Preparing Analysis Datasets

Overview

You will work through a few typical data construction tasks in this hands-on session. The typical tasks in this stage include:

1. Planning construction outputs
2. Prepare data for construction:
   - Dealing with outliers
   - Standardizing units
3. Creating indicators
4. Save data

We won't do all of this, but a few of them to show you the workflow. This exercise will use the clean dataset you created during the data processing tutorial.

Exercise 1: Standardize units for land area and currencies

During data acquisition, land area and consumption values are often reported in different units. These need to be converted to standardized units for analysis.

- **Standardize land area to acres:**
    - If the land area is reported in acres, keep the value as is.
    - If the land area is in hectares, convert the value to acres using the conversion factor (2.47) (create a variable 'area_acre').
- **Standardize currency to USD:**
    - Set a variable for the USD conversion rate at the beginning of the script.
    - Use this rate to convert all consumption values (both 'food_cons' and 'nonfood_cons') to USD (create variables 'food_cons_usd' and 'nonfood_cons_usd').
- If there is a change, simply update the variables, and the code will automatically reflect the new values.

Exercise 2: Deal with Outliers

- During the data processing stage, we identified outliers in the area and food variables. Now after standardization: `area_acre`, `food_cons_usd`, and `nonfood_cons_usd`.
- **Winsorize outliers**: Winsorization involves replacing extreme values with those at the 5th percentile (for lower values) and 95th percentile (for upper values). This helps to retain high-value observations without allowing them to disproportionately influence the results.

- After creating the indicators from clean datasets, merge the treatment dataset to prepare the dataset for analysis
- Save the resulting dataframe. You will be using this in the data analysis session.

THANK YOU