Applied Multilevel Regression Modeling

Day 9: Multilevel Regression with Post-stratification

Constantin Manuel Bosancianu

WZB Berlin Social Science Center
Institutions and Political Inequality
bosancianu@icloud.com

August 8, 2019

Estimating sub-national public opinion

A few reasons for pursuing this:

- ✓ we can test theories about sub-national variation
- ✓ crudely put: it boosts sample size
- ✓ practically, very useful information for NGOs, government agencies, parties, or campaign organizations

Very challenging to do in practice, as it would require large amounts of data.

Keep in mind, that the accuracy of a sample is most often a function of sample size, not of sample size relative to population size (Freedman, Pisani, & Purves, 2007, p. 367).

Alternatives

Past attempts were made, but with varying degree of success.

Disaggregation: collect multiple polls using the same question over a period of time, and split up respondents based on sub-national unit.

Problems of insufficient sample size for small states, and inability to tackle most questions of interest remain.

Impossible to go at a lower level of aggregation due to varying ways of recording education, income etc.

MRP

The state-of-the-art (1)

Individual responses are modeled as (1) nested within states and, (2) nested within socio-demographic subgroups.

With a sufficient number of random effects, we can generate predictions that vary between states and between sub-groups.

Advantage: groups from small states can borrow power from similar groups in larger states $\Rightarrow$ we can get estimates even for very small groups!

The state-of-the-art (2)

In a second stage, we adjust the estimates by weighting them with the proportion of the subgroup in the population of the state (for constructing aggregates).

Has been shown to be superior to alternatives, such as disaggregation, or state-by-state analyses.

Has been show to yield workable results at state level even with samples of 1,500 voters, and at CD level with about 2,500 voters.

The model

Assume 3 groups, but this can easily extend to 6–7: income (j_1), ethnic group (j_2), and state (j_3).

The analysis is carried out at the level of groups, not individual-level; predictors are the characteristics of the groups.

$$log(\frac{\theta_j}{1-\theta_j}) = \alpha^0 + \alpha^1_{j1} + \alpha^2_{j2} + \alpha^3_{j3} + \alpha^{1,2}_{j_1,j_2} + \alpha^{2,3}_{j_2,j_3} + \alpha^{1,3}_{j_1,j_3} + \alpha^{1,2,3}_{j_1,j_2,j_3} \tag{1}$$

To this, you begin to add predictors at the levels of the 3 groups: income, ethnic group, and state.

The predictors

TABLE 1 Variables in the `lmer()` Model, Along with Analogous Terms from the Statistical Model

lmer() Variable	Description	Type	Number of Groups	Coefficient in Statistical Model
y	Vote choice (1 = McCain, 0 = Obama)	Output variable	–	–
z.incstt	State-level income	Linear predictor	–	Part of $\beta^1, \beta^3, \beta^4$
z.repprv	State-level Republican vote share from previous election	Linear predictor	–	Part of $\beta^1, \beta^3, \beta^4$
z.inc	Income (included as a linear predictor)	Linear predictor/ Varying slope	–	Part of $\beta^2, \beta^3, \beta^4$
inc	Income	Varying intercept	5	α^1
eth	Ethnicity	Varying intercept	4	α^2
stt	State	Varying intercept	51	α^3
reg	Region of the country	Varying intercept	5	α^4
inc.eth	Income × ethnicity interaction	Varying intercept	$4 \times 5 = 20$	$\alpha^{1,2}$
inc.stt	Income × state interaction	Varying intercept	$4 \times 51 = 204$	$\alpha^{1,3}$
inc.reg	Income × region interaction	Varying intercept	$5 \times 5 = 25$	$\alpha^{1,4}$
eth.stt	Ethnicity × state interaction	Varying intercept	$5 \times 51 = 255$	$\alpha^{2,3}$
eth.reg	Ethnicity × region interaction	Varying intercept	$4 \times 5 = 20$	$\alpha^{2,4}$

Post-stratification and other adjustments

for John McCain for president. In any case, label N_j as the relevant population in cell j, and suppose we are interested in θ_S: the average of θ_j's within some set J_S of cells. The poststratified estimate is simply

$$\theta_S = \sum_{j \in J_S} N_j \theta_j \Big/ \sum_{j \in J_S} N_j. \qquad (1)$$

voters for each state and perform a simple adjustment so that overall turnout matches the state totals, as follows. Let ξ_s indicate the number of voters for each state $s = 1, \ldots, 51$, and let S denote the set of cells such that j is in state s. We derive the adjusted turnout estimate θ_j^* for each cell $j \in S$ as follows:

$$\delta_s = \min\left(\text{abs}\left(\xi_s - \sum_S \left(N_j \,\text{logit}^{-1}\left(\text{logit}(\theta_j) + \delta\right)\right)\right)\right) \qquad (7)$$

$$\theta_j^* = \theta_j + \delta_s \ \forall \ j \in S, \qquad (8)$$

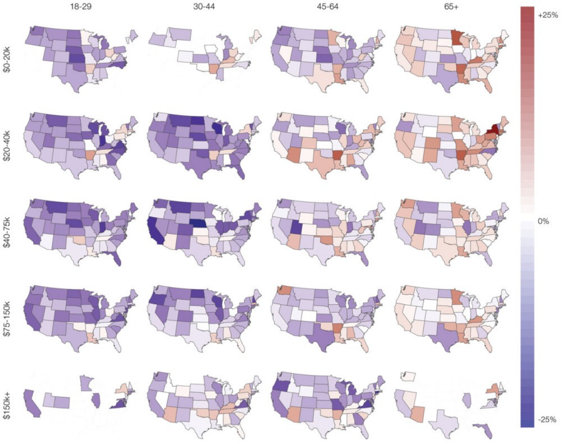
where abs() is the absolute value function and min() is a function that finds the δ that minimizes the expression. This process simply applies a constant logistic adjustment δ_s to each cell in state s to make sure that the total number of estimated voters is correct. We assume here that

The post-stratification part is needed so as to be able to do accurate aggregations.

Where available, a correction can also be done so as to adjust estimates to actual aggregate totals.

Presenting results (1)

Figure 4 McCain 2008 Minus Bush 2004 among Non-Hispanic Whites



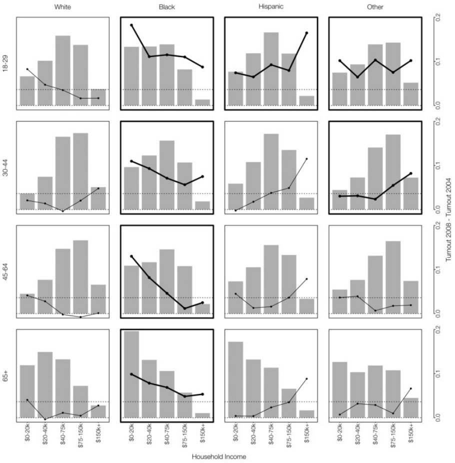
Note: State-by-state shift toward McCain (red) or Obama (blue) among white voters broken down by income and age. Red = McCain better than Bush; Blue = McCain worse than Bush. Only groups with >1% of state voters shown. Although almost every state moved

Though it's possible to break down findings very finely, we're still limited in the case of very small groups (estimates are very erratic).

The plot is made possible by running MRP on both 2004 and 2008 data, and producing differences.

Presenting results (2)

Figure 5 Turnout Swing Mainly Isolated to African Americans and Young Minorities



Note: Turnout change shown in line graphs; population distribution shown as bar graphs. Turnout changes in the 2008 election were not consistent across demographic subgroups. African Americans and young minorities increased turnout almost uniformly, but white voters did not. Groups with a total turnout change over 5% are highlighted with a thick box and trend line.

Similar presentation, but without focusing on geography.

MRSP

Data requirements (1)

Two challenges related to data availability.

With every grouping we add, we need to find census data which contains the respective variable which matches the categorization of the one in our survey.

We can include only socio-demographic predictors, even though for phenomena like turnout or party preference, what matter a lot are party ID, political interest etc.

Data requirements (2)

TABLE 1 Census Data Requirement Example of Classic MrP

Gender \ Education	No High School	High School	College	Postgraduate	Total
Men	$\mathbf{N_{11}}$	$\mathbf{N_{12}}$	$\mathbf{N_{13}}$	$\mathbf{N_{14}}$	$N_{1\cdot}$
Women	$\mathbf{N_{21}}$	$\mathbf{N_{22}}$	$\mathbf{N_{23}}$	$\mathbf{N_{24}}$	$N_{2\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	$N_{\cdot 4}$	N

"True" joint distributions are needed for each subnational unit.

MRSP

MR-Synthetic-P works with "synthetic" joint distributions, derived from information on marginal distributions.

TABLE 2 Example of True and Simple Synthetic Joint Distributions

v1 \ v2	i=1	i=2	
j=1	60%	0%	60%
j=2	0%	40%	40%
	60%	40%	100%

(a) True Joint Distribution

v1 \ v2	i=1	i=2	
j=1	36%	24%	60%
j=2	24%	16%	40%
	60%	40%	100%

(b) Simple Synthetic Joint Distribution

Simple version: joint distributions are computed as products of marginal distributions.

Problematic when the marginal distributions refer to variables that are correlated, but the authors find the bias is not major. *Adjusted version*: does away with the assumption of independence.

Thank you for the kind attention!

References

Freedman, D. A., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). New York: W. W. Norton & Co.