Multilevel Regression with Poststratification: A Tutorial

Julian Perry

Harvard University
Stat 160: Introduction to Survey Sampling and Estimation
Professor Kelly McConville

May 2024

1. Overview

Generally, survey sampling and estimation take the following structure:

- Define target population U and parameter of interest θ
- Take sample $S \subset U$ (ideally via probability sampling)
- Using values known for observations in S, produce estimate $\hat{\theta}$ of θ

Suppose, however, U is made of J mutually exhaustive subpopulations U_j, such that

$$U = \bigcup_{j=1}^{J} U_j,$$

and for $j \neq j'$,

$$U_j \cap U_{j'} = \emptyset.$$

These subpopulations' characteristics θ_j are often of interest, in addition to θ. The process of trying to estimate them is known as **small-area estimation**.[1]

One popular small-area estimation approach — with applications from election forecasting (Park et al. 2004) to assessing public health (Downes et al. 2018) — is **multilevel regression with poststratification** (**MRP**). MRP requires the following ingredients:

(1) (Categorical) characteristics X_i for each sampled observation that are related to study variable y_i, and the subpopulation U_j to which observations belong

(2) The number of population units in every U_j with each combination of characteristics in X_i

(3) A multilevel model that provides "reasonable"[2] predictions of y_i among sampled observations, using covariates X_i, membership in group U_j, and known group-level characteristics X_j

[1] It isn't necessary that subpopulations correspond to spatial areas, though this will be the case in examples here.
[2] Meaning either low bias, low variance, or some combination.

With these, it's possible to construct a poststratification table with a row for every combination of characteristics in X_i for every group U_j, along with the population size of that "cell" and any auxiliary X_j of group U_j used in the model. Then, using the model, estimate the conditional expectation of y_i for the cell in each row,

$$\mathbb{E}[y_i|i \in U_c] = \hat{\mu}_{yc}.$$

Then, aggregate $\hat{\mu}_{yc}$'s within each U_j (weighted by cell population N_c) to estimate group means,

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

Additionally, when MRP is used not for small-area estimation but to correct for non-representative samples (see Wang et al. 2015), these estimates can be used to estimate a population mean

$$\hat{\mu}_y = \frac{\sum_j N_j \hat{\mu}_{yj}}{\sum_j N_j}.$$

2. Demonstration

To demonstrate how this works in practice, I start by estimating a parameter whose true values are known, the percent of each state's adults who voted for Biden in 2020. I use the 2020 Cooperative Election Study (Schaffner et al. 2021), a survey of adults before and after that year's election. 51,550 completed both rounds; however, I'll follow the example of Lopez-Martin et al. (2022) and demonstrate that MRP works on smaller subset — in particular, a random sample of 6,004 of them. This is equal to the number of respondents under 30, a subset I'll analyze later.

For the poststratification table, I use 2020 ACS (Ruggles et al. 2024) estimates of the approximate population size of each unique combination of the variables

{state} x {education} x {race} x {age} x {gender},

creating 10,200 cells (Appendix A contains exact definitions of categories). This set of covariates (and sometimes citizenship too) is standard in MRP models in political science, because of their known population distributions and their ability to explain a large portion of variance in Americans' political alignments (see Kastellec et al. 2014, Lopez-Martin et al. 2022, Park et al. 2004). After modifying CES data so the corresponding variables are grouped into the same categories, I merge auxiliary state-level data into both datasets, with Hillary Clinton's share of the 2016 vote (MIT Election Data and Science Lab 2017) and the state's region, using EPA definitions (Regional and Geographic Offices 2024).

As a dependent variable, I code each sampled respondent 1 if they are a validated voter[3] and said they voted for Biden, 0 otherwise. Given the binary nature of this variable, one possibility would be to use a standard logit regression, with indicators for non-baseline categories of each variable. However, the inclusion of so many indicators, some applying to just a few observations, would overfit the model, worsening out-of-sample predictions. Further, state indicators would be perfectly collinear with state-level characteristics, posing a challenge to identification.

[3]Absent validation using voter records, survey respondents exaggerate their own turnout in elections, which would bias estimates of vote totals (Ansolabehere and Hersh 2012)

The solution to these problems that makes MRP effective is using *multilevel* models, which assume values of the dependent variable are correlated within groups, and assume group-level effects come from a particular distribution. Thus, as the sample size for a particular group diminishes, the estimated group effect shrinks, accounting for the possibility that observed observed values for that group may be the result of sampling variance. In this case I use a weighted multilevel logit model, with parameters

α^{state} for every state

α^{race_eth} for race/ethnicity categories

α^{age_cat} for age categories

α^{edu_cat} for education categories

$\alpha^{race\_eth\_edu\_cat}$ for race-education interactions[4]

β^{male}

β^{2016_vote}

β^{region} for indicators of each non-baseline region (EPA definitions)

which I estimate using a Bayesian approach. I begin with the prior that the effects of categories for each α are normally distributed about 0. Using the `stan_glmer` command in R's `rstanarm` package, I run the following code to estimate posterior distributions of possible values for each parameter, using the CES weights for respondents who completed both rounds:

```
model_bidenvotes <- stan_glmer(voted_biden ~ (1 | state_fips) +
        (1 | race_eth) + # "(1 | variable)" denotes variables estimated as group-level effects
        (1 | edu_cat) +
        gender + # just two gender groups in data, so not estimating as group-level effects
        (1| age_cat) +
        (1 | edu_cat:race_eth) + # education-race interaction
        dem_share_16 +
        factor(epa_region), # Lopez-Martin et al suggest regular factors for state-level vars
        family = binomial(link = "logit"),  # telling R we're using logit
        data = CES_s, # sample of 6004 from CES
        weights = commonpostweight, # using weights for subset who completed both rounds
        prior = normal(0, 1, autoscale = TRUE), #prior: group effects normal about 0, SD scaled
        adapt_delta = 0.99, # Lopez-Martin et al say higher values stabilize estimates
        seed = 160) # setting seed for random components of estimation process
```

Because I use a Bayesian approach, the result is not a single estimate of each parameter but a posterior distribution of possible values and their likelihoods, after adjusting priors based on the observed data. As demonstrated by Alexander (2023), it is possible to use the `add_epred_draws` function in R's `tidybayes`

[4]The political gap between college graduates and non-college graduates tends to be larger among white Americans (Marble 2023). Gelman and Ghitza (2013) argue that multilevel models make it reasonable to use more interaction terms than "classical" regression.

package to repeatedly draw parameter values from posterior distributions, and then, as if those were true parameter values, produce state estimates using the poststratification table.

The following code to repeats this process 4,000 times, taking the mean of each state's 4,000 estimates at the end along with the quantiles to produce a 95% confidence interval:

```
results_w_estimates <-
    model_bidenvotes %>%
    add_epred_draws(newdata = ps_table,ndraws = 4000) %>% #take 4000 draws, poststratify to ps_table
    rename(biden_votes = .epred) %>%
    mutate(biden_votes_cell = biden_votes * n) %>% # multiplying probability by pop size of cell
    ungroup() %>%
    summarise(state_biden_votes = sum(biden_votes_cell),
            .by = c(state_fips, .draw)) %>% # adding up votes for cells in each state (for each draw)
    summarise(
        state_votes_est = mean(state_biden_votes), # mean of each state's 4000 draws
        CI_lower = quantile(state_biden_votes, 0.025), # lower bound on 95% confidence interval
        CI_upper = quantile(state_biden_votes, 0.975), # upper bound on 95% confidence interval
        .by = state_fips
        )  %>%
    left_join(actual_results, by = "state_fips") %>% #merging w/ true results and adult populations
    mutate(pct_est = 100*state_votes_est/adult_pop) %>% # converting # of Biden votes to % of adults
    mutate(pct_est_upper = 100*CI_upper/adult_pop) %>%
    mutate(pct_est_lower = 100*CI_lower/adult_pop) %>%
    mutate(pct_voting_biden = 100*dem_votes/adult_pop) #converting true vote total to %
```

Figure 1 shows these estimates (Appendix B contains code for producing all figures).

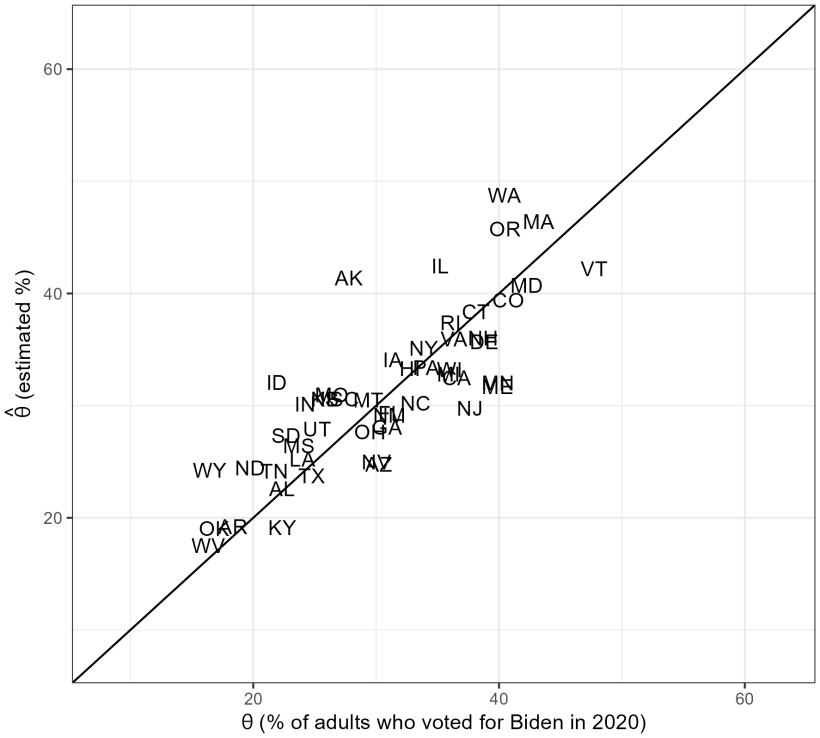
Compared to the "direct estimator" (Lohr 2022) — the weighted share (as a percent) voting for Biden among observations in each state — the MRP estimates have much lower mean squared error (MSE), 21.49 versus 77.87. This is unsurprising for small states, which have as few as seven observations, but even among the 25 states with the largest adult populations, MRP still produces lower MSE than direct estimates (16.93 versus 28.04).

In every case, the 95% confidence intervals contain the true θ_j (Figure 2), giving credibility to MRP confidence intervals in settings where the estimand isn't known. Turning to such an example now, I use a similar model to estimate younger Americans' attitudes toward gun laws.

Because there are only 6,004 respondents under 30 who completed both CES rounds, the constraint to that sample size is now binding. Using this subset, I estimate a similar model (code in Appendix B) with a new dependent variable, coded 1 if the respondent supports a hypothetical proposal to "Make it easier for people to obtain concealed-carry permit" (I also remove `age`, now that only one level is present).

I repeat the process described earlier of drawing from posterior distributions to produce estimates and confidence intervals (code in Appendix B). Figure 3 shows these results, and Figure 4 shows them as a map made with R's `urbnmapr` package. The confidence intervals are large, but the prior example gives them credibility as bounds for the estimates' uncertainty. While I opted to add no new regressors going from the first to second model, in practice it would be reasonable to incorporate much more auxiliary data (state-level gun ownership, etc), which, if it is strongly predictive, could add precision to estimates.

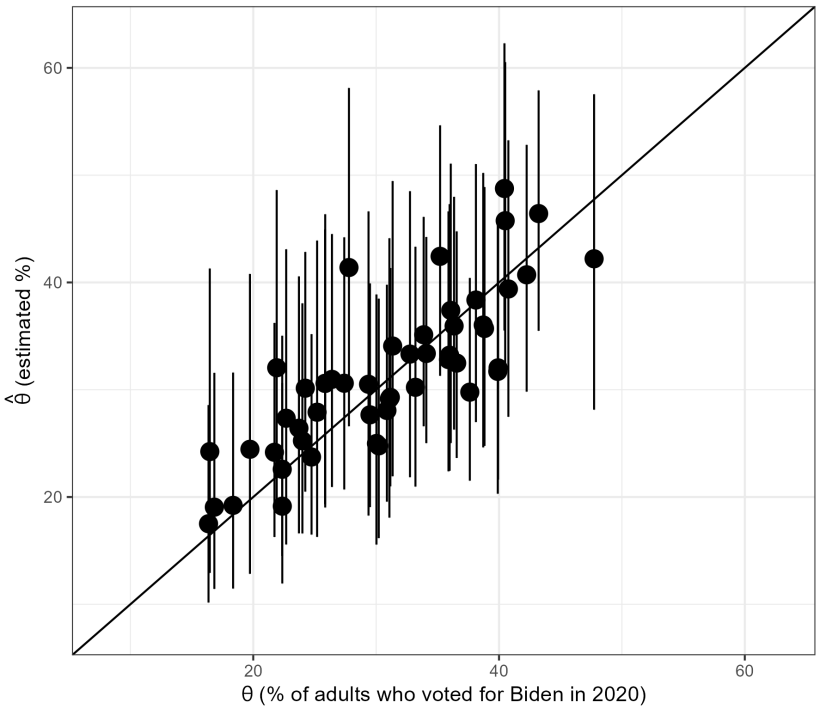
Figure 1: MRP estimates vs. actual % of adults voting for Biden



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

Figure 2: MRP estimates vs. actual % of adults voting for Biden

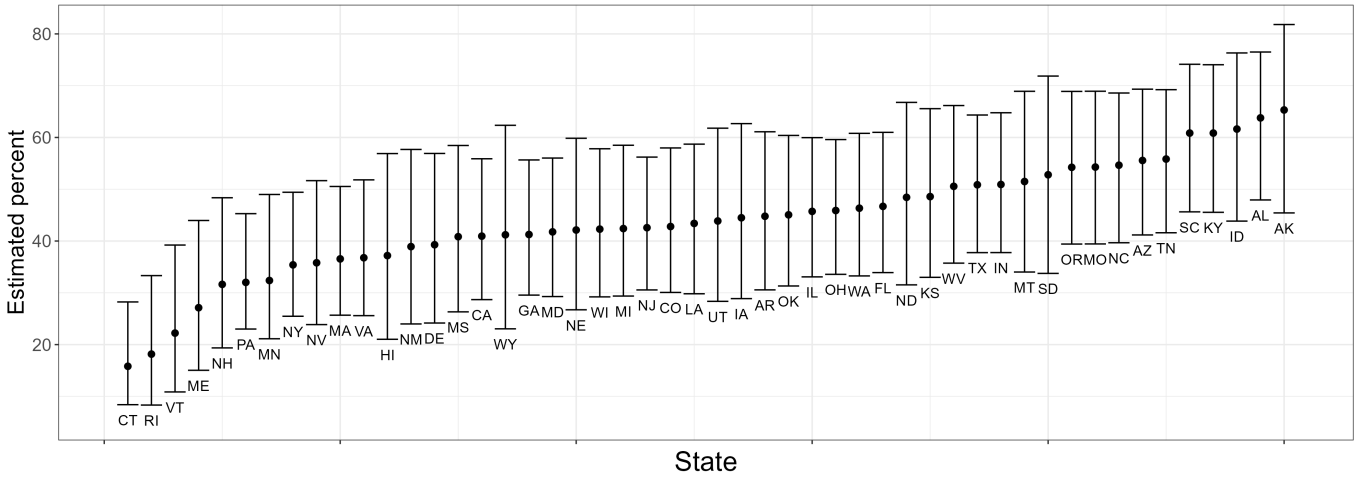
95% confidence intervals shown



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

Figure 3: Percent of Americans ages 18–29 who believe it should be easier to obtain a concealed carry permit

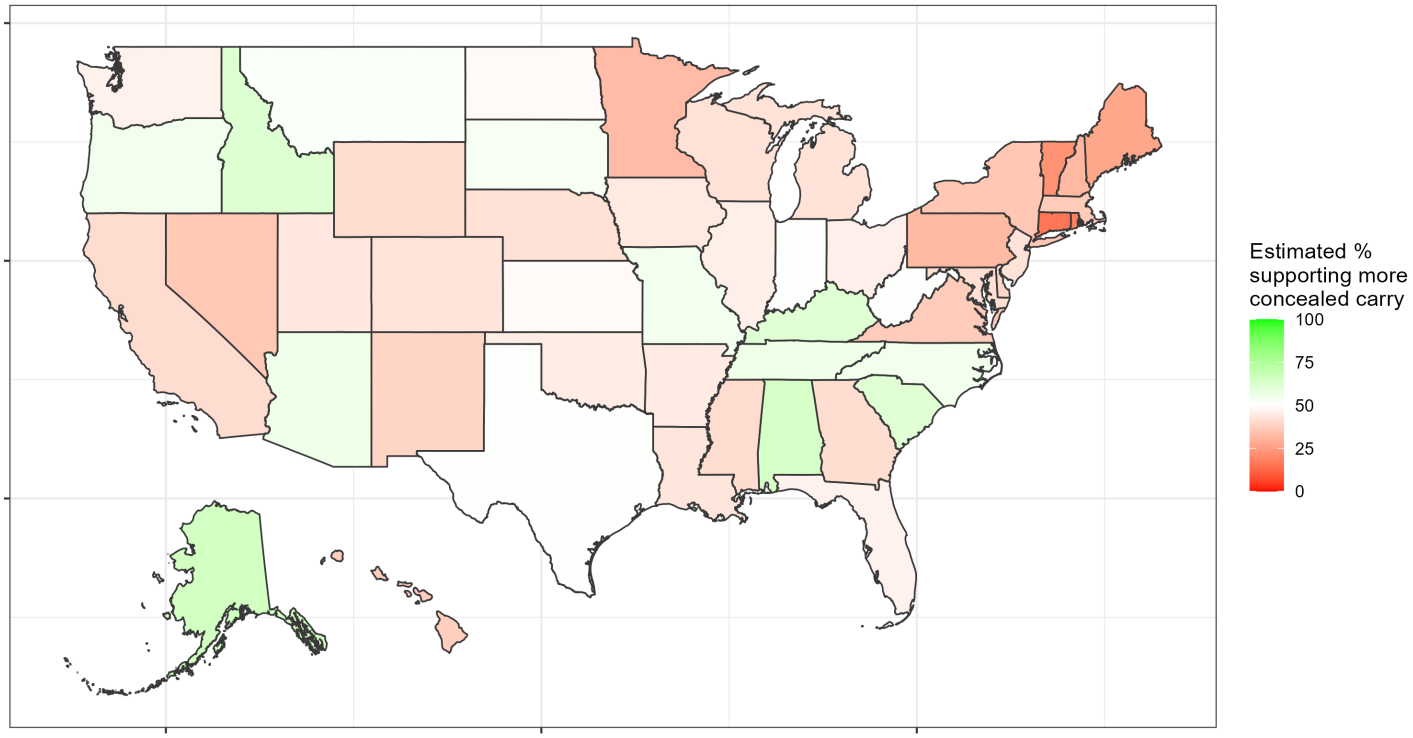
MRP estimates, 95% confidence intervals shown



Estimates calculated using 2020 ACS data and respondents under 30 from 2020 Cooperative Election Study (n = 6004)

Figure 4: Percent of Americans ages 18–29 who believe it should be easier to obtain a concealed carry permit

Estimated via multilevel regression with poststratification



Estimates calculated using 2020 ACS data and respondents under 30 from 2020 Cooperative Election Study (n = 6004)

Appendix A: Defining variables

I construct variables that have common definitions for my model and the poststratification table, defined below:

- `state` has values for all fifty states
- `race_eth` has four categories: 'Black', 'white', 'Hispanic', and 'other'
- `edu_cat` has categories for respondents whose highest level of education are graduate school, a bachelor's degree, some college (without a bachelor's degree), high school, and less than high school
- `age_cat` has categories for respondents aged 18–29, 30–39, 40–49, 50–65, and 65+ (calculating age as 2020 - birth year)

Additionally, though I do not explicitly construct them as variables, group effects are estimated for each interaction between `race_eth` and `edu_cat` as a result of how I specify `stan_glmer()`.

Other variables I include as regressors (but not as group-level effects) are

- `gender`, which Lopez-Martin et al. (2022) recommend not estimating as a group-level effect, since there are just two categories in ACS data
- `region`, a factor variable identifying the region (EPA definition) to which each state belongs (Lopez-Martin et al. 2022 recommend coding variables that vary at the state level as regular indicator variables, hence not estimating this as a group effect when I run `stan_glmer()`)
- `2016_vote`, the share of that state's major-party votes that went to Hillary Clinton in 2016

Appendix B: Additional code

```
# not including every tedious bit of cleaning the data
# but commands for model specification and graphics are here

# code to produce Figure 1:
# state_po has state's two-letter po abbreviation

ggplot(results_w_estimates, aes(x = pct_voting_biden, y = pct_est,label=(state_po))) +
  geom_text() +
  geom_abline(intercept = 0, slope = 1) + # 45-degree line, along which ideal estimates would fall
  xlim(8, 63) +
  ylim(8, 63) + theme_bw() +
  xlab(TeX("$\\theta$ (% of adults who voted for Biden in 2020)"))+
  ylab(TeX("$\\hat{\\theta}$ (estimated %)")) + labs(
    title = "Figure 1: MRP estimates vs. actual % of adults voting for Biden",
    caption = "Estimates calculated using 2020 ACS and sample of 6,004 \nrespondents from 2020 Cooperati
# forgive run-on code, ends "respondents from 2020 Cooperative Election Study")"

# code to produce Figure 2:
```

```
# (same as Figure 1, but with points + 95% confidence intervals instead of PO codes)

ggplot(results_w_estimates, aes(x = pct_voting_biden, y = pct_est)) +
  geom_point(size = 4) +
  geom_errorbar(aes(ymax = pct_est_upper, ymin = pct_est_lower)) +
  geom_abline(intercept = 0, slope = 1) +
  xlim(8, 63) +
  ylim(8, 63)+ theme_bw() +
  xlab(TeX("$\\theta$ (% of adults who voted for Biden in 2020)"))+
  ylab(TeX("$\\hat{\\theta}$ (estimated %)")) + labs(
    title = "Figure 2: MRP estimates vs. actual % of adults voting for Biden",
    subtitle = "95% confidence intervals shown",
    caption = "Estimates calculated using 2020 ACS and sample of 6,004 \nrespondents from 2020 Cooperati



# code for second model:
# only differences from first model:
#       new dependent variable (1 = wants concealed carry to be easier)
#       different sample (now just respondents under 30)
#       age parameter dropped (no longer any varance in tht)

model_concealed_carry <- stan_glmer(concealed_carry ~ (1 | state_fips) +
    (1 | race_eth) +
    (1 | edu_cat) +
    gender +
    (1 | edu_cat:race_eth) + # education-race interaction
    dem_share_16 +
    factor(epa_region),
    family = binomial(link = "logit"),  # using logit
    data = CES_under30, # using sample of just people under 30
    weights = commonpostweight, # using weights for subset who completed both rounds
    prior = normal(0, 1, autoscale = TRUE), # prior: group effects normal about 0, SD scaled
    adapt_delta = 0.99, # again, higher values stabilize estimates
    seed = 160) # setting seed for random components of estimation

# generating estimates and confidence intervals for second model:
# basically the same as before
#   difference is we're using new poststratification table of just cells age 18-29
#   and denominator for % is population in that age group
#   also we don't have "true" parameters to merge in for comparison

youth_concealed_carry_estimates <-
  model_concealed_carry %>%
```

```
  add_epred_draws(newdata = under30_ps_table,ndraws = 4000) %>% # new ps table, just cells in age group
  rename(concealed_carry_support = .epred) %>%
  mutate(concealed_carry_supporters = concealed_carry_support * n) %>%
  ungroup() %>%
  summarise(concealed_carry_supporters = sum(concealed_carry_supporters),
            .by = c(state_fips, .draw)) %>%
  summarise(
    state_votes_est = mean(concealed_carry_supporters),
    CI_lower = quantile(concealed_carry_supporters, 0.025),
    CI_upper = quantile(concealed_carry_supporters, 0.975),
    .by = state_fips
  ) %>%
  left_join(youthpops_2020, by = "state_fips") %>%
  mutate(pct_est = 100*state_votes_est/youth_pop) %>%
  mutate(pct_est_upper = 100*CI_upper/youth_pop) %>%
  mutate(pct_est_lower = 100*CI_lower/youth_pop)


# code for Figure 3:
# putting in "order" variable, by estimate magnitude

youth_concealed_carry_estimates <-
    youth_concealed_carry_estimates[
    order(youth_concealed_carry_estimates$pct_est),
    ] %>%
      mutate(order = row_number())


# adding amount by which to nudge each label, don't want overlap with error bar

youth_concealed_carry_estimates$lab_nudge <- NA
youth_concealed_carry_estimates$lab_nudge <-
    youth_concealed_carry_estimates$pct_est_lower - youth_concealed_carry_estimates$pct_est - 3

# figure itself

ggplot(youth_concealed_carry_estimates, aes(x = order, y = pct_est,label=(state_po))) +
  geom_point(size = 2) +
  geom_errorbar(aes(ymax = pct_est_upper, ymin = pct_est_lower)) +
  geom_text(nudge_y = youth_concealed_carry_estimates$lab_nudge)+
  theme_bw() +
  xlab("State")+
  theme(
```

```
    plot.title = element_text(size = 19),
    plot.subtitle = element_text(size = 15),
    plot.caption = element_text(size = 12),
    axis.title.x = element_text(size = 20),
    axis.text.x = element_text(size = 0.1),
    axis.text.y = element_text(size = 13),
    axis.title.y = element_text(size = 17))+
  ylab("Estimated percent") + labs(
    title = "Figure 3: Percent of Americans ages 18-29 who believe it should be easier to obtain a conc
    subtitle = "MRP estimates, 95% confidence intervals shown",
    caption = "Estimates calculated using 2020 ACS data and respondents \nunder 30 from 2020 Cooperativ

# title line (which runs on here) ends...
#       "...to obtain a concealed carry permit","

# and caption line ends...
#       "...from 2020 Cooperative Election Study (n = 6004)")"



# and code for figure 4 (my favorite!)

# loading pre-made outline using command from urbnmapr package
states_urbnmapr <- get_urbn_map(map = "states")
states_urbnmapr$state_fips <- as.integer(states_urbnmapr$state_fips)

map <- left_join(states_urbnmapr, youth_concealed_carry_estimates, by = "state_fips")

ggplot() + geom_polygon(data = map, mapping = aes(x = long, y = lat, group = group, fill = pct_est)) +
theme_bw() +
geom_polygon(data = map, mapping = aes(x = long, y = lat, group = group),
fill = NA, color = "grey22", linewidth = 0.4) +
theme(axis.text.x = element_blank(), axis.text.y = element_blank(),
axis.title.x = element_blank(),axis.title.y = element_blank()) +
labs(fill="Estimated %\nsupporting more\nconcealed carry",
title="Figure 4: Percent of Americans ages 18-29 who believe it should be easier to obtain a concealed
subtitle = "Estimated via multilevel regression with poststratification",
caption = "Estimates calculated using 2020 ACS data and respondents \nunder 30 from 2020 Cooperative El
scale_fill_gradient2(low="red", high="green", midpoint=50, limits=c(0,100)) # setting gradient

# again, title and caption run on too long here
# title line ends:
#   "...obtain a concealed carry permit","
```

```
# caption line ends:
#   "...respondents \nunder 30 from 2020 Cooperative Election Study (n = 6004)") +"
```

Works Cited

Alexander, Rohan. "Multilevel regression with post-stratification." *Telling Stories with Data*, 1st ed., vol. 1, CRC Press, 2023, pp. 503–24, 9781032134772, https://tellingstorieswithdata.com/15-mrp.html.

Ansolabehere, Stephen, and Eitan Hersh. "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis*, vol. 20, no. 4, 2012, 437–459. https://doi.org/10.1093/pan/mps023.

Downes, Marnie, et al. "Multilevel Regression and Poststratification: A Modeling Approach to Estimating Population Quantities From Highly Selected Survey Samples." *American Journal of Epidemiology*, vol. 187, no. 8, Apr. 2018, pp. 1780–90. *https://academic.oup.com/aje/article-pdf/187/8/1780/25369165/kwy070.pdf*, https://doi.org/10.1093/aje/kwy070, https://doi.org/10.1093/aje/kwy070.

Ghitza, Yair, and Andrew Gelman. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American journal of political science*, vol. 57, no. 3, 2013, pp. 762–76.

Kastellec, Jonathan P, et al. Estimating State Public Opinion with Multi-Level Regression and Poststratification Using R. 2014. https://jkastellec.scholar.princeton.edu/publications/mrp_prime.

Lohr, Sharon L. *Sampling : design and analysis*. Third edition., CRC Press, 2022, 9781000478266. Chapman Hall CRC texts in statistical science.

Lopez-Martin, Juan, et al. *Multilevel Regression and Poststratification: A Practical Guide and New Developments*. 2022, https://bookdown.org/jl5522/MRP-case-studies/.

Marble, William. What Explains Educational Polarization Among White Voters? 2023. https://williammarble.co/docs/EducPolarization.pdf.

MIT Election Data and Science Lab. U.S. President 1976–2020. 2017. https://doi.org/10.7910/DVN/42MVDX, https://doi.org/10.7910/DVN/42MVDX.

Park, David K., et al. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis*, vol. 12, no. 4, 2004, 375–385. https://doi.org/10.1093/pan/mph024.

"Regional and Geographic Offices," https://www.epa.gov/aboutepa/regional-and-geographic-offices. Accessed 4 May 2024.

Ruggles, Steven, et al. IPUMS USA. 2024. https://doi.org/https://doi.org/10.18128/D010.V15.0, usa.ipums.org.

Schaffner, Brian, et al. Cooperative Election Study Common Content, 2020. 2021. https://doi.org/10.7910/DVN/E9N6PH, https://doi.org/10.7910/DVN/E9N6PH.

Wang, Wei, et al. "Forecasting elections with non-representative polls." *International Journal of Forecasting*, vol. 31, no. 3, 2015, pp. 980–91. https://doi.org/https://doi.org/10.1016/j.ijforecast.2014.06.001, https://www.sciencedirect.com/science/article/pii/S0169207014000879.