Problem Set 3: Multiple Linear Regression

Part 1: Paper Analysis

Some studies have found that when reading academic papers, more experienced researchers tend to focus on the *methods*and *results* sections, while undergraduates and junior researchers focus more on the paper's *narrative*. Evaluating research empirically is a skill that takes time to develop, and this exercise is designed to help you cultivate that skill.

Evaluate Fearon and Laitin's (2003) APSR paper, *"Ethnicity, Insurgency, and Civil War,"* and answer the following questions.
I encourage you to complete this part of the exercise with classmates.

1. **Research Goals**
   Is the goal of the study causal inference, description, prediction, or something else? Have the authors clearly stated their goals? Describe any strengths or weaknesses in how the authors articulate their research objectives.

2. **Estimands**
   Have the authors sufficiently defined their theoretical and empirical estimands? Discuss what these estimands are and explain how the authors could clarify them if necessary.

3. **Identification Strategy**
   The way you connect your theoretical estimand to your empirical estimand is known as *identification*—in other words, what does the research do to ensure that the empirical estimand is a good measure of the theoretical estimand? Describe the authors' identification strategy.

4. **Assessment of Findings**
   Provide an overall assessment of the paper and its conclusions. Does the identification strategy support the authors' claims? For example, could the regression coefficients be credibly interpreted as causal effects if causal inference is the goal? Does the model adequately represent the real-world data-generating process? Does the data credibly measure the phenomena being studied?

5. **Broader Contribution**
   Despite any weaknesses, can this research still inform our understanding of the world? If so, how?

Part 2: Data Analysis

A common way to measure how people feel about other groups is through a survey instrument known as the *feeling thermometer*. Respondents are asked to rate how "warm" they feel toward a group on a scale from 0 to 100, with 100 representing very warm and positive feelings, and 0 representing very cold and negative feelings.

For this exercise, you will analyze a dataset that contains feeling thermometers for several groups, along with demographic information about the survey respondents.

1. Load the thermometers.csv data from the data folder on the github repo. Use the birth_year variable to create a new age variable (Note: This survey was taken in 2017).
2. Pick one of the feeling thermometers and one of the categorical demographic. variables (sex, race, party_id, or educ). Describe the spread and central tendency of the feeling thermometer both for all observations, and for each category in the demographic variable you chose. Use histograms or density plots to visualize the distribution.
3. Fit a regression model to estimate the conditional mean of the feeling thermometer for each category in the demographic variable you chose.
4. Create a new dataframe that only contains rows for Democrats and Republicans. Create a new binary variable for party_id.
5. Use multiple linear regression to build a model that predicts your binary party_id variable. Use any combination of variables you like, but you should include at least one feeling thermometer and one interaction term. Justify your model.
6. The coefficients in your model represent the change in what?
7. Select one of the feeling thermometers in your model and plot how your predicted values change as the feeling thermometer changes. Interpret your results. Can this reasonably be interpreted as a causal effect?