Multilevel Regression And Poststratification (A Primer)

Joseph T. Ornstein*

October 26, 2021

1 Running Example

To demonstrate how MRP works, we'll consider an example where we know the "real" answer, and can explore how various refinements to the model improve predictive accuracy. The approach we'll use mirrors that in Buttice and Highton (2013), taking responses from a large scale US survey of voters (schaffner ref).[1]

```
library(tidyverse)
library(ggrepel)

load('data/CES-2020.RData')
```

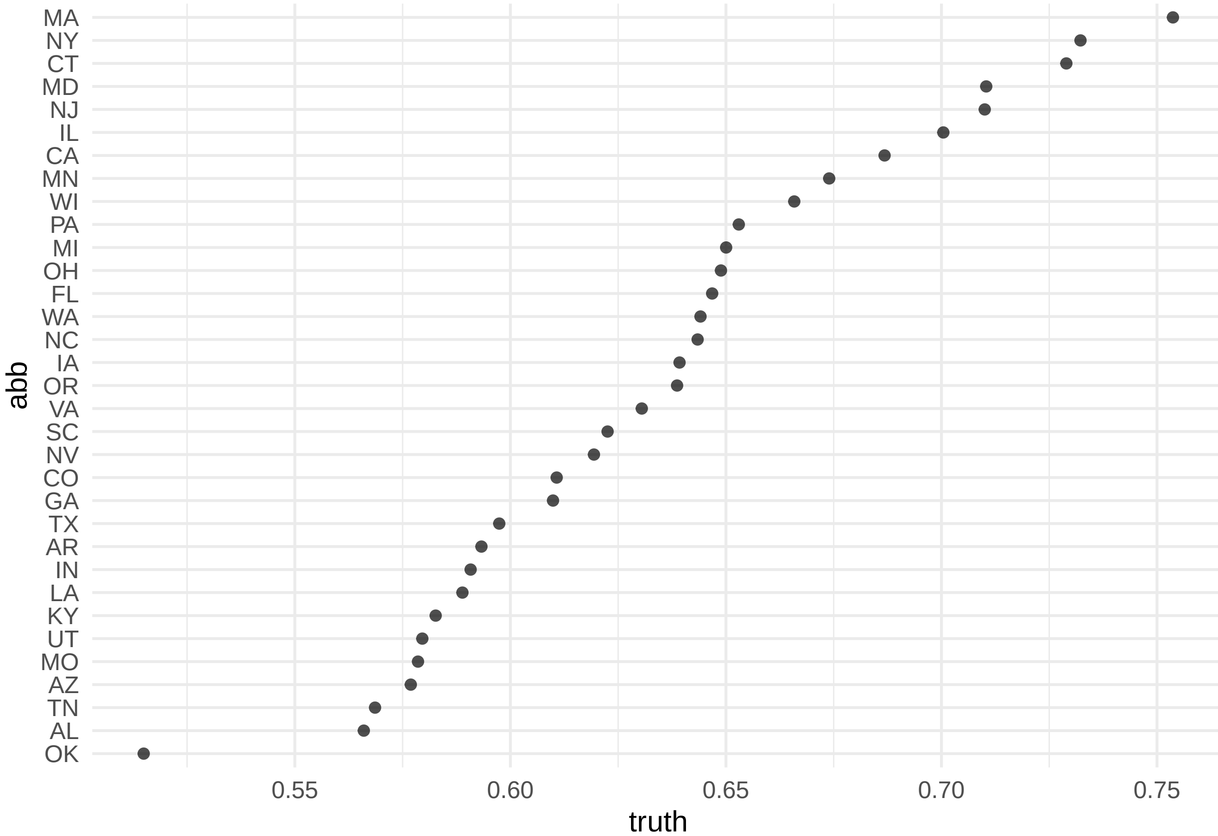
The data is available here, and we'll be using a tidied up version of the dataset created by `R/cleanup-ces-2020.R`. This tidied version of the data only includes the 33 states with at least 500 respondents.

1.1 The Truth

```
truth <- ces %>%
  filter(!is.na(assault_rifle_ban)) %>%
  group_by(abb) %>%
  summarize(truth = sum(assault_rifle_ban == 'Support') / n())

# plot
truth %>%
  # reorder abb so the chart is organized by percent who support
  mutate(abb = fct_reorder(abb, truth)) %>%
  ggplot(mapping = aes(x=truth, y=abb)) +
  geom_point(alpha = 0.7) +
  theme_minimal()
```

*Department of Political Science, University of Georgia

[1] Throughout, I will use R functions from the "tidyverse" to make the code more human-readable.



Note what I mean by the "truth" here is the true percentage of CES respondents who supported the assault rifle ban. That's our target. This overstates the percent of the total population that support such a ban, since the CES sample is not a simple random sample.

1.2 Draw a Sample

Step 1: draw a sample.

```
sample_data <- ces %>%
  slice_sample(n = 500)
```

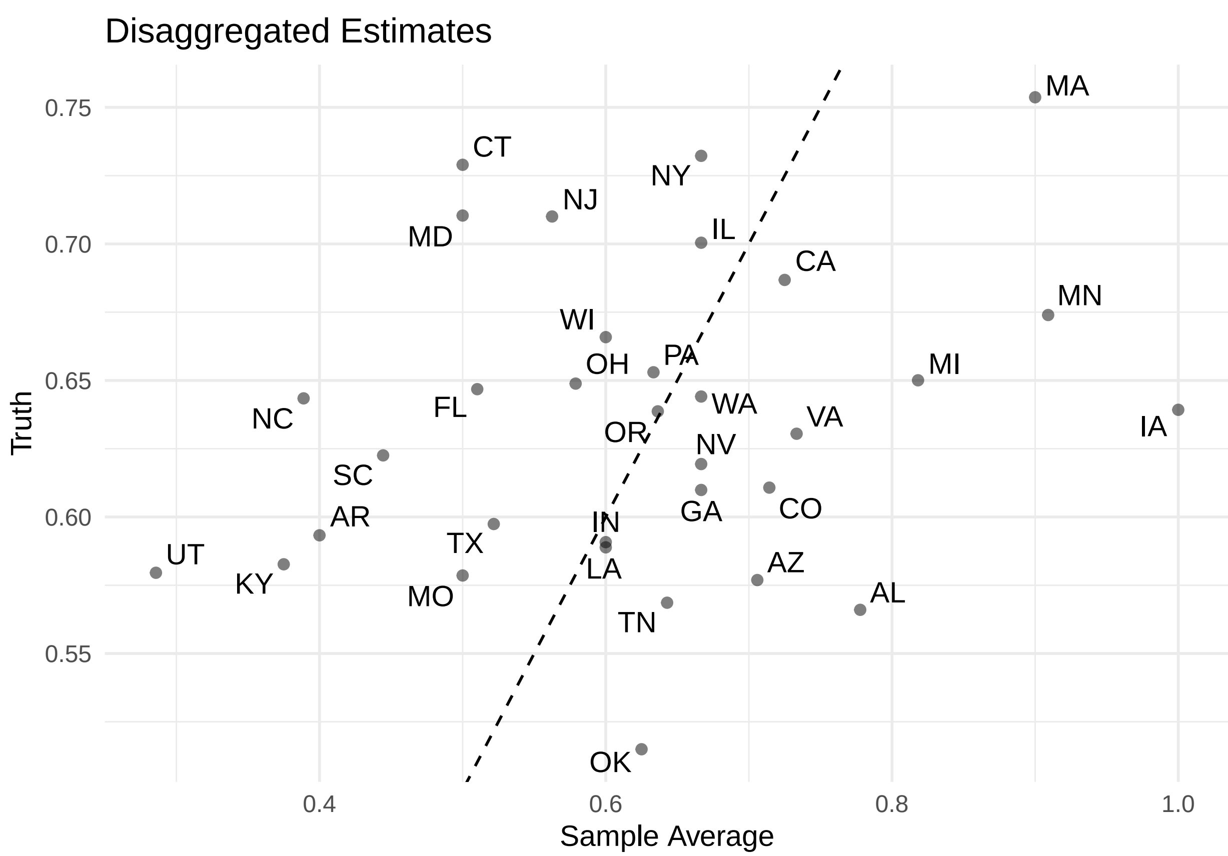
```
sample_summary <- sample_data %>%
  filter(!is.na(assault_rifle_ban)) %>%
  group_by(abb) %>%
  summarize(pct_support = sum(assault_rifle_ban == 'Support') / n(),
            num = n())

sample_summary
```

```
## # A tibble: 33 x 3
##    abb   pct_support   num
##    <chr>       <dbl> <int>
##  1 AL          0.778     9
##  2 AR          0.4       5
##  3 AZ          0.706    17
##  4 CA          0.725    40
##  5 CO          0.714     7
```

```
##   6 CT          0.5        2
##   7 FL          0.510     49
##   8 GA          0.667     18
##   9 IA          1          5
## 10 IL          0.667     21
## # ... with 23 more rows
```

For readers who are less familiar with American politics, rest assured that this is an unrepresentative draw from the state of Iowa. So simply disaggregating and taking sample means will not yield good estimates, as you can see by comparing the percent of respondents from the sample who supported the ban against the percent of CES respondents.

```
sample_summary %>%
  left_join(truth, by = 'abb') %>%
  ggplot(mapping = aes(x=pct_support,
                       y=truth,
                       label=abb)) +
  geom_point(alpha = 0.5) +
  geom_text_repel() +
  theme_minimal() +
  geom_abline(intercept = 0, slope = 1, linetype = 'dashed') +
  labs(x = 'Sample Average',
       y = 'Truth',
       title = 'Disaggregated Estimates')
```



1.3 Multilevel Regression

```r
# TODO: multilevel model; show how partial pooling fixes a bunch here.

# logistic model
model <- glm(as.numeric(assault_rifle_ban == 'Support') ~
               gender + educ + race + age,
             data = sample_data,
             family = 'binomial')

summary(model)
```

```
##
## Call:
## glm(formula = as.numeric(assault_rifle_ban == "Support") ~ gender +
##     educ + race + age, family = "binomial", data = sample_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2034  -1.1676   0.6614   0.9510   1.6246
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.035670   0.897739   1.154   0.2486
## genderMale                 -0.956595   0.204229  -4.684 2.81e-06 ***
## educHigh school graduate -1.187415   0.615309  -1.930   0.0536 .
## educSome college           -1.373742   0.616990  -2.227   0.0260 *
## educ2-year                 -0.613440   0.639581  -0.959   0.3375
## educ4-year                 -0.488137   0.621276  -0.786   0.4320
## educPost-grad              -0.395022   0.653821  -0.604   0.5457
## raceBlack                   0.936116   0.724075   1.293   0.1961
## raceHispanic                0.001496   0.703360   0.002   0.9983
## raceOther                  -0.774865   0.778793  -0.995   0.3198
## raceWhite                   0.022585   0.648816   0.035   0.9722
## age                         0.013543   0.005617   2.411   0.0159 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 665.03  on 499  degrees of freedom
## Residual deviance: 607.47  on 488  degrees of freedom
## AIC: 631.47
##
## Number of Fisher Scoring iterations: 4
```

1.4 Poststratification

```r
psframe <- ces %>%
  count(abb, gender, educ, race, age)

head(psframe)
```

```
## # A tibble: 6 x 6
```

```
##   abb   gender educ  race    age     n
##   <chr> <chr>  <fct> <chr> <dbl> <int>
## 1 AL    Female No HS Black    28     1
## 2 AL    Female No HS Black    29     1
## 3 AL    Female No HS Black    34     1
## 4 AL    Female No HS Black    54     1
## 5 AL    Female No HS Black    64     1
## 6 AL    Female No HS Other    36     1
```

Append predicted probabilities to the poststratification frame.

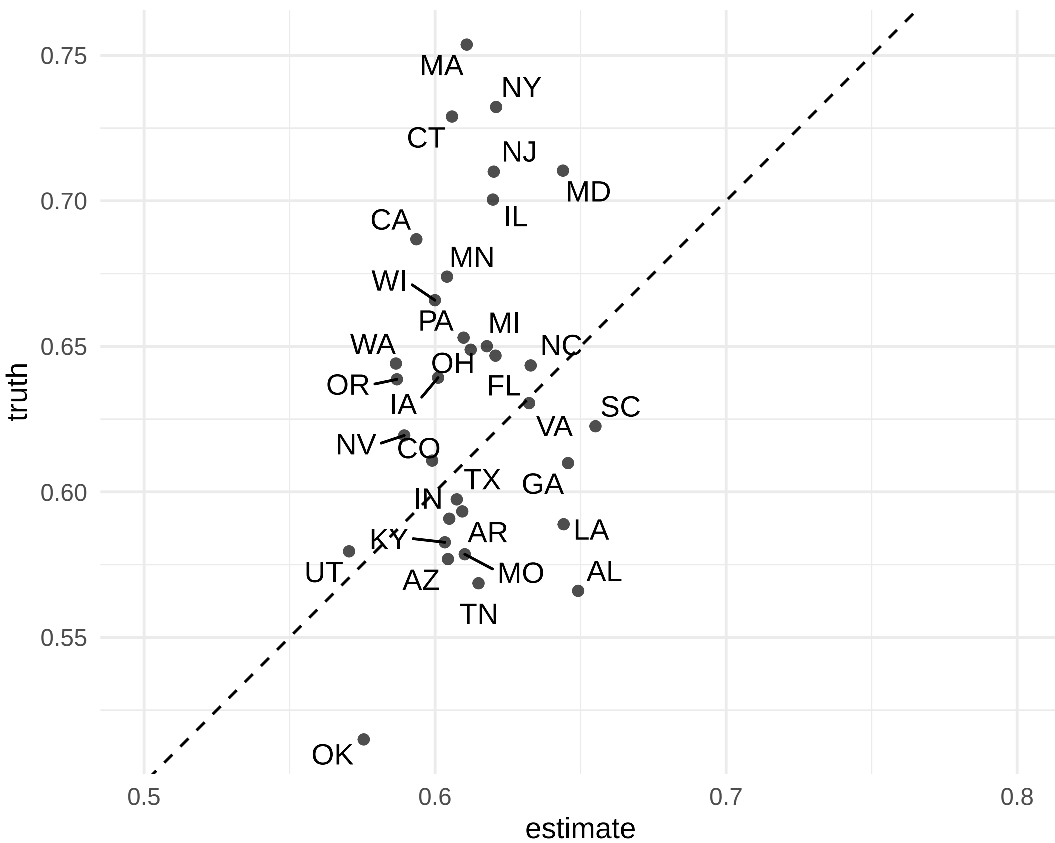
```r
psframe <- psframe %>%
  mutate(predicted_probability = predict(model, psframe, type = 'response'))
```

Poststratified estimates are the population-weighted predictions

```r
poststratified_estimates <- psframe %>%
  group_by(abb) %>%
  summarize(estimate = weighted.mean(predicted_probability, n))
```

Merge and compare:

```r
d <- left_join(poststratified_estimates, truth,
               by = 'abb')

library(ggrepel)

ggplot(data = d,
       mapping = aes(x = estimate,
                     y = truth,
                     label = abb)) +
  geom_point(alpha = 0.7) +
  geom_text_repel() +
  theme_minimal() +
  geom_abline(intercept = 0, slope = 1, linetype = 'dashed') +
  scale_x_continuous(limits = c(0.5, 0.8)) +
  coord_equal()
```



References

Buttice, Matthew K., and Benjamin Highton. 2013. "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" *Political Analysis* 21 (4): 449–67. https://doi.org/10.1093/pan/mpt017.