Research Assistant Data Assessment

5/5/23

Introduction

This take home coding/analytics assessment is designed to assess the candidate's ability to wrangle and data generated in the ORL Basic Income RCT. The assessment uses dummied data that matches the form of data collected from our midline survey.

Please do not spend more than 2 hours on this assessment.

If you have questions, where you feel you cannot move forward with the assessment without clarification, please do not hesitate to reach out to patrick@openresearchlab.org. Otherwise, please document any questions you have or any assumptions/analytical decisions you make to complete the assessment in your code.

Instructions

- Create an R script to complete the task described below.
    - Please annotate your code and document your approach within the script (including any decision points or areas where you would have liked to have more information)
- Create a output data file that includes only id and new cleaned variables generated

Task: Create cleaned final monthly expense variables

For all raw expense variable (e4:e37) variables included in midline_responses_raw.Rds,

1. Generate a new numeric variable for each that recodes the expenses to **monthly**
    - Based on expense unit variables (e4a:e37)
    - Ensure new variable name differentiates from raw variable
2. Clean new variables
    - Please recode "don't know" or "refused" values to missing
    - Replace missing values with the *within-treatment arm* sample mean
    - Winsorize at the 99 percentile to address outliers

Data Notes

- Questions only asked to specific participants
    - E17 & E18 - IF num_childu5>0

- E19 & E22 - IF num_child5_17>0

Extra Credit

- Write a general purpose function for data validation
  - ex. `validate_replace_missing(df, raw_var, clean_var)` that confirms missing values in raw_var have been replaced with the *within-treatment arm* mean in clean_var

Output

1. R script including all annotated cleaning code
2. Final dataset generated in R script

Data Provided

File	Description
Midline_responses_raw.Rds	Raw output of midline survey questionnaire (Subset of selected variables for this assessment)
treatment_assignment.Rds	Program treatment indicator
data_dictionary.csv	Data dictionary for all variables in midline_raw and treatment_assignment
Midline Instrument.pdf	Includes survey question and range of response values for all midline variables included in midline_responses_raw.Rds