Project Part 3

Describe (in a maximum of 3 *new* pages) the project that you proposed in part 1 (300 points)

a. **Include (the same) a short description describing the project that you proposed in part 1.**
This research study investigates the relationship between habits (such as caffeine intake, frequency of physical activity, and sleep hours) and university students' productivity. Using a correlational study design with a target sample of 30-50 students, anonymous surveys will collect data on variables like caffeine consumption patterns, exercise frequency, sleep duration, and daily productivity ratings (on a 0-100 scale). The original proposal included tracking over time. However, based on feedback, the study will focus on data collection "on a given day" rather than weekly measurements due to survey anonymity requirements. The analysis will primarily use multiple regression and correlation analyses to examine relationships between these variables.

b. **State what kind of analysis you did, and what your result is for every "test statistic" that you conducted (ie a 2 way ANOVA has a test statistic for each of the two main effects, and one for the interaction, so what was the result of each). This needs to be written like a results section: no commentary, just the results in APA format.**

   i. **Multiple regression analysis:**
   A multiple regression analysis was conducted to predict daily productivity based on work/study hours, sleep hours, energy fluctuations, age, consistent sleep, and exercise frequency. The overall model was significant [$R^2 = 0.597$, $F(6, 37) = 9.12$, $p < .001$]. Significant predictors included work/study hours ($B = 0.075$, $SE = 0.028$, $t(37) = 2.67$, $p = .011$), sleep hours ($B = -0.309$, $SE = 0.093$, $t(37) = -3.33$, $p = .002$), and energy fluctuations ($B = 0.374$, $SE = 0.144$, $t(37) = 2.60$, $p = .013$). These findings indicate that these predictors significantly contribute to daily productivity while holding all other variables constant. Non-significant predictors included age ($B = -0.031$, $SE = 0.020$, $t(37) = -1.56$, $p = .131$), consistent sleep ($B = 0.164$, $SE = 0.090$, $t(37) = 1.82$, $p = .078$), and exercise frequency ($B = 0.017$, $SE = 0.051$, $t(37) = 0.35$, $p = .729$).

   For productivity consistency, the regression model was also significant [$R^2 = 0.541$, $F(6, 37) = 7.26$, $p < .001$]. Energy fluctuations significantly predicted productivity consistency ($B = 0.829$, $SE = 0.170$, $t(37) = 4.88$, $p < .001$), and exercise frequency showed a marginally significant effect ($B = 0.111$, $SE = 0.058$, $t(37) = 1.93$, $p = .064$). Age ($B = -0.024$, $SE = 0.022$, $t(37) = -1.03$, $p = .308$), work/study hours ($B = -0.013$, $SE = 0.033$, $t(37) = -0.41$, $p=.684$), sleep hours ($B = -0.177$, $SE = 0.111$, $t(37) = -1.59$, $p = .112$, and consistent sleep ($B = -0.036$, $SE = 0.106$, $t(37) = -0.34$, $p = .737$) were not significant predictors above and beyond other variables.

   ii. **Pearson correlation analysis:**

A Pearson correlation analysis revealed significant positive correlations between energy fluctuations and daily productivity (r = .543, p < .001), energy fluctuations and productivity consistency (r = .674, p<.001), and work/study hours and daily productivity (r = .529, p < .001). Consistent sleep was positively correlated with daily productivity (r = .375, p = .038) and productivity consistency (r = .311, p = .048). Sleep hours are negatively associated with daily productivity (r = −0.436, p = .004).

These results indicate that while energy fluctuations, work/study hours, and sleep hours were significant predictors of daily productivity in the regression model, energy fluctuations were the strongest predictor of productivity consistency. Other predictors, while showing significant correlations, were not significant in the regression model when holding all other variables constant.

c. **Did the results support your hypothesis or not? That is, could you reject the null hypothesis, and thereby show support for your hypothesis? What type of error might you be making (type I or II), and what is the chance of that error? How do you know (ie where/what did you do to find the chance of that error you might be making)?**

Support for Hypothesis:

1. Regression Analysis:
    a. Daily Productivity:
        i. The regression analysis provided partial support for the hypothesis. Significant predictors included:
            a. Work/Study Hours (B = 0.075, SE = 0.028, t(37) = 2.67, p = .011)
            b. Sleep Hours (B = −0.309, SE = 0.093, t(37) = −3.33, p = .002)
            c. Energy Fluctuations (B=0.374,SE=0.144,t(37)=2.60,p=.013)
        ii. These results allow for the rejection of the null hypothesis for these predictors, indicating that they contribute significantly to daily productivity above and beyond other variables.
        iii. Non-significant predictors in the regression model included:
            1. Age (p=.131)
            2. Consistent Sleep (p=.078)
            3. Exercise Frequency (p=.729)
    b. Productivity Consistency:
        i. The regression model identified Energy Fluctuations as a significant predictor (B = 0.829, SE = 0.170, t(37) = 4.88, p < .001), providing strong evidence to support the hypothesis.
        ii. Exercise Frequency was marginally significant (B = 0.111, SE = 0.058, t(37) = 1.93, p = .064), suggesting weak evidence.
        iii. Non-significant predictors included:
            1. Age (p=.308)
            2. Work/Study Hours (p=.684)
            3. Sleep Hours (p=.112)

            4. Consistent Sleep (p=.737)

2. Pearson Correlation Analysis:
    a. Pearson correlation analysis revealed significant bivariate relationships, supporting the hypothesis for several predictors:
        i. Daily Productivity:
            1. Energy Fluctuations (r = .543, p < .001)
            2. Work/Study Hours (r = .529, p < .001)
            3. Sleep Hours (r = −.436, p = .004)
            4. Consistent Sleep (r = .375, p = .038)
        ii. Productivity Consistency:
            1. Energy Fluctuations (r = .674, p < .001)
            2. Consistent Sleep (r = .311, p = .048)

Type of Error:

1. Type I Error (False Positive):
    a. A Type I error occurs when the null hypothesis is wrongly rejected.
    b. The probability of a Type I error is controlled by the significance level (α = 0.05), which applies to both the regression and correlation analyses.
    c. For significant results, such as work/study hours, sleep hours, energy fluctuations, and their corresponding correlations, there is a 5% chance of making a Type I error.
2. Type II Error (False Negative):
    a. A Type II error occurs when a false null hypothesis is not rejected.
    b. For non-significant predictors in the regression analysis (e.g., age, consistent sleep, exercise frequency), the likelihood of a Type II error increases due to limited statistical power (1 − β), small effect sizes, and modest sample size (n = 37).

Chance of Error:

1. Type I Error:
    a. The probability of a Type I error is set at α = 0.05 for all tests conducted in both the regression and correlation analyses.
2. Type II Error:
    a. Type II errors are more likely for predictors with non-significant results in the regression model, such as consistent sleep and exercise frequency.
    b. These errors may occur due to insufficient power or small effect sizes, as indicated by higher p-values, such as p = 0.131.

How It Was Determined:

1. Regression Analysis:
    a. The significance level (α = 0.05) was used to evaluate each predictor in the regression model.
    b. Significant results indicated a low likelihood of Type I error for predictors such as work/study hours, sleep hours, and energy fluctuations.

2. Pearson Correlation Analysis:
    a. Bivariate relationships were explored using Pearson r, which provided additional insights into the predictors' relationships with the outcomes.
    b. Significant correlations confirmed the relationships for predictors like energy fluctuations and consistent sleep that were not consistently significant in the regression model.
3. Power of the Test (1−β):
    a. Power was not explicitly computed but inferred from the sample size and observed effect sizes.
    b. Larger p-values (e.g., p=.308) and marginal significance (p = .064) suggest potential Type II errors due to insufficient power.

d. If you did not need to do a power analysis for c above, conduct a power analysis now to see the likelihood you had of finding an effect as big as the one you found (also what was your effect size? – make sure you present Pearson's R or Cohen's D)

Power Analysis:

1. Main Effect of Work/Study Hours on Daily Productivity:
    - Power: The power analysis for the main effect of work/study hours indicated a high power of 0.99, suggesting a robust likelihood of detecting the observed effect if it truly exists.
    - Effect Size (Cohen's d): d = 1.96
    - Effect Size (Pearson's r): r = 0.529
    - Common Language Effect Size: 83.54%
2. Main Effect of Sleep Hours on Daily Productivity:
    - Power: The power for the main effect of sleep hours was high at 0.97, indicating a strong likelihood of detecting the observed effect if it exists.
    - Effect Size (Cohen's d): d = 1.96
    - Effect Size (Pearson's r): r = −0.436
    - Common Language Effect Size: 78.14%
3. Main Effect of Energy Fluctuations on Productivity Consistency:
    - Power: The power for the main effect of energy fluctuations was substantial at 1.00, indicating a very low risk of Type II error.
    - Effect Size (Cohen's d): d = 2.31
    - Effect Size (Pearson's r): r = 0.674
    - Common Language Effect Size: 89.91%
4. Interaction Effect of Evening Caffeine and Exercise Frequency on Daily Productivity:
    - Power: The power for the interaction effect of evening caffeine and exercise frequency was moderate at 0.78, indicating a 22% chance of a Type II error (1−Power = 0.221).
    - Effect Size (Cohen's d): d = 0.90
    - Effect Size (Pearson's r): r = 0.418
    - Common Language Effect Size: 81.19%

Interpretation:

- The high power for the main effects of work/study hours, sleep hours, and energy fluctuations indicates a strong ability to detect the observed effects, and the effect sizes suggest substantial practical significance (all Cohen's d > 1.0).
- For the interaction effect of evening caffeine and exercise frequency, the moderate power of 0.78 indicates a small risk of failing to detect a true effect. The effect size (d = 0.90) suggests it is moderately related to the interaction on daily productivity.
- The Common Language Effect Size adds an intuitive understanding of the practical significance, with values above 80% for the main effects and interaction indicating a meaningful proportion of non-overlapping distributions.
- The findings emphasize the importance of work/study hours, stable energy, and targeted caffeine use for productivity. Larger sample sizes are needed for better exploration of interaction effects.

e. **Write a short discussion section like what you might see in a paper. What are the *big* take away conclusions from your results? I.e., what do they mean? Interpret them for readers. This is the commentary that you left out of b above. <u>What are the limitations of your study?</u> *For example: Did you have any confounds? Was your sample representative? Other limitations? Also mention from d above if your power was too low (If your power was high but you still didn't find an effect, what does this mean?).* Think critically about your own design and study.**

Discussion:

The results of our study provide valuable insights into the factors influencing daily productivity and productivity consistency. The main finding reveals a significant effect of energy fluctuations on both daily productivity and productivity consistency, with a substantial effect size (d = 2.31). This suggests that stable energy levels are strongly associated with improved productivity outcomes. The observed effect size implies a large practical significance, indicating a meaningful contribution of energy fluctuations to consistent performance.

Work/study hours also showed a significant positive effect on daily productivity, with a large effect size (d = 1.96). This finding highlights the importance of structured time allocation in enhancing productivity. Conversely, sleep hours are negatively related to daily productivity (d = 1.96), suggesting a potential short-term trade-off between sleep and productivity, though this raises concerns about long-term sustainability.

Comparison Between Regression and Correlation Findings:

Notably, consistent sleep and exercise frequency did not emerge as significant predictors in the regression model, despite showing significant moderate correlations with productivity metrics in the Pearson correlation analysis. For instance, consistent sleep was positively correlated with daily productivity (r = .375, p = .038) and productivity consistency (r = .311, p = .048). This discrepancy suggests that while consistent sleep may play a role when examined in isolation, its impact is less pronounced when accounting for other variables in the regression model. This distinction highlights the added value of multivariate analyses in identifying independent effects of predictors.

Similarly, exercise frequency was marginally significant (p = .064) in predicting productivity consistency in the regression model but did not reach statistical significance in the correlation analysis. These differences underscore the importance of considering both bivariate and multivariate contexts when interpreting predictors' effects.

Limitations:

Several limitations of our study should be noted. First, the sample size (n = 37) was relatively small, potentially limiting the generalizability of the findings to broader populations. The reliance on self-reported data introduces biases, such as over- or under-reporting productivity and related behaviors. Additionally, the cross-sectional design prevents us from establishing causal relationships, making it unclear whether the predictors directly drive productivity or are consequences of it.

Furthermore, potential confounding variables, such as stress levels, dietary habits, and work environment, were not accounted for and may have influenced the results. Future studies should address these factors by including them as covariates in the analysis to better isolate the effects of the predictors.

Power Analysis and Interpretation:

The power analysis underscores the importance of interpreting non-significant results with caution. For most predictors, the power was high (Power > 0.95), suggesting that the lack of significance reflects a true absence of independent effects rather than insufficient sensitivity. However, the moderate power (Power = 0.78) for the interaction effect of evening caffeine and exercise frequency leaves a 22% chance of a Type II error, highlighting the need for further investigation with larger samples.

The significant predictors, such as energy fluctuations, work/study hours, and sleep hours, demonstrated robust effects in both the regression model and the correlation analysis, reinforcing their critical roles in productivity outcomes. In contrast, the lack of significance for consistent sleep and exercise frequency in the regression model suggests their contributions may be more context-dependent.

Future Directions:

To address these limitations, future research should:

1. Incorporate Objective Measures: Use wearable devices to track sleep patterns, energy levels, and physical activity to reduce self-report biases.
2. Account for Confounders: Include variables like stress levels, dietary habits, and workload as covariates to better isolate the predictors' effects.

3. Adopt Longitudinal Designs: Conduct studies over time to establish causal relationships and assess the long-term sustainability of trade-offs between factors like sleep and productivity.
4. Increase Sample Diversity: Recruit larger and more diverse samples to improve generalizability and ensure that findings are representative of various populations.

Conclusion:

In conclusion, our study highlights the critical role of energy fluctuations and structured work/study hours in enhancing productivity. While some predictors, such as consistent sleep and exercise frequency, were not significant in the regression model, their potential contributions should not be dismissed, as suggested by their moderate correlations with productivity metrics. Future research should aim to overcome the limitations identified here to provide a more comprehensive understanding of the complex dynamics underlying productivity and consistency. By doing so, researchers can better inform interventions to enhance productivity in both personal and professional contexts.

f. Include as an appendix that includes both previous assignments (part 1 and 2), which you can update based on comments you've received and/or new understanding making sure you highlight all changes in red font. If you don't have any changes you can make, then put a note in red at the top of the appendix indicating that. As a second appendix, also provide a screenshot of G*power showing one of your power analyses. These appendices dont count towards your page "limit"

Project Part 1 Feedback:

Overall, this is good work. However, ALL sentences need to be complete sentences; there are a lot of places where your bullets don't make complete sentences. Furthermore, please take care of all your continuous variables, and make sure that the order is correct for each- like constant productivity isnt the end of that morning-to-night scale, so perhaps ask them to choose the MOST productive, even if its mostly constant. Then you can treat these as continuous variables, and as we talked about in class, you should do so. Please go back to the analysis section and correct when you have treated continuous variables as categorical, as it will allow you to fully use all the data rather than reducing the informativeness of it. Also, it is not practical to have volunteer participants to repeat this survey, and also to track them over time accordingly, especially while making it anonymous. You have not explained how you would do so either. Please go ahead instead and remove that plan, which seems to just be mentioned at the end in analysis anyway. So change all mentions to on a given day, rather than any mention of weekly. Also once you specify all the DVs as continuous that can be, please readdress your plan and your justification as to that they are the right analysis. This is implied, but be clearer about it in each set of analyses. Also, be careful with your language to use exactly what was given in class regarding the analyses: for example, "Pearson Correlation Analysis - Detect Continuous DV Effects over Productivity": one this needs to be a complete sentence, two the phrase itself is unclear/awkward, and three, its to detect a relationship between continuous *IVs* and productivity as your DV. There are also no main effects when you talk about correlation, this is only a needed term for analyses when there are interactions. Please go carefully though all analyses and check for correctness with the lectures.

Project part 2 feedback:

Great work! The progress brought you right up to par. Unfortunately, you did not include the whole assignment for part 1 (missing question 1). But overall very good job to make all those changes and improvements! Its also better not to use the word "influence" and rather use "related to" or "associated with" or "predicts" because influence can be taken as causal language. I prefer if your "bullets" dont try to make incomplete sentences, as you have in the Part 1 still. Also you still have a placeholder in there "such as XYZ"

Project Part 2

Describe (in a maximum of 2 *new* pages) the project that you proposed in part 1 (200 points)

1. **Include a short description outlining the project that you proposed in part 1.**
   This research study investigates the relationship between habits (such as caffeine intake, frequency of physical activity, and sleep hours) and university students' productivity. Using a correlational study design with a target sample of 30-50 students, anonymous surveys will collect data on variables like caffeine consumption patterns, exercise frequency, sleep duration, and daily productivity ratings (on a 0-100 scale). The original proposal included tracking over time. However, based on feedback, the study will focus on data collection "on a given day" rather than weekly measurements due to survey anonymity requirements. The analysis will primarily use multiple regression and correlation analyses to examine relationships between these variables.
2. **Describe the dataset you will have. What are the variables? Which are IVs, and which are DVs? Within IVs and DVs, which are categorical and which are continuous?**
   The dataset for this research study will consist of variables related to university students' habits and productivity levels. The focus will be on understanding how variables like hours of sleep, frequency of physical activity, and caffeine consumption are related to daily productivity. Below is a detailed description of the variables included in the dataset, categorized into independent variables (IVs) and dependent variables (DVs).

Independent Variables (IVs):

1. **Continuous Variables:**
   a. Caffeine consumption frequency:
   Participants will track the exact number of servings they consume each day.
   b. Caffeine consumption timing:
   Participants will track caffeine timing in hours.
   c. Physical activity frequency in Hours:
   This variable measures the frequency of Activity, measured as the total hours of exercise per week.

d. Physical activity timing:
   Like caffeine consumption timing, this variable captures physical activity timing in hours after activity for a continuous scale.
e. Energy level Rating:
   Energy Level Variations will be measured on a continuous scale from 0 to 100.
f. Sleep schedule consistency:
   This variable assesses how consistent participants are with their sleep schedules; it tracks the number of days per week that the sleep schedule is consistent.
g. Hours of sleep per night:
   This variable measures the average hours of sleep participants get each night.

2. **Categorical Variables:**
   a. **None** (As we are working only with Multiple Regression Analysis and Pearson Correlation Analysis)

Dependent Variables (DVs):

1. **Continuous Variables:**
   a. Daily productivity score:
      Participants will rate their overall productivity for the day on a continuous scale from 0 to 100.
   b. Time-specific productivity levels:
      Measures productivity at different times (morning, afternoon, evening) on a continuous scale from 0 to 100.
   c. Productivity Consistency:
      Record productivity consistency as the number of days the productivity level is steady
2. **Categorical Variables:**
   a. **None** (As we are working only with Multiple Regression Analysis and Pearson Correlation Analysis)

3. **Describe the data pre-processing steps that you believe you will need to do. What data cleaning? Will you impute values from missing data? If so, how? What other pre-processing might you need to do, if any?**
   Several data pre-processing steps will be necessary to ensure the dataset is clean and ready for analysis. These steps are critical to ensure the integrity and quality of the data, ultimately affecting the validity of the results.

   1. **Data Cleaning**
      Data cleaning involves identifying and correcting errors or inconsistencies in the dataset. For this study, potential issues might include:
      - **Outliers:** Check for outliers in continuous variables such as hours of sleep or physical activity frequency. Outliers could result from participants entering incorrect or exaggerated values (e.g., reporting 20 hours of sleep per night or 15 days of physical activity in a week). These values should be flagged and either corrected (if possible) or removed if they are erroneous.
      - **Inconsistent Responses:** Ensure that all responses align with the continuous nature of the variables. For example, if a participant provides a categorical response for caffeine consumption (e.g., "Never"), this needs to be converted into a continuous measure (e.g., 0 cups of caffeine per day).

- **Duplicate Entries:** Check for duplicate responses from participants, ensuring no participant has submitted multiple surveys.

2. **Handling Missing Data**

   Missing data is common in survey-based studies. It is important to decide how to handle missing values to avoid bias in the analysis. There are several strategies for dealing with missing data:

- **Imputation:** If there are missing values for continuous variables such as hours of sleep or physical activity frequency, imputation can be used. A common approach is:
  - **Mean Imputation:** Replace missing values with the mean value of that variable across all participants. This method works well when the amount of missing data is small and when the data is normally distributed.
  - **Median Imputation:** If the data is skewed (e.g., sleep hours), using the median instead of the mean may be more appropriate.
  - **Mode Imputation:** For categorical variables like caffeine consumption timing or physical activity timing, missing values can be replaced with the most frequent category (mode).
- **Listwise Deletion:** If a participant has too many missing responses (e.g., more than 50% of their survey is incomplete), their entire response may be removed from the dataset. This method should be used sparingly to avoid reducing the sample size too much.

3. **Bucketing and Transformation Adjustments**
   a. **Bucketing:**
      i. **Caffeine Intake:** To simplify analysis and capture usage patterns, caffeine intake can be bucketed into categories such as "Low," "Moderate," and "High," based on daily consumption frequency. These buckets will help in understanding productivity correlations for different intake levels.
      ii. **Sleep Duration:** Bucket hours of sleep into ranges, such as "Short" (0-4 hours), "Moderate" (5-7 hours), and "Optimal" (8+ hours), to analyze sleep's impact on productivity more distinctly.
      iii. **Physical Activity Frequency:** Physical activity could be categorized into "Sedentary," "Moderate," and "Active" based on weekly exercise hours. These buckets provide an easier way to examine productivity variations related to activity levels.
   b. **Transformation:**
      i. **Scaling:** Normalize continuous variables like hours of sleep, caffeine intake, and physical activity using min-max scaling. This approach brings all variables to a common scale (0-1), ensuring regression and correlation analysis consistency.
      ii. **Log Transformation:** If data for caffeine intake frequency or productivity is skewed, a log transformation can be applied to reduce skewness and stabilize variance, which enhances the model’s robustness.
   c. **Handling Outliers:**
      i. **Bucketing as an Outlier Solution:** For variables with extreme values, such as high caffeine consumption, bucketing also contains outliers within a specific range.

4. **Converting Categorical Variables to Continuous**

Since both Multiple Regression and Pearson Correlation require continuous variables, any categorical variables need to be transformed into continuous measures:

- **Caffeine Intake:** Instead of using categories like "Never," "Rarely," or "Daily," caffeine intake should be recorded as a continuous variable based on the number of caffeinated beverages consumed per day (e.g., cups of coffee, tea, energy drinks). If participants have provided categorical responses, these can be converted into approximate daily consumption values (e.g., "Never" = 0 cups/day, "Rarely" = 0.5 cups/day, etc.).
- **Physical Activity Frequency:** Convert physical activity frequency into total hours per week rather than days per week. This provides a more precise measure of physical activity that can be treated as continuous.
- **Sleep Consistency:** If sleep schedule consistency was originally measured on an ordinal scale (e.g., "Very Consistent" to "Very Inconsistent"), this needs to be transformed into a continuous metric. For instance, participants could report how many days per week they maintain a consistent sleep schedule (0–7 days).

5. **Normalization/Scaling**

   For continuous variables such as hours of sleep and physical activity frequency, normalizing or scaling the data may be necessary before conducting regression analysis, especially if these variables have widely different ranges.

   - **Min-Max Scaling:** This technique scales all continuous variables to a range between 0 and 1, ensuring that no variable is disproportionately related to the results due to its scale.
   - **Z-score Normalization:** This method standardizes continuous variables by subtracting the mean and dividing by the standard deviation, resulting in a distribution with a mean of 0 and a standard deviation of 1.

   Normalization is particularly important when running algorithms like regression that assume all predictors are on comparable scales.

6. **Addressing Survey Response Bias**

   Since this study relies on self-reported data, there may be biases such as social desirability bias or recall bias. While these biases cannot be eliminated through pre-processing, it's important to:

   - **Check for Response Patterns:** Identify any participants who may have given identical responses across all questions (e.g., selecting "Neutral" for every item), which could indicate careless responding.
   - **Cross-check Responses:** For questions that should logically align (e.g., caffeine intake frequency and timing), cross-check responses for consistency.

Conclusion

In summary, the pre-processing steps will include:

1. Cleaning data by addressing outliers, inconsistencies, and duplicates.
2. Handling missing data through imputation or listwise deletion.
3. Bucketing and Transformation Adjustments
4. Converting categorical variables into continuous measures where necessary.

5. Normalizing/scaling continuous variables if necessary.
6. Checking for response patterns to mitigate survey biases.

These steps will ensure that the dataset is clean and ready for robust statistical analysis, allowing us to accurately explore relationships between habits like caffeine intake, physical activity, sleep quality, and productivity among university students.

4. State what kind of analysis you intend to do, and what your hypothesis is for every "test statistic" that you will generate (ie a 2 way ANOVA has a test statistic for each of the two main effects, and one for the interaction, so generate a hypothesis for each).

We intend to do Multiple Regression Analysis and Pearson Correlation Analysis

Hypotheses for Test Statistics

With the help of t-tests, we will isolate each of the relations in the regression model, determining the strength of each of the independent variables about the dependent one, as well as employing multiple regression analysis and Pearson correlation analysis to determine how three independent variables related to the dependent variable.

Multiple Regression Analysis and T-tests Hypotheses

Multiple regression analysis shall be carried out assisted with t-tests for each of the independent variables in order to test whether or not each of them is significantly related to productivity provided that other variables are already controlled.

1. Caffeine Intake (IV1)
   - ○ Null Hypothesis (H0): under this hypothesis, the regression coefficient for caffeine intake β1 is equal to zero which means that caffeine intake does not significantly contribute towards enhancing daily productivity in relation to physical activity and sleep quality.
     - ■ H0:β1=0
   - ○ Alternative Hypothesis (H1): under this hypothesis, the regression coefficient for caffeine intake β1 is not equal to zero which means that caffeine intake significantly contributes towards augmenting one's productivity levels on a daily basis in relation to physical activity and sleep quality.
     - ■ H1:β1!=0
2. Physical Activity (IV2)
   - ○ Null Hypothesis (H0): the regression coefficient for physical activity, β2, is equal to a zero which implies that physical activity is not viewed as an important determinant of daily productivity level after the caffeine dosage and the quality of sleep has been considered.
     - ■ H0:β2=0
   - ○ Alternative Hypothesis (H1): the regression coefficient for physical activity, β2, is not equal to a zero and this implies that physical activity is practiced on all days as a

contributor to the daily productivity level even with the consideration of the caffeine dosage and the quality of sleep.

- H1:β2!=0

3. Sleep Quality (IV3)
    - Null Hypothesis (H0): Sleep quality has a coefficient, β3, that is equal to zero. This means that when we factor in the caffeine dosage and physical activity, sleep quality is not one of the key factors explaining the daily productivity level.
        - H0:β3=0
    - Alternative Hypothesis (H1): The regression coefficient for sleep quality (β3)is not equal to zero, meaning sleep quality has a significant effect on daily productivity after accounting for caffeine intake and physical activity.
        - H1:β3!=0

Pearson Correlation Analysis Hypotheses

We will also conduct a Pearson correlation analysis to assess the relationship between each independent variable (IV) and the dependent variable (DV).

1. Correlation between Caffeine Intake and Productivity
    - Null Hypothesis (H0): There is no significant correlation between caffeine intake and productivity.
        - H0:r=0
    - Alternative Hypothesis (H1): There is a significant correlation (either positive or negative) between caffeine intake and productivity.
        - H1:r!=0
2. Correlation between Physical Activity and Productivity
    - Null Hypothesis (H0): There is no significant correlation between physical activity and productivity.
        - H0:r=0
    - Alternative Hypothesis (H1): There is a significant correlation between physical activity and productivity (either positive or negative).
        - H1:r!=0
3. Correlation between Sleep Quality and Productivity
    - Null Hypothesis (H0): There is no significant correlation between sleep quality and productivity.
        - H0:r=0
    - Alternative Hypothesis (H1): There is a significant correlation between sleep quality and productivity (either positive or negative).
        - H1:r!=0

Summary of Analysis Approach:

- We will conduct multiple regression analyses to understand the combined effects of caffeine intake, physical activity, and sleep quality on productivity.
- For each independent variable in the regression model, we will perform a t-test to determine if the regression coefficient (β) is significantly different from zero, meaning that the variable significantly impacts productivity after controlling for other variables.
- We will also conduct Pearson correlation analysis to determine the strength and direction of the relationships between each IV (caffeine intake, physical activity, sleep quality) and DV (productivity).
  - The Null hypothesis for correlation analysis (H0:r=0)implies no relationship.
  - An Alternative hypothesis for correlation analysis (H1:r!=0) implies a significant relationship.

By using multiple regression analysis and Pearson correlation, we will be able to determine each variable's unique contribution to productivity and the overall strength and direction of the relationship between these lifestyle habits and productivity levels.

5. **Include as an appendix the entire assignment from part 1, which you can update based on comments you've received and new understanding, making sure you highlight all changes in red font. This doesn't count towards your page "limit"… also if you don't have any changes you can make, then put a note in red at the top of the appendix indicating that**

Project Part 1

CHANGES MADE TO PART 1

1. **A short summary (1/2 page to 1 page) of:**

   Shruti Subramanyam

   a. **Your interests**

      From early in my career, I've been drawn to how data can uncover patterns and improve decision-making. Working across industries, I saw firsthand how data-driven solutions boost efficiency. A key experience was at Rakuten, where applying market basket analysis improved promotions, directly enhancing customer satisfaction and sales. This deepened my interest in using data for business growth and addressing societal challenges, particularly in healthcare. The combination of problem-solving and making a broader impact continues to drive my passion for data science.

   b. **The reasons why you choose your current degree and major**

      I chose to pursue a Master's in Applied Data Science at USC to deepen my technical skills and explore emerging areas like Generative AI, Natural Language Processing, and Computer Vision. While my data science career has been rewarding, I wanted a structured academic approach to build a stronger foundation in advanced topics. USC's hands-on projects and research opportunities align perfectly with my growth goals in this field.

   c. **The reasons why you decided to take this class**

I enrolled in the "Research Methods and Analysis for User Studies" course to learn how to design and conduct research involving human subjects. While my work has focused on data analytics, I recognized the need to understand the ethical and methodological aspects of working with human data. This course will help me ensure that my future research is both scientifically sound and ethically responsible.

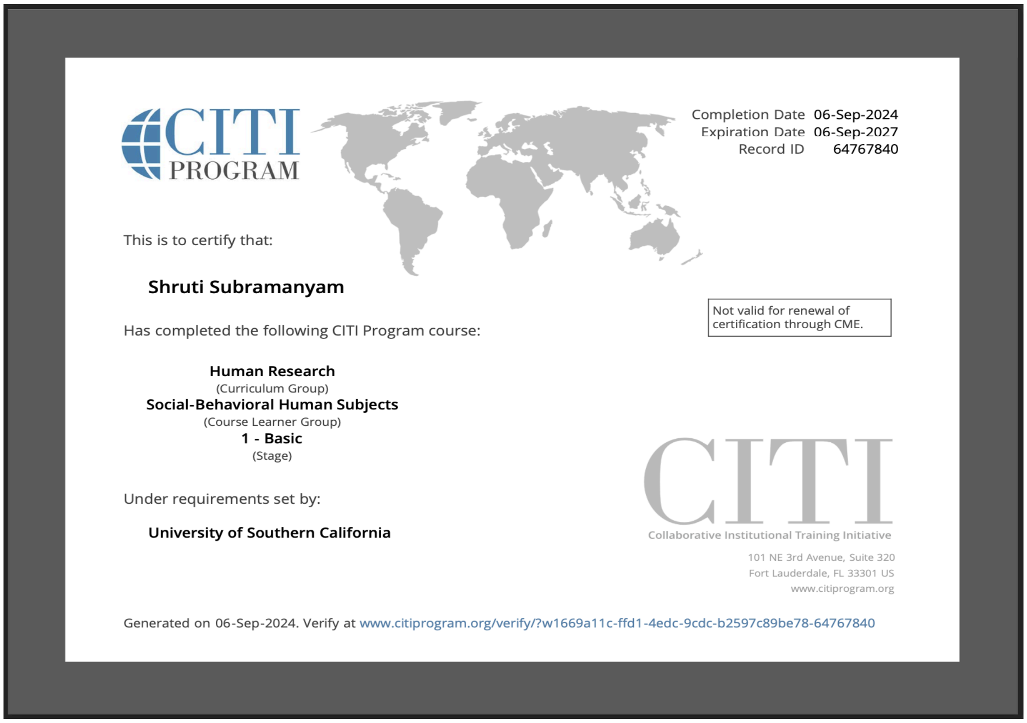
d. **Your personal ambitions to change the world**

I aspire to use my data science skills to make a meaningful impact on healthcare systems. One of my key ambitions is to develop AI-powered solutions that can revolutionize healthcare diagnostics and patient care, making it more personalized and accessible. I believe that through data-driven research and innovations, we can significantly improve the quality of life for individuals, especially in underserved populations where access to healthcare is limited.

e. **The reasons why you are interested in the topic you have chosen for your project.**

I chose this topic because I'm fascinated by how daily habits like caffeine intake, physical activity, and sleep can impact productivity and well-being. This research aligns with my interest in using data to improve everyday life. By studying these correlations, I aim to identify trends that could lead to practical recommendations for healthier, more productive lifestyles, which makes this project exciting for me.

f. **Show me a screenshot of your CITI certification for human subjects research.**



CITI PROGRAM

Completion Date 06-Sep-2024
Expiration Date 06-Sep-2027
Record ID 64767840

This is to certify that:

Shruti Subramanyam

Not valid for renewal of certification through CME.

Has completed the following CITI Program course:

Human Research
(Curriculum Group)
Social-Behavioral Human Subjects
(Course Learner Group)
1 - Basic
(Stage)

Under requirements set by:

University of Southern California

CITI
Collaborative Institutional Training Initiative

101 NE 3rd Avenue, Suite 320
Fort Lauderdale, FL 33301 US
www.citiprogram.org

Generated on 06-Sep-2024. Verify at www.citiprogram.org/verify/?w1669a11c-ffd1-4edc-9cdc-b2597c89be78-64767840

2. **Sketch out the plan for the user study that you will conduct this term, including details such as:**

   a. **What variables are you going to collect?**

      i. **Caffeine Intake:** For **Caffeine Intake**, the variables collected will include **Frequency of Consumption**, which will be measured on a continuous scale based on the **number of caffeinated beverages** consumed daily, such as cups of coffee, tea, or energy drinks.

Participants will track the exact number of servings they consume each day. The **Timing of Consumption** will be recorded as the number of **hours after waking up** when caffeine is consumed. This replaces categorical time blocks like "Morning" or "Afternoon" and allows for more granular data. Participants will log the exact time (e.g., 1 hour after waking, 3 hours after waking) at which they consume caffeine.

ii. **Physical Activity:** For **Physical Activity**, the key variables include **Frequency of Activity**, measured as the total **hours of exercise per week**, turning the data into a continuous variable. Participants will log the total time spent engaging in physical activity each week, allowing for a more accurate measurement than simply counting days of activity. The **Timing of Activity** will be recorded as the **number of hours after waking** when participants begin their physical activity session, providing precise data on when exercise is most likely to occur. Finally, **Energy Level Variations** will be measured on a continuous scale from 0 to 100, with participants rating their energy levels at different points of the day (e.g., morning, afternoon, evening), creating a continuous variable that can show fluctuations in energy throughout the day.

iii. **Sleep Quality:** For **Sleep Quality**, the variable **Hours of Sleep per Night** will be recorded as a continuous measure of the average hours of sleep participants get each night. This data will be collected using a **sleep tracker** or manual logs to ensure precision. **Consistency of Sleep Schedule** will be measured on a continuous scale, focusing on the **regularity** of a participant's sleep schedule. Instead of categorical terms like "Consistent" or "Inconsistent," participants will track the **number of days** they maintain a consistent sleep schedule throughout a week, giving a more nuanced view of their sleep habits.

iv. **Daily Productivity:** For **Daily Productivity**, the variable **Self-reported Productivity Rating** will be measured on a continuous scale from 0 to 100, where 0 indicates no productivity and 100 represents maximum productivity. Participants will self-assess their overall productivity at the end of each day. The **Productivity Consistency** will be recorded as the number of days a participant maintains **consistent productivity levels** throughout the week, rather than as a simple categorization of consistency. Finally, **Variation in Productivity** will be assessed by tracking productivity at different times of day (morning, afternoon, and evening) on a continuous scale from 0 to 100. Participants will rate their productivity levels throughout the day, providing detailed insight into how productivity fluctuates during various periods.

b. **What design is your study (experimental vs. correlational; if experimental, what factors are between subjects vs. within subjects)?**

The study will use a **correlational design** to examine the relationships between variables such as caffeine intake, physical activity, sleep quality, and daily productivity.

Justification:

i. It’s correlational as there is no active manipulation of variables, and does not involve random assignment. Also, the participants' daily routine is observed as it occurs naturally.

ii. Our study examines the relationship between daily habits (caffeine intake, physical activity, sleep quality) and productivity.

iii. We aim to identify trends and associations between these variables without manipulating them. Participants will self-report their habits and productivity over time, and we will analyze naturally occurring variations to see if they correlate.

c. **Given those answers, out of those variables which are your IV(s) and DV(s)?**

i. **Independent Variables:**

1. **Caffeine Intake:** For **Caffeine Intake**, the variables collected will include **Frequency of Consumption**, which will be measured on a continuous scale based on the **number of caffeinated beverages** consumed daily, such as cups of coffee, tea, or energy drinks. Participants will track the exact number of servings they consume each day. The **Timing of Consumption** will be recorded as the number of **hours after waking up** when caffeine is consumed. This replaces categorical time blocks like "Morning" or "Afternoon" and allows for more granular data. Participants will log the exact time (e.g., 1 hour after waking, 3 hours after waking) at which they consume caffeine.
2. **Physical Activity:** For **Physical Activity**, the key variables include **Frequency of Activity**, measured as the total **hours of exercise per week**, turning the data into a continuous variable. Participants will log the total time spent engaging in physical activity each week, allowing for a more accurate measurement than simply counting days of activity. The **Timing of Activity** will be recorded as the **number of hours after waking** when participants begin their physical activity session, providing precise data on when exercise is most likely to occur. Finally, **Energy Level Variations** will be measured on a continuous scale from 0 to 100, with participants rating their energy levels at different points of the day (e.g., morning, afternoon, evening), creating a continuous variable that can show fluctuations in energy throughout the day.
3. **Sleep Quality:** For **Sleep Quality**, the variable **Hours of Sleep per Night** will be recorded as a continuous measure of the average hours of sleep participants get each night. This data will be collected using a **sleep tracker** or manual logs to ensure precision. **Consistency of Sleep Schedule** will be measured on a continuous scale, focusing on the **regularity** of a participant's sleep schedule. Instead of categorical terms like "Consistent" or "Inconsistent," participants will track the **number of days** they maintain a consistent sleep schedule throughout a week, giving a more nuanced view of their sleep habits.

ii. **Dependent Variables:**

1. **Daily Productivity:** For **Daily Productivity**, the variable **Self-reported Productivity Rating** will be measured on a continuous scale from 0 to 100, where 0 indicates no productivity and 100 represents maximum productivity. Participants will self-assess their overall productivity at the end of each day. The **Productivity Consistency** will be recorded as the number of days a participant maintains **consistent productivity levels** throughout the week, rather than as a simple categorization of consistency. Finally, **Variation in Productivity** will be assessed by tracking productivity at different times of day (morning, afternoon, and evening) on a continuous scale from 0 to 100. Participants will rate their productivity levels throughout the day, providing detailed insight into how productivity

fluctuates during various periods.

d. **What are the operational definitions going to be for your IV(s) and DV(s)? (ie how are you going to measure or manipulate the variables)?**

Our study is **correlational**, and we will use self-reported data collected through surveys to measure each variable. The operational definitions of both **independent** and **dependent variables** are designed to capture precise, continuous data, providing a detailed view of participants' lifestyle habits and their impact on daily productivity.

Independent Variables (IVs):

1. **Caffeine Consumption**:
   To assess caffeine intake, participants will report the **number of caffeinated beverages** they consume per day, such as coffee, tea, or energy drinks. This will be measured as a continuous variable (e.g., 1, 2, or 3 cups per day). This provides a direct measure of the volume of caffeine consumed. Additionally, participants will record the **timing of caffeine consumption**, measured as the **number of hours after waking** that they consume caffeine (e.g., 1 hour after waking, 3 hours after waking). This allows us to examine the relationship between the timing of caffeine intake and its potential effects on productivity.
2. **Physical Activity**:
   For physical activity, we will measure **frequency** by asking participants to report the **total hours** of exercise they engage in per week (e.g., 3.5 hours per week). This gives a continuous measure of physical activity, capturing both light and vigorous exercise. We will also record the **timing of physical activity**, which will be measured as the **number of hours after waking** that participants begin their exercise sessions (e.g., 2 hours after waking, 5 hours after waking). This data will help us understand how the timing of physical activity relates to productivity levels throughout the day.
3. **Sleep Quality**:
   For sleep quality, participants will provide the **average number of hours** they sleep per night, measured as a continuous variable (e.g., 6.5 hours per night). This allows us to analyze sleep duration about daily productivity. In addition, we will assess **sleep schedule consistency** by asking participants to report how **regular** their sleep patterns are. Rather than using categories like "consistent" or "inconsistent," we will measure consistency on a scale (e.g., the number of days per week they maintain a consistent sleep schedule). This will provide a more accurate view of sleep patterns and their potential **effects on** productivity.

Dependent Variables (DVs):

1. **Daily Productivity**:
   For daily productivity, participants will rate their **overall productivity** each day on a continuous scale from 0 to 100, where 0 indicates no productivity and 100 represents maximum productivity. This allows for more precise measurement than traditional Likert

scales. We will also measure **productivity consistency** by asking participants how consistent their productivity is daily. This will be tracked as a continuous variable based on the number of days they feel their productivity remains consistent throughout the week (e.g., 0 to 7 days). Finally, to capture variations in productivity, participants will report their productivity at different times of the day (morning, afternoon, and evening), using a continuous scale from 0 to 100. This will allow us to see how productivity fluctuates during the day and how it relates to other lifestyle factors like caffeine consumption and physical activity.

e. **What is your population? How are you going to get participants from that population? How many are you planning to recruit for the study?**

i. The population for this study consists of university students who are likely to consume caffeine daily, engage in physical activity, and experience varying sleep habits. This demographic is ideal because they often manage their productivity in academic or work environments.

ii. Recruitment will target students from the University of Southern California (USC), where the study is being conducted. We will share the survey through Email, Student organization platforms, and word of mouth.

iii. A sample of 30-50 participants is a reasonable size for detecting patterns in correlational research while maintaining manageable data collection and analysis efforts for our study.

3. **Sketch out your plan for analysis:**

a. **State your research question(s), and discuss how it could be answered by analyzing the data that you listed in the previous question. That is, affirm for me that your research question is answerable using the data you will collect.**

i. **Research question:** How do caffeine intake, physical activity, and sleep quality relate to self-reported productivity levels among university students on a daily basis?

ii. We will answer this question by analyzing the correlations between our independent and dependent variables (productivity). Our collected survey data will measure specific aspects of each habit and productivity metric, allowing us to examine relationships between these variables. We can determine which factors are strongly associated with productivity outcomes by comparing the relationships between different habits and productivity measures.

iii. For **Data Collection Mapping**, the primary **dependent variables (DVs)** in the survey will include participants' **daily productivity rating**, which will be measured on a continuous scale from 0 to 100, where 0 indicates no productivity and 100 represents maximum productivity. This will provide a precise measure of productivity throughout the day. Additionally, **productivity consistency** will be assessed by asking participants how consistent their productivity is over a week, measured on a scale from 0 to 7 days, where 0 indicates no consistency and 7 indicates perfect consistency. Finally, **time-of-day productivity variations** will be captured by asking participants to rate their productivity at different times of the day (morning, afternoon, and evening) on a scale from 0 to 100. This will provide detailed insights into how productivity fluctuates throughout the day and whether it is **related to other** factors.

b. **Describe in your own words what kinds of analysis could be done with the data to answer each question. Be specific about what analysis -within null hypothesis significance testing- you would use and why.**

Since this is a correlational study with multiple variables, we will employ several statistical analyses within null hypothesis significance testing:

i. **Multiple Regression Analysis - Combined IV Test**

1. **Purpose:** To examine the combined effects of caffeine intake, physical activity, and sleep quality on productivity ratings.
2. **Hypothesis:**
   1. **Caffeine Intake:**
      a. Null Hypothesis (H0): There is no unique effect of caffeine intake (β1) controlling for Physical activity & Sleep quality.
      b. Alternative Hypothesis (H1): There is a unique effect of caffeine intake (β1) controlling for Physical activity & Sleep quality.
   2. **Physical Activity:**
      a. Null Hypothesis (H0): There is no unique effect of Physical Activity (β2) controlling for Caffeine Intake & Sleep quality.
      b. Alternative Hypothesis (H1): There is a unique effect of Physical Activity (β2) controlling Caffeine Intake & Sleep quality.
   3. **Sleep Quality:**
      a. Null Hypothesis (H0): There is no unique effect of Sleep Quality (β3) controlling for Caffeine Intake & Physical activity.
      b. Alternative Hypothesis (H1): Sleep Quality (β3) has a unique effect on controlling Caffeine Intake & Physical activity.
3. **Significance Level:** α = 0.05 (as per general trend)
   a. The **independent variables (IVs)** in this study will include **Caffeine Intake**, measured by the frequency of caffeinated beverage consumption, which reflects how often participants consume caffeine throughout the day. **Physical Activity** will be assessed by the frequency of physical exercise per week(0-7 scale), indicating how many days per week participants engage in exercise. Finally, **Sleep Quality** will be measured by two factors: the average number of hours of sleep participants get per night and the consistency of their sleep schedule, reflecting how regular and stable their sleep patterns are. These independent variables are expected to be related to productivity ratings, the primary focus of the study.
   b. This study's **dependent variable (DV)** will be the **Overall Productivity Rating**, which will be assessed through participant self-reports. They will rate their daily productivity on a scale, indicating how productive they feel throughout the day. This variable will help determine how factors such as caffeine intake, physical activity, and sleep quality are related to perceived productivity.

ii. **Pearson Correlation Analysis** - Detects Continuous DV Effects over Productivity (positive, negative, or neutral) for a given day.

1. **Purpose:** To determine the strength and direction of relationships between continuous variables.
2. **Correlation between Caffeine Intake and Productivity:-**
   The hypotheses for this study are as follows: The **null hypothesis (H_0)** suggests no correlation between **caffeine intake frequency** and **productivity ratings** ($r = 0$), meaning caffeine consumption does not affect productivity. The **alternative hypothesis (H_1)** proposes a significant correlation ($r \neq 0$), indicating that the frequency of caffeine intake is **related to** productivity ratings. In this analysis, the **independent variable (IV)** is **caffeine intake frequency**, and the **dependent variable (DV)** is the **overall productivity rating**.
3. **Correlation between Physical Activity and Productivity:**
   The hypotheses for the relationship between physical activity and productivity are as follows: The **null hypothesis (H_0)** states that there is no correlation between **physical activity frequency** and **productivity ratings** ($r = 0$), meaning exercise frequency does not affect productivity. The **alternative hypothesis (H_1)** suggests a significant correlation ($r \neq 0$), indicating that the frequency of physical activity related to productivity ratings. In this context, the **independent variable (IV)** is the **frequency of physical exercise per week**, and the **dependent variable (DV)** is the **overall productivity rating**.
4. **Correlation between Sleep Quality and Productivity:**
   The hypotheses for the relationship between sleep and productivity are as follows: The **null hypothesis (H_0)** posits no correlation between **hours of sleep** and **productivity ratings** ($r = 0$), meaning the amount of sleep a person gets does not impact their productivity. The **alternative hypothesis (H_1)** suggests a significant correlation ($r \neq 0$), indicating that the number of hours of sleep significantly affects productivity ratings. In this analysis, the **independent variable (IV)** is the **average sleep per night**, and the **dependent variable (DV)** is the **overall productivity rating**.

The analysis focuses on two key components:

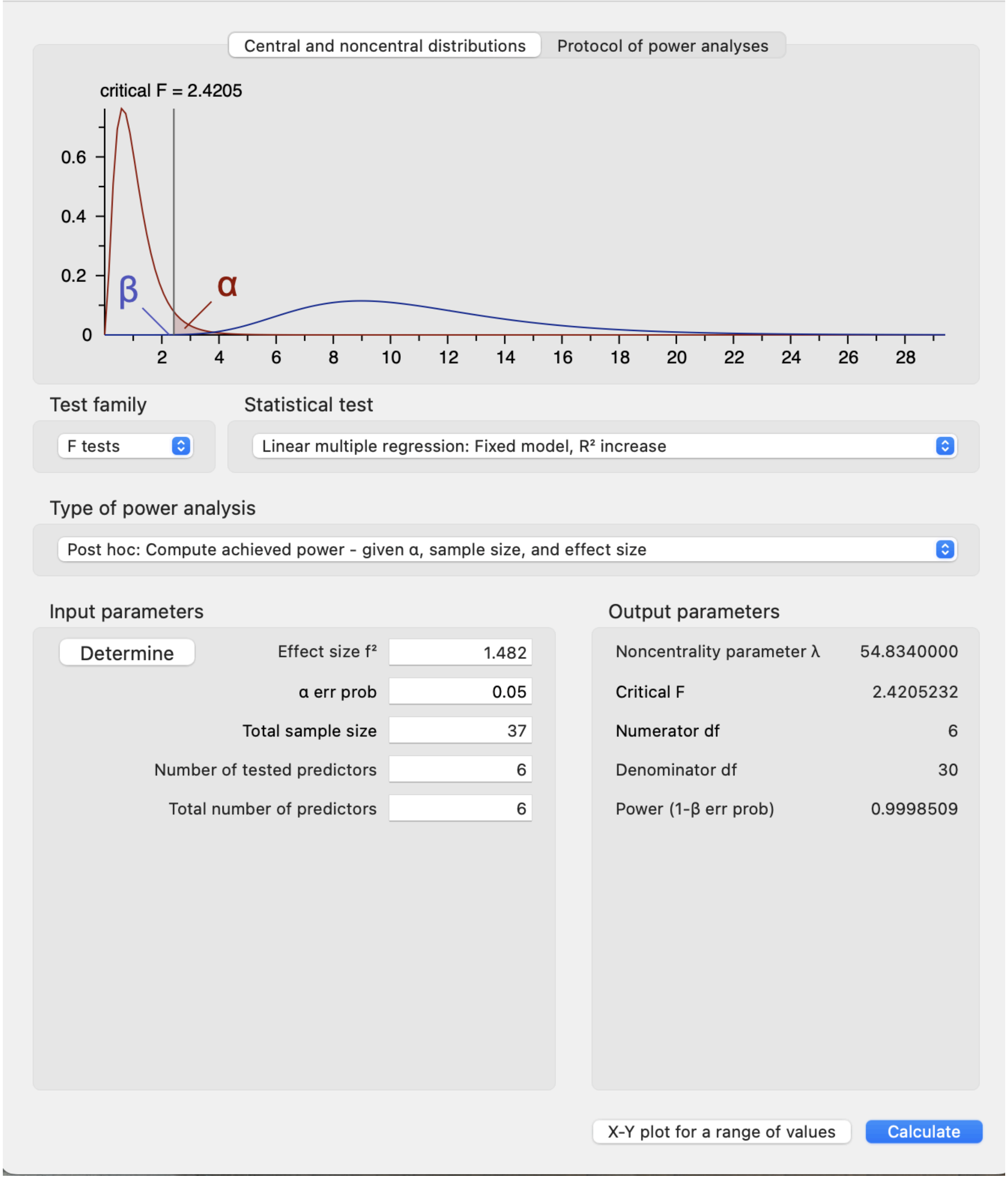
1. Overall predictive relationships through multiple regression
2. Individual correlations between continuous variables

This comprehensive analysis approach will allow us to:

- Identify which habits have the strongest relationships with productivity
- Understand how the timing of different activities relates to productivity patterns
- Determine the role of consistency in habits and its relationship with consistent productivity
- Account for both continuous and categorical variables in our dataset

By examining these relationships through the above-mentioned statistical approaches, we can provide robust insights into how caffeine intake, physical activity, and sleep quality correlate with productivity among university students.

Screenshot of G*power:


Central and noncentral distributions
Protocol of power analyses
critical F = 2.4205
0.6
0.4
0.2
0
β
α
2
4
6
8
10
12
14
16
18
20
22
24
26
28
Determine
X-Y plot for a range of values
Calculate

Test family: F tests

Statistical test: Linear multiple regression: Fixed model, R² increase

Type of power analysis: Post hoc: Compute achieved power - given α, sample size, and effect size

Input parameters

Effect size f²	1.482
α err prob	0.05
Total sample size	37
Number of tested predictors	6
Total number of predictors	6

Output parameters

Noncentrality parameter λ	54.8340000
Critical F	2.4205232
Numerator df	6
Denominator df	30
Power (1-β err prob)	0.9998509