# MRP for NYC Community Health Survey

Gio Circo

2023-05-24

## MRP Example Code

Pull data from the **NYC Community Health Survey.** Code takes about 15-20 minutes to run end-to-end (pulling data from PUMS is very, very slow).

```
library(tidyverse)
library(tidycensus)
library(brms)
library(sf)

# load survey data
df <-
  haven::read_sas("https://www.nyc.gov/assets/doh/downloads/sas/episrv/chs2019_public.sas7bdat")

# load puma shapefile of NYC
# plot map
nyc_geo <- st_read("geo_export_a92160ac-6718-4bd7-85a8-dac03d8ee421.shp")

# load Public-Use Micro Data
# ky = '<API KEY>'
puma = get_pums(
  state = 'NY',
  variables = c("PUMA", "AGEP", "SEX", "RAC1P", "FHISP", "SCHL"),
  rep_weights = "person",
  key = ky
) %>%
  mutate(SCHL = as.numeric(SCHL))
```

## Data Setup

### Recode PUMA data into strata

```
# Recode PUMA variables to match survey strata
nyc_puma <-
  puma %>%
  mutate(
    age = case_when(
      between(AGEP, 18, 24) ~ "18-24",
      between(AGEP, 25, 44) ~ "25-44",
      between(AGEP, 45, 64) ~ "45-64",
      AGEP >= 65 ~ "65+"
    ),
    sex = ifelse(SEX == 1, "male", "female"),
```

```
    race = case_when(
      RAC1P == 1 & FHISP == 0 ~ "white",
      RAC1P == 2 ~ "black",
      FHISP == 1 ~ "hispanic",
      FHISP == 6 ~ "asian",
      TRUE ~ "other"
    ),
    edu = case_when(
      SCHL %in% 1:15 ~ "less_hs",
      SCHL %in% 16:17 ~ "hs",
      SCHL %in% 18:20 ~ "some_college",
      SCHL %in% 20:24 ~ "college"
    )
  ) %>%
  count(PUMA, age, sex, race, edu, wt = PWGTP) %>%
  na.omit()
```

**Recode survey data**

```
# construct the survey
svy_df <- df %>%
  mutate(
    age = case_when(
      agegroup == 1 ~ "18-24",
      agegroup == 2 ~ "25-44",
      agegroup == 3 ~ "45-64",
      agegroup == 4 ~ "65+"
    ),
    sex = case_when(birthsex == 1 ~ "male",
                    birthsex == 2 ~ "female"),
    race = case_when(
      newrace == 1 ~ "white",
      newrace == 2 ~ "black",
      newrace == 3 ~ "hispanic",
      newrace == 4 ~ "asian",
      TRUE ~ "other"
    ),
    edu = case_when(
      education == 1 ~ "less_hs",
      education == 2 ~ "hs",
      education == 3 ~ "some_college",
      education == 4 ~ "college"
    ),
    health_good = case_when(generalhealth %in% 1:3 ~ 1,
                            TRUE ~ 0)
  ) %>%
  select(age, sex, race, edu, health_good) %>%
  na.omit()

# post strat table
# 160 = 4*2*5*4
post_strat <-
  svy_df %>%
```

```
  expand(age, sex, race, edu)
```

### Run HLM

```
# Regression Step
# Multi-level regression predicting the probability a respondent
# says their health is "excellent", "very good" or "good"

bprior <- c(prior(normal(0, 2), class = "Intercept"),
            prior(normal(0, 2), class = "sd"))

# simple random intercepts model, no interactions
# although we would likely want to do sex*race, and age*sex
fit1 <- brm(
  health_good ~ 1 +
    (1 | age) +
    (1 | sex) +
    (1 | race) +
    (1 | edu),
  family = bernoulli(),
  data = svy_df,
  chains = 4,
  cores = 4,
  iter = 2000,
  control = list(adapt_delta = .95)
)
```

```
## Compiling Stan program...
```

```
## Start sampling
```

```
summary(fit1)
```

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: health_good ~ 1 + (1 | age) + (1 | sex) + (1 | race) + (1 | edu)
##    Data: svy_df (Number of observations: 8722)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~age (Number of levels: 4)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.38      0.75     0.56     3.32 1.00     1519     1784
##
## ~edu (Number of levels: 4)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.10      0.67     0.43     2.88 1.00     1579     2302
##
## ~race (Number of levels: 5)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.55      0.34     0.22     1.47 1.00     1150     2121
##
## ~sex (Number of levels: 2)
```

```
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.80      0.96     0.05     3.51 1.00     1104     1323
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.93      1.07    -1.28     3.02 1.00     1799     1940
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
# get predictions
# draw from posterior, compute means
pp <- posterior_predict(fit1, post_strat)
pred_df <- tibble(post_strat, pred = apply(pp, 2, mean))

head(pred_df)
```

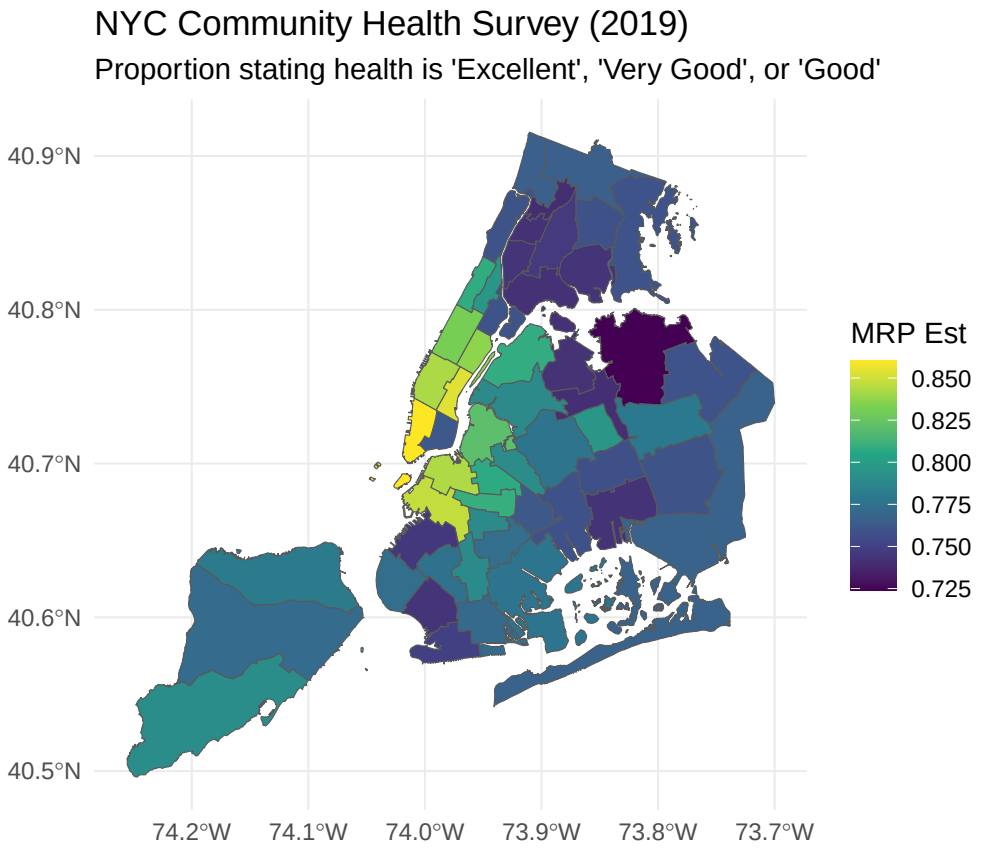
```
## # A tibble: 6 x 5
##   age   sex    race  edu           pred
##   <chr> <chr>  <chr> <chr>        <dbl>
## 1 18-24 female asian college      0.913
## 2 18-24 female asian hs           0.822
## 3 18-24 female asian less_hs      0.677
## 4 18-24 female asian some_college 0.847
## 5 18-24 female black college      0.956
## 6 18-24 female black hs           0.917
```

## Post Strat Estimates

```
# post stratify
mrp <-
  nyc_puma %>%
  mutate(PUMA = substr(PUMA, 2, 5)) %>%
  filter(PUMA %in% nyc_geo$puma) %>%
  left_join(pred_df) %>%
  mutate(mrp_est = n * pred) %>%
  group_by(PUMA) %>%
  summarise(prop = sum(mrp_est) / sum(n))
```

```
## Joining, by = c("age", "sex", "race", "edu")
```

```
# plot
nyc_geo %>%
  left_join(mrp, by = c("puma" = "PUMA")) %>%
  ggplot() +
  geom_sf(aes(fill = prop)) +
  scale_fill_viridis_c() +
  theme_minimal() +
  labs(title = "NYC Community Health Survey (2019)",
       subtitle = "Proportion stating health is 'Excellent', 'Very Good', or 'Good'",
       fill = "MRP Est")
```


NYC Community Health Survey (2019)
Proportion stating health is 'Excellent', 'Very Good', or 'Good'
40.9°N
40.8°N
40.7°N
40.6°N
40.5°N
74.2°W
74.1°W
74.0°W
73.9°W
73.8°W
73.7°W
MRP Est
0.850
0.825
0.800
0.775
0.750
0.725