Approaches to Small Area Estimation:

Multilevel Regression with Poststratification (MRP)

Julian Perry

Stat 160

Spring 2024

The general problem

- Define population U, parameter θ, sample S
- Suppose U a union of disjoint sub-populations, $U = \bigcup_{j=1}^{J} U_j$
- What if we want to estimate θ_j for some U_j?
  - If $\theta =$ unemployment rate, may be curious about $\theta_{\text{Massachusetts}}$, θ_{Maine}, θ_{Vermont}, etc. in addition to overall θ

The general problem (continued)

- **Small area estimation** addresses how how we can produce estimators for θ_j
- Could simply use same estimator as with U, but on subset of S belonging to U_j
  - $n_j < n$, so $\text{Var}(\hat{\theta}_j)$ could be (very) large
- Want to find "better" $\hat{\theta}_j$, particularly for when $n_j << n$

Big Picture: Multilevel Regression with Poststratification

- Identify categorical respondent attributes x_i in survey that are also available in census data (including area identifiers[1])
- Gather auxiliary area-level variables ($\underline{x}_j$) we that may be predictive of y_i
- Use (bespoke) regression to estimate the effects of individual characteristics (x_i) and area characteristics ($\underline{x}_j$) on y_i
- Construct "poststratification table" with cell (row) for every combination such attributes and population size N_c of that cell
- Calculated predicted value $\hat{\mu}_{yc}$ (estimator of $\mathbb{E}[y_i | i \in U_c]$) for each cell in poststratification table

$$\text{Then we say... } \hat{\mu}_{yj} = \frac{\sum_{c \in j} N_c \hat{\mu}_{yc}}{\sum_{c \in j} N_c}$$

[1] Technically can be applied any categorical grouping, not just spatial ones.

Example multilevel logit

- Assume data follows generating process

$$\Pr(y_i = 1) = \frac{e^{(\alpha_j+\beta_1 x_i)}}{1 + e^{(\alpha_j+\beta_1 x_i)}}, \text{ for observation } i \text{ in group } j$$

- Try to estimate β and a separate intercept α_j for each group j
  - **Prior**: α_j's follow some distribution (often $\alpha_j \sim N(0, \sigma^2)$)
  - **Posterior**: adjusts for groups with large-enough n_j
    - For groups with small n_j and outlier y_i distribution, accounts for possibility of sampling variance than outlier α_j

In the news


News Opinion Sport Culture Lifestyle More

UK World Climate crisis Ukraine Football Newsletters Business Environment **UK politics** Education Society Science Tech

Opinion polls

How accurate are MRP polls predicting huge Tory losses in next general election?

Firms combine polling and census-type data to predict outcomes but with sometimes differing results

Demonstration:

Question: what do younger adults think of gun control, by state?

- Largest survey of political attitudes: Cooperative Election Study
  - 51,550 adults completed both rounds in 2020
  - Just 6,004 of them under age 30
    - In individual states, $n_{\text{under-30}}$ very small
- To see if $n = 6,004$ reasonable for state estimates, can first try estimating something we know: 2020 vote totals

Model (2020 vote)

$y_i = 1$ if voted for Biden, 0 if voted for someone else / didn't vote

Use 'stan_glmer' in R to estimate posterior distributions of (weighted) multilevel logit parameters...

α^{state} for every state

α^{race} for every race category

α^{age} for every age category

$\alpha^{education}$ for every education category

$\alpha^{race\_education}$ for every race x education interaction

β^{male}

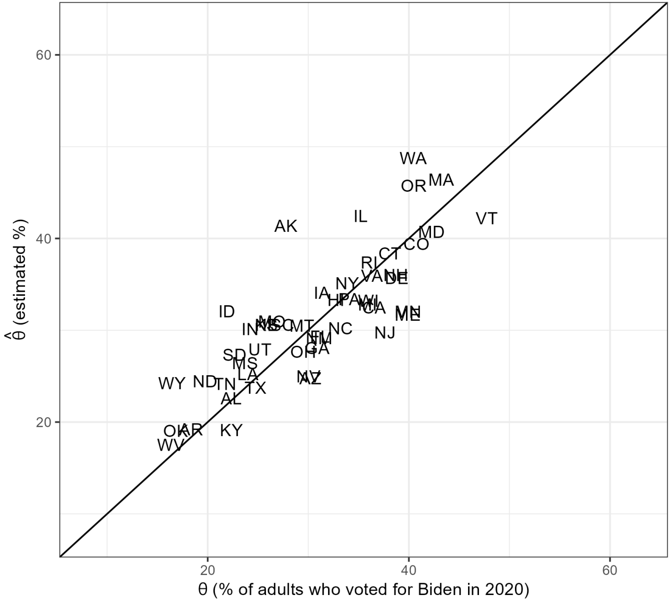
β^{2016_vote}

β^{region} for non-baseline regions (EPA definitions)

Computing estimates

- Construct poststratification table using Census IPUMS
  - N_c for each cell $\{$state$\}x\{$gender$\}x\{$race$\}x\{$education$\}x\{$age$\}$
    - 10,200 rows!
  - Merge in 2016 vote data and add regions
- Take 4,000 draws of parameter values from posterior distributions, computing every state's mean each time
  - For each state take mean and quantiles 2.5 and 97.5

MRP estimates vs. actual % of adults voting for Biden



Estimates calculated using 2020 ACS and sample of 6,004 respondents from 2020 Cooperative Election Study

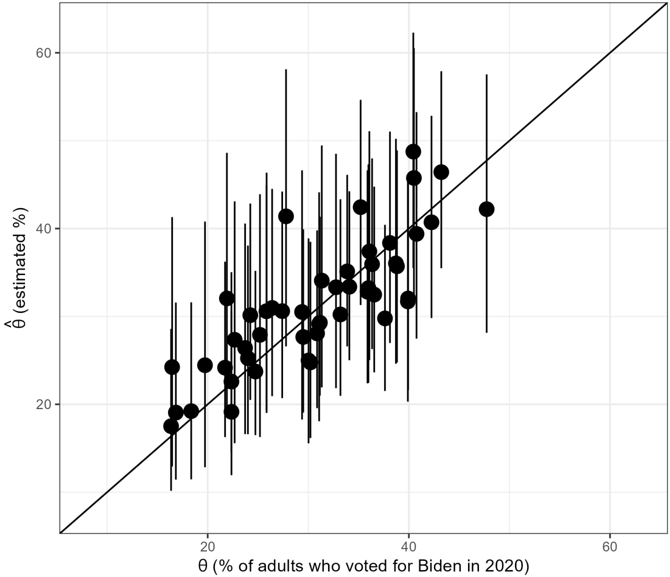
MSE lower than direct estimator

	States	MRP_MSE	Direct_MSE
1	All states	21.48948	77.86618
2	25 largest states	16.92803	28.03674
3	25 smallest states	26.05093	127.69562

- Much lower MSE than "direct estimator" (weighted mean of 'biden_vote' by state)
  - True even of larger states

MRP estimates vs. actual % of adults voting for Biden in 2020

95% confidence intervals shown



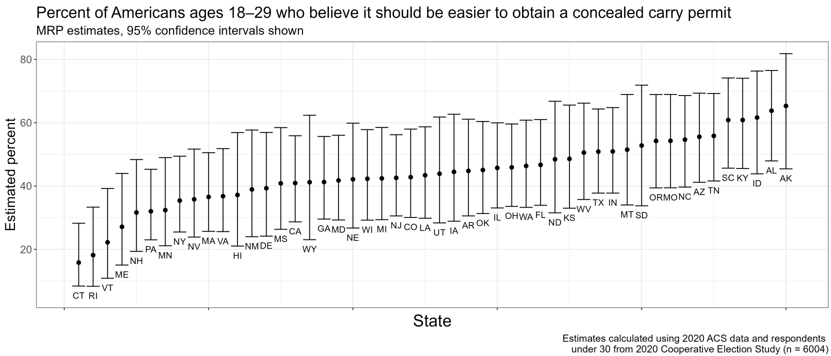
Estimates calculated using 2020 ACS and sample of 6,004 respondents from 2020 Cooperative Election Study

Gun control

Now back to young adults' views of gun control

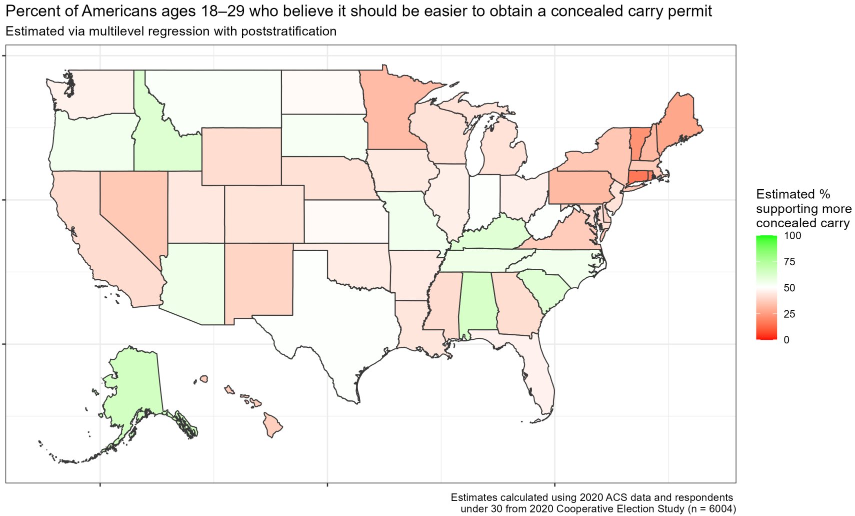
- $n = 6,004$ respondents aged 18–29
- Each asked whether they support policy to "make it easier for people to obtain concealed-carry permit"
- Run same model using this as y_i, and dropping age parameter

Estimates



Percent of Americans ages 18–29 who believe it should be easier to obtain a concealed carry permit
MRP estimates, 95% confidence intervals shown
Estimated percent
80
60
40
20
CT RI VT ME NH PA MN NY NV MA VA HI NM DE MS CA WY GA MD NE WI MI NJ CO LA UT IA AR OK IL OH WA FL ND KS WV TX IN MT SD OR MO NC AZ TN SC KY ID AL AK
State
Estimates calculated using 2020 ACS data and respondents
under 30 from 2020 Cooperative Election Study (n = 6004)

Best part of small area estimation: maps!



Further reading:

- General guides to MRP:
  - Lopez-Martin et al. (2022), "Multilevel Regression and Poststratification Case Studies"
  - Alexander (2023), *Telling Stories with Data*, chapter 16
- On multilevel models:
  - Gelman and Hill (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*
  - Snijders and Bosker (2012), *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*
- Data this would not have been possible without:
  - 2020 ACS IPUMS
  - 2020 Cooperative Election Study