**Homework 4: Mixed Effect Models**

**Due: Mar 8th at Midnight**

**Instructions:** Load the dataset **GACTT\_RESULTS\_ANONYMIZED\_HW4.xlsx**. This assignment focuses on modeling and comparing different regression approaches. The outcome variable of interest is **most\_willing\_for\_cup**, and the main predictor of interest is **hh\_income**. Conduct the analyses outlined below and provide interpretations of your results.

**Question 1: Normal OLS model (Benchmark) (2 points)**

In this section, you will establish a benchmark model by fitting a **Ordinary Least Squares (OLS) regression model** that treats all observations as independent.

1. Fit an **OLS model** predicting most\_willing\_for\_cup using hh\_income.
2. Interpret the coefficient for hh\_income. What does it suggest about the relationship between income and willingness?
3. Plot the relationship for the two variables, showing the trend for each age groups and the average trend across all age groups. What do you observe? Do you think OLS is an appropriate model to capture the relationship?

**Question 2: Fixed Effects Models (2 points)**

Fit two **fixed effects** models and compare them to the pooled model.

1. **One-way fixed effects model:** Include **age group (age\_cat)** as a fixed effect.
2. **Two-way fixed effects model:** Include **age group (age\_cat)** and **state** as fixed effects.
3. Compare the results from both fixed effects models to the **pooled OLS model** from Question 1:
   * How does the coefficient and your interpretation for hh\_income change?
   * What do these changes indicate about the role of age group and state in influencing most\_willing\_for\_cup?

**Question 3: Mixed Effects Models (3 points)**

In this section, you will fit **mixed effects models** to account for hierarchical data structures.

1. **Random intercept model:** Include a **random intercept** for **age group (age\_cat)**.
2. **Random slope model:** Include a **random slope** for **hh\_income** (varying across age groups) and a **fixed intercept**.
3. Compare the **random intercept model (3.1)** with the **one-way fixed effects model (2.1)**:
   * How do the interpretations of parameters differ?
   * Does one model provide a better explanation than the other? Why?
4. Compare the **random intercept model (3.1) and the random slope model (3.2)**:
   * What changes when allowing hh\_income to have a random slope across age groups?
   * How does this affect the interpretation of hh\_income?

**Question 4: More Mixed Effects Model (3 points)**

Now, fit a **random effects Cmodel** that includes both:

* A **random intercept** for **age group (age\_cat)**
* A **random slope** for **hh\_income** (across age groups)

1. Compare this model to the two **mixed effects models** from Question 3 and explain how these random effects model differ
2. Among all models you have estimated (pooled OLS, fixed effects, mixed effects models):
   * Which model provides the best fit?
   * Justify your selection based on appropriate criteria (e.g., AIC, BIC, R-squared, interpretability, theoretical considerations, and you may talk about several).

**5. Bonus task (up to 3 points):**

Participate in the [Kaggle competition](https://www.kaggle.com/t/c7c554eb2d35473181197ea084238b94): you need to use the training data set to fit the model and make predictions on the test set.

You can work on your device or use Kaggle notebooks. However, you'll need to submit the script of your final model to Kaggle (and note a separate deadline of March 12th). Check the separate Assignment page for bonus task for more information.

You can use any methods we covered in this class (linear, generalized linear, linear mixed models, dimension reduction methods, and variable selection methods), but not any unsupervised methods (i.e., decision trees, neural networks, SVMs, etc.).