Multilevel Regression and Poststratification (mister P or MRP)

M. Fuat Kına

Why we need MRP?

- Sampling bias
- Social desirability bias
- Sub-units in the population
    - Small area estimation
    - Hard-to-reach groups

- Traditional version of MRP contains only the P

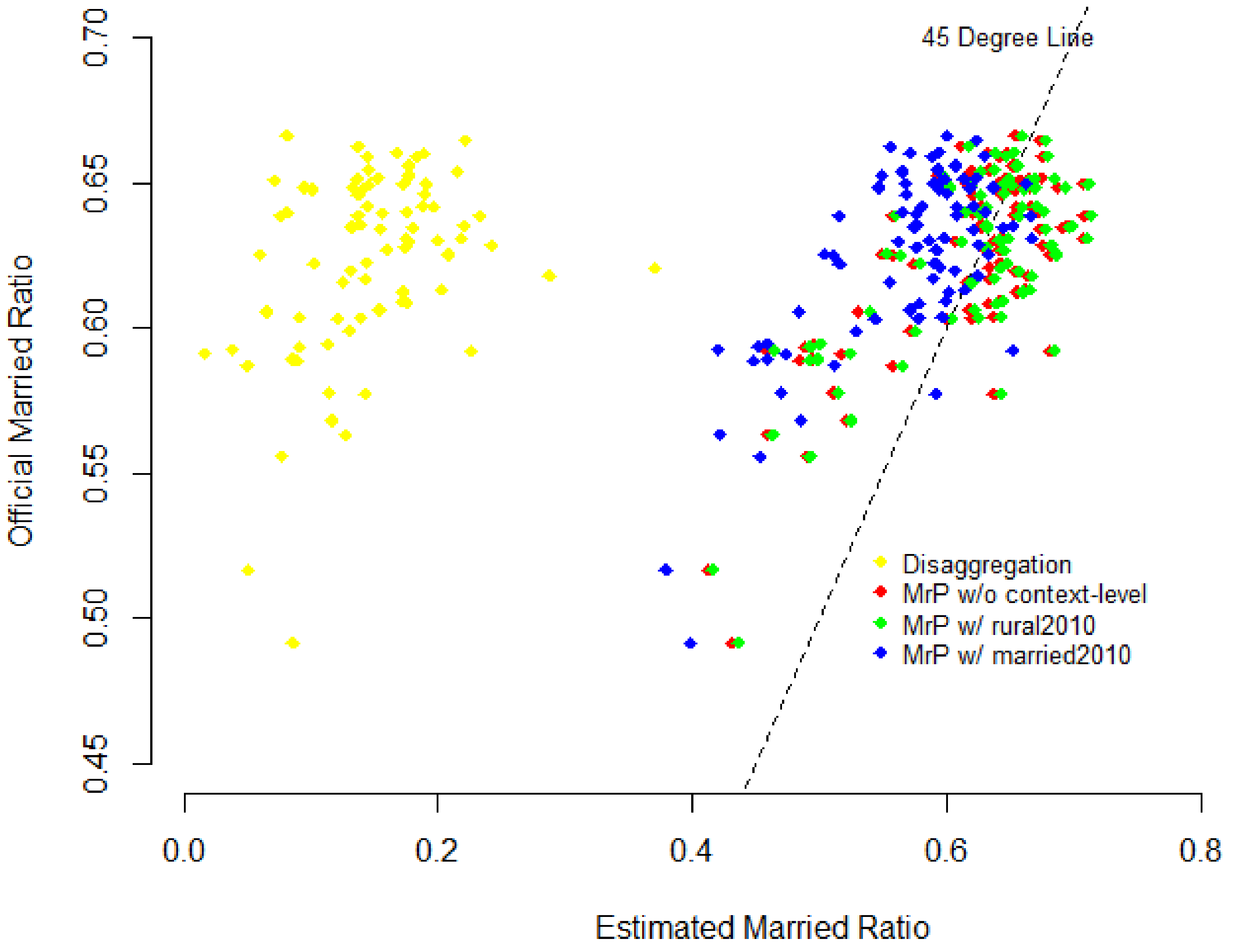
Post-stratification to multi-level regression

- Post-stratification is mathematically more complicated version of weighting
- Let’s assume
    - Our aim is to understand the average income for population, and we only have a nonrepresentative sample.
    - However, we know much information about individuals for both the sample and the population, except for income. We intend to use these information as weighting variables.
    - We have K number of weighting variables (e.g., age, gender, education, class, etc.).
    - They have N1, N2, N3, N4 ,…, Nk numbers of subcategories (young-middle-old, male-female, etc.),
    - which means the number of weighting cells: N1*N2*N3*N4*…*Nk.
    - Then we calculate the mean of Y for each cell.
    - And multiply these numbers with the actual weights of cells in the targeted population.
    - The summation of these weighted means will present the post-stratified Y.

Post-stratification to multi-level regression

- However, if we have too many cells, it is hard to find enough observations for each. Remember we have a biased data; we probably do not have any observations for some.
- Multilevel regression works to predict random effects for each factor, and we use the effect of each subcategory.
  - Because multilevel models contain a mix of fixed effects and random effects, they are sometimes known as mixed-effects models.
  - Different geographical units (nested levels) might have autonomies.
  - Generalizability to a wider population
- Bayesian multilevel regression works to increase the performance.

Marital Status of Twitter users: gender, age, location, married



What if we are interested in understanding city level sentiments?

Park, David K., Andrew Gelman, and Joseph Bafumi. (2004). “*Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls*.”

- Predictors: gender, race, age, education, state.
- Outcome variable: Votes of George Bush
- Bayesian logistic regression:

$$
\begin{aligned}
y^{pred} = logit^{-1}(\beta^{0} &+ \beta^{female}\cdot female_{j} + \beta^{black}\cdot black_{j} \\
&+ \beta^{female.black}\cdot female_{j}\cdot black_{j} + \beta^{age}_{age(j)} \\
&+ \beta^{edu}_{edu(j)} + \beta^{age.edu}_{age(j),edu(j)} + \beta^{state}_{state(j)})
\end{aligned}
$$

- Total number of categories: 3264

A python pymc3 trial for MRP analysis:

https://austinrochford.com/posts/2017-07-09-mrpymc3.html