep Quality: Fair	0.981	8.75×10^{-15}
Sleep Quality: Good	0.612	2.47×10^{-7}
Sleep Quality: Poor	1.33	1.67×10^{-22}

Table 1: Logistic Regression of High Caffeine Intake (with Sleep Quality Predictor)

Table 2 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
(Intercept)	-1.98	5.22×10^{-11}
BMI	-0.00124	0.873
Heart Rate	0.00944	0.00230
Stress Level: Low	-0.816	$< 2 \times 10^{-16}$
Stress Level: Medium	-0.351	0.000459

Table 2: Logistic Regression of High Caffeine Intake (with Stress Level Predictor)

Table 1's and Table 2's results indicate that several physiological and behavioral factors meaningfully predict high caffeine consumption, defined as intake above 400 mg per day. In both Table 1 and 2, Heart rate emerged as a significant positive predictor, with each unit increase in heart rate increasing the odds of high caffeine intake ($p = 0.003$). Sleep quality in Table 1 also showed strong and consistent associations, where individuals reporting fair, good, or poor sleep quality had significantly higher odds of high caffeine consumption compared to the reference group, as reflected by large positive coefficients and highly significant p-values (all greatly less than 0.05). Stress level was similarly influential in Table 2, with both low and medium stress categories significantly decreasing the odds of high caffeine intake relative to the high-stress reference group (both $p < 0.001$), suggesting that individuals experiencing higher stress are more likely to consume large amounts of caffeine. In contrast, BMI was not a significant predictor ($p = 0.97$) in either Table 1 or Table 2, indicating no meaningful relationship between body mass and caffeine consumption in this dataset. Overall, the results show that caffeine consumption has a significant relationship with sleep quality, stress, and heart rate, while BMI plays no significant role.

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—and to determine whether these relationships differ across demographic groups. Our logistic regression results partially address these goals. We successfully identified several physiological and behavioral variables that predict high caffeine intake, but we were unable to evaluate demographic differences due to issues with multicollinearity and category imbalance.

The results of our logistic regression models directly address our goal of identifying which health and lifestyle markers are associated with high caffeine consumption. Across the models, heart rate, sleep quality, and stress level consistently emerged as significant predictors of consuming more than 400 mg of caffeine per day. This aligns with our research questions by revealing that caffeine intake is not evenly distributed across individuals but is tied to physiological and psychological states, particularly stress and sleep-related outcomes. Importantly, BMI was not a significant predictor, indicating that caffeine consumption level is not meaningfully associated with this dataset.

Our interpretation of the statistical models is grounded clearly and correctly in the results produced by the GLM outputs. For example, the significant positive coefficient for heart rate suggests that individuals with higher resting heart rates are more likely to consume high levels of caffeine—a finding consistent with known physiological responses to stimulants. The statistically significant effects of sleep quality and stress level further support the conclusion that caffeine intake is intertwined with sleep disturbances and heightened stress.

However, our models do not fully address our secondary goal of identifying demographic patterns. Variables such as age, gender, and nationality could not be included due to multicollinearity and the structure of the synthetic dataset. As a result, we were unable to determine whether certain demographic groups are at greater risk of consuming excessive caffeine or experiencing related health consequences.

Logistic regression was an appropriate starting point for predicting a binary threshold of caffeine intake. However, several limitations emerged. Category imbalance and redundancy between variables caused singularities in the model, preventing 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, future work done in assessing multicollinearity formally using variance inflation factors would 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 demographic groups and producing artificial linear relationships (Melton, 2025). In addition, due to being synthetic data, contextual information about the caffeine intake, like type of caffeinated beverage, timing of consumption 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.

Overall, our findings suggest that high caffeine consumption is strongly associated with elevated heart rate, poor sleep quality, and higher stress, but not BMI. While the results partially meet the goals of the study, limitations in data quality and model constrain the conclusions that can be drawn. Future studies employing larger, real datasets and more flexible modeling techniques would better capture the behavioral and physiological mechanisms underlying caffeine use.

Sources

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

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

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

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

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

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

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

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/

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

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 (Abingdon, England), 113(7), 1333–1338. https://doi.org/10.1111/add.14154

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

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 3: Box–Tidwell Test for Logit Linearity Test

Table 3 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.

Appendix B

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, the model output indicated 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.