**Homework 3: Analysis of Monthly Spending and Preferences**

**Due Feb 22 at midnight**

**Instructions:** Load the dataset GACTT\_RESULTS\_ANONYMIZED\_HW3.csv. Answer the following questions by carrying out the required analyses and providing interpretations. You should perform necessary data cleaning (missing data, reassigning values) before conducting analysis.

**Question 1: Modeling Monthly Spending (5pt)**

For this question, we are interested in understanding how demographic factors influence monthly spending.

1. **Simple Linear Regression**  
   (1) Choose a key demographic predictor that you believe might influence monthly spending. Fit a simple linear model to regress monthly spending on this predictor.
   * Interpret the coefficient of the predictor and discuss its significance.
   * What does the R-squared value tell you about the model’s fit?
2. **Building Nested Models**  
   (1) Sequentially add **at least two but up to four control variables** to your model to account for potential confounders (e.g., marital status, employment status, region). Explain why you selected these variables as controls.  
   (2) Compare the nested models using relevant metrics.  
   (3) Which model do you believe offers the best explanation of monthly spending? Why? Does controlling for additional variables change your interpretation compared to the simple model?
3. **Modeling Interaction Effects**  
   (1) Introduce an interaction term into the model using a moderator of your choice (e.g., interaction between income and age).  
   (2) Interpret all relevant coefficients, particularly the interaction term.
   * How does the interaction affect your conclusions about the relationship between the main predictor and monthly spending?
4. **Visualize your results using predicted values from the interaction model.**

**Question 2: Coffee Preference Analysis (5pt)**

In this question, we’ll analyze how demographic predictors influence preference for different coffee types, starting with a binary outcome.

1. **Binary Indicator for Coffee Preference**  
   (1) Based on the abc\_prefer variable, create a binary indicator for preference for **Coffee A** (1 = prefers Coffee A, 0 = prefers B or C).  
   (2) Select a demographic predictor (it can be the same as in Question 1 or a different one).
2. **OLS Modeling**  
   (1) Fit an **OLS model** to predict preference for Coffee A using the demographic predictor without any controls.  
   (2) Fit another OLS model with **two to four controls** and justify your selection.  
   (3) Compare the results of the two models. Do the added controls change the interpretation of your main predictor?
3. **Binary Logistic Regression**  
   (1) Fit a **logistic regression** model to predict the preference for Coffee A (with the same set of controls used in the OLS models).  
   (2) Compare the results of logistic regression with those from OLS.
   * How does the interpretation of coefficients differ between OLS and logistic regression?
   * Which model do you believe is more appropriate for this type of outcome? Why?

**Question 3: Bonus Task (2pt)**

In Question 2, we focused on **Coffee A vs. B or C** using binary outcomes. Now, let’s expand this to analyze preferences for **A vs. B vs. C** using a **multinomial logit model**. This will help us understand the factors that influence preferences across all three coffee types.

1. **Self-Learning Exercise:** Research and implement the multinomial logit model in R. In your report, include:
   * Statistical notation for the multinomial logit model and underlying assumptions.
   * How to interpret coefficients in a multinomial logit model.
   * The empirical analysis (with and without controls).
   * Visualizations of predicted probabilities.
   * Cite any references or resources you used.
2. **Fit Models and Interpret Results:**  
   (1) Fit multinomial logit models using the same demographic predictor from Question 2, first without controls and then with some controls (different number of covariates).  
   (2) Interpret your findings. How do the results compare to the binary logistic model in Question 2?