A Multilevel Regression and Post-Stratification (MRP) Model

The methodological framework for the analysis here is borrowed from Leemann and Wasserfallen (2017). The model is based on one binary dependent variable, three weighting variables (one of those is for geographical units). Three weighting variables are planned as follows: gender, age and var_geo. We will benefit those weighting variables as random effects.

The model will take two datasets (one user/individual level and one administrational population data) as inputs and extracts a list that includes prediction for each geographical unit, and a number that shows aggregate average.

- Which R packages does the model require?
    - *foreign, lme4, arm, extrafont, readxl, dplyr*

- What does the model include, and how should the inputs be structured?

1. Parameters:

    a. The gender variable is considered as having two subcategories (as female=1). The numbers of subcategories for the other three variables will be calculated by the model. For clarification, let us briefly explain what they stand for.
        i. *N_age*: the number of categories under *age*
        ii. *N_geo*: the number of geographical units
    b. Accordingly the model will calculate two more parameters:
        i. *N_cat = 2 x N_age* (mathematical multiplication)
        ii. *N_total = 2 x N_age x N_geo* (mathematical multiplication)

2. Datasets:

    a. User/individual level dataset, might be gained from social media (*user_data*)
        i. Includes 5 variables
            1. One user identifier: *user_id,*
            2. One dependent variable: *dep_var,*
            3. Two weighting variables, might be demographics: *gender (0=male, 1=female), age*
            4. One geographical identifier: *var_geo,*
        ii. The matrix size equals to: 5 x Number of users plus one (with a header row)

    b. Administrational population data (*pop_data*), which depends on geographical distribution of the two weighting variables: *gender, age*.
        i. This *pop_data* includes geographical units (var_geo) as a first row, and hence, the column number equals to the number of geographical units (*N_geo).* Note that the header row only consists of geographical codes, and there is no other header.
        ii. Then, each row after the first one represents every combination of variable categories (which means *N_cat* number of additional rows)
        iii. Therefore, the matrix size of the dataset should be equal to: (*N_cat*+1) x *N_geo*

iv. Each cell stands for the number of people that belongs to the specific combination of categories living in that geographical unit.
v. The order of combinations should be listed in the same order with the variable names, as in Table 1. Note that each row in this table will be the unwritten row identifier in the *pop_data.*
vi. Also, the order of geographical units from left to right (columns) needs to be ordered from 1 to *N_geo.*

Table 1: The order of combinations

gender	age
0	1
1	1
0	2
1	2
0	3
1	3
...	...

References:

Leemann, L., & Wasserfallen, F. (2017). Extending the use and prediction precision of subnational public opinion estimation. American journal of political science, 61(4), 1003-1022.