**Research Replicability and Workflow Management (RRWM)**

**Replication report**

* ***Part 1: Comments by Mahjoube on Camila’s project program***

Program with Associated Do-Files

1. Clean and generate the dependent and key independent variables I. Dependent variable is the self-rated health variable (srh\_110) recoded into a new binary variable called good\_health. Categorize the respondents' self-assessed health into two categories: "good health" (combining those who rate their health as 1excellent, 2very good, or 3good) and "not good health" (combining those who rate their health as 4fair or 5poor). Missing values or outliers are left as missing. II. Drop missing data for good\_health. III. The key independent variable is household size variable (hsdsizec) recoded into a new binary variable called living\_alone. Categorize respondents based on whether they live alone or not. Respondents who have a household size of 1 are categorized as "living alone" (coded as 1), while those with a household size between 2 and 6 are categorized as "not living alone" (coded as 0). Missing values or outliers are left as missing. IV. Drop the missing data for living\_alone.
2. Clean and generate control variables I. The demographic control variables include gender, age group, and visible minority status. Gender is recoded into a binary variable (female), based on the original variable (sex), where respondents who identify as female are coded as 1, and males are coded as 0. Age is categorized into four age groups (agegrp), based on the original variable (agegr10): 0 for respondents aged 54 and below (who will later be dropped), 1 for those aged 55–64, 2 for those aged 65–74, and 3 for those aged 75 and older. Respondents under the age of 55 are excluded from the analysis. Visible minority status is recoded into a binary variable (vis\_minority), based on the original variable (vismin), where respondents who identify as visible minorities are coded as 1, and those who do not are coded as 0. II. Drop the missing values for the demographic control variables (only visible minority has missing) and for agegrp 0 because we are only interested in observing older adults. III. The socioeconomic control variables include educational attainment and income level. Educational attainment is recoded into a categorical variable (edu\_level), based on the original variable (ehg3\_01b), with four categories: 1 for less than high school, 2 for high school or equivalent, 3 for trade, college, or non-university certificate, and 4 for a university degree or higher. Income level is recoded into a categorical variable (income), based on the original variable (famincg2), with three categories: 1 for low income (less than $50,000), 2 for middle income ($50,000 to $99,999), and 3 for high income ($100,000 or more). IV. Drop the missing values for the socioeconomic control variables (only education level has missing).
3. Save the cleaned dataset. I. Keep only created variables. II. Keep person survey weights (wght\_per) III. Save data “gss\_clean”
4. Generate analysis do-file. I. Use cleaned data file “gss\_clean” II. Install estout (for comparison tables) III. Set survey using wght\_per
5. Create a weighted summary table with all variables (t) I. Generate a list of summary statistics for key variables: living arrangement (living\_alone), health status (good\_health), gender (female), age group (agegrp), visible minority status (vis\_minority), educational attainment (edu\_level), and income level (income). II. Apply the survey weights using the variable wght\_per.
6. Create logistic regression models (t) I. Perform a bivariate logistic regression to analyze the relationship between living alone and good health, using good\_health as the outcome variable and living\_alone as the key independent variable. Store the model for later comparison table. II. Perform additional bivariate logistic regressions to explore the relationship between good\_health as the outcome variable and living\_alone as the key independent variable and control for demographic variables such as gender (female), age group (agegrp), and visible minority status (vis\_minority). This model is also stored for later comparison table. III. Perform final multivariate logistic regression by adding socioeconomic status (SES) variables, including educational attainment (edu\_level) and income level (income), along with the demographic predictors. This multivariate model assesses the combined effect of living arrangements, demographics, and SES on health outcomes. This model is stored for later comparison table. IV. Generate a comparison table for both the bivariate and multivariate models (the three store models from above), allowing for comparison of the effects of each predictor on good health. The results are presented in terms of odds ratios for each model

* ***Part 2: Report and Discussion (Mahjoube)***

***Poll: Were you able to reproduce the work based only on the program (not the code)?***

* **Yes, entirely**
* **Yes, with only a few errors**
* **I got something that looks similar**
* **No, my numbers are off**
* **Program doesn’t even make a table that looks like what I was sent**

Yes, with a few minor errors. I was able to reproduce the project without any issues. However, there is no result table or graph in her git repository to compare our results.

1. **Where did it go off the rails? Was there one error that compounded everything?**

In most cases, divergence comes from the different variables we choose and may result in different outcomes; if we could have a table of variables we want to consider for each step, it would be easier.

1. **What pieces of the program left you guessing as to what to do?**

Because the formats and exact outcome of the GSS clean were not clearly indicated, the transition between recoding and cleaning and analysis requires a bit of thought.

1. **Any comments or reflections about the reproducibility exercise? For instance, has it changed your mind about the benefits (or costs) of open science? For those in fields where replication materials are not readily available, could an emphasis on reproducibility disadvantage researchers who use qualitative methods or restricted (e.g., administrative) data?**

Although I know the main purpose of this exercise is kind of for broader scope like when a journal or a reviewer asks for this kind of file with clear steps, I was always thinking about the joint project I'm working on with my peers and how it makes communicating with the research group easier, especially since we have different backgrounds.