A Dynamic Multilevel Regression and Post-Stratification (MRP) Model

Using the data extracted from online networks for election prediction is often criticized as having a solid demographic bias, referring to a well-known phenomenon that participation in online networks is strongly affected by users' age, sex, education level, race, income level, etc. To deal with this demographic bias, it is suggested to use the multilevel regression and post-stratification method (MRP), which is based on an adjustment of every possible combination of characteristics according to their actual representations in the population. The method has been developed to predict national-level data by non-representative polls or data for smaller areas from representative surveys (called small area estimation).

The methodological framework for the analysis here is borrowed from Leemann and Wasserfallen (2020). The model is based on one binary dependent variable, four weighting variables (one of those is for geographical units), and one additional variable for a geographical characteristic. Four weighting variables are planned as follows: gender, age, var3 (which is any other variable like education, religiosity, etc.), var_geo. We will benefit those weighting variables as random effects and employ the variable for a geographical characteristic (var_context) to hold for the fixed effect in the MRP model.

The model will take two datasets (as a user/individual level and an administrational population data) and three parameters (as the number of subcategories of the variables) 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*

- 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 the inputs of the model.
        i. *N_age*: the number of categories under *age*
        ii. *N_3*: the number of categories under *var3*
        iii. *N_geo*: the number of geographical units

    b. Then accordingly the model will calculate two more parameters:
        i. *N_cat = 2 x N_age x N_3* (mathematical multiplication)
        ii. *N_total = 2 x N_age x N_3 x N_geo* (mathematical multiplication)

2. Datasets:

    a. User/individual level dataset, might be gained from social media (*user_data*)
        i. Includes 7 variables
            1. One user identifier: *user_id,*
            2. One dependent variable: *dep_var,*
            3. Three weighting variables, might be demographics: *gender, age, var3,*

4. One geographical identifier: *var_geo,*
5. One geographical characteristic: *var_context,*

ii. The matrix size equals to: 7 x Number of users

Note: This *var_context* variable is not coming from the original individual level data (which is probably extracted from social media). Rather it will be taken from another data source, but is needed to be attached to the *user_data* via geographical identifier (*var_geo*) before this analysis.

b. Administrational population data (*pop_data*), which depends on geographical distribution of the three weighting variables: *gender, age, var3*.

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)*

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	var3
0	1	1
1	1	1
0	2	1
1	2	1
0	3	1
1	3	1
0	1	2
1	1	2
0	2	2
…	…	…

Leemann, L., & Wasserfallen, F. (2020). Measuring Attitudes–Multilevel Modeling with Post-Stratification (MrP). The SAGE Handbook of Research Methods in Political Science and International Relations, 371-384.