17.1

Nikhil Gopal

2/27/2022

```
rm(list = ls())
library(rstanarm)
```

```
## Warning: package 'rstanarm' was built under R version 4.0.5
```

```
## Warning: package 'Rcpp' was built under R version 4.0.5
```

```
library(rosdata)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'stringr' was built under R version 4.0.4
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
library(tidyr)
```

17.1

Regression and poststratification: Section 10.4 presents some models predicting weight from height and other variables using survey data in the folder Earnings. But these data are not representative of the population. In particular, 62% of the respondents in this survey are women, as compared to only 52% of the general adult population. We also know the approximate distribution of heights in the adult population: normal with mean 763. inches and standard deviation 2.7 inches for women, and normal with mean 69.1 inches and standard deviation 2.9 inches for men.

- a:

Use poststratification to estimate the average weight in the general population, as follows: (i) fit a regression of linear weight on height and sex;

i:

```r
election <- rosdata::poll

data <- rosdata::earnings

fit <- stan_glm(weight~height+factor(male), data = data, refresh=0)

print(fit)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      weight ~ height + factor(male)
##  observations: 1789
##  predictors:   3
## ------
##                Median MAD_SD
## (Intercept)   -107.6   16.6
## height           3.9    0.3
## factor(male)1   11.9    2.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 28.7    0.5
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

ii:

(ii) use posterior_epred to make predictions for men and women for each integer value of height from 50 through 80 inches;

```r
p_male <- 1 - 0.52
pop1 <- data.frame(male = c(0, 1), N = c(1 - p_male, p_male))
pop2 <- NULL
cut <- c(0, seq(50, 80, 1), 999)
for (i in 1:(length(cut)-1)){
   # male height distribution in the population
   m_h <- pnorm(cut[i + 1], 69.1, 2.9) -  pnorm(cut[i], 69.1, 2.9)
   # female height distribution
   f_h <- pnorm(cut[i + 1], 63.7, 2.7) -  pnorm(cut[i], 63.7, 2.7)
   pop2 <- bind_rows(pop2,
                     data.frame(height = cut[i], male = 1, N = m_h),
                     data.frame(height = cut[i], male = 0, N = f_h))
}

pop2$N <- ifelse(pop2$male == 1, p_male * pop2$N, (1 - p_male) * pop2$N)

#expected value
epreds <- posterior_epred(fit, newdata = pop2)
```

iii:

(iii) poststratify using a discrete approximation to the normal distribution for heights given sex, and the known proportion of men and women in the population. Your result should be a set of simulation draws representing the population average weight. Give the median and mad sd of this distribution: this represents your estimate and uncertainty about the population average weight

```
poststrat2 <- epreds %*% pop2$N
median(poststrat2)
```

```
## [1] 154.4314
```

```
mad(poststrat2)
```

```
## [1] 0.7959909
```

The mean was 154.3979 and the mad sd was 0.7585279.

- b:

Repeat the above steps, this time including the height:female interaction in your fitted model before post-stratifying.

```
fit2 <- stan_glm(weight ~ height+factor(male)+ height:factor(male), data = data, refresh=0)

print(fit2)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      weight ~ height + factor(male) + height:factor(male)
##  observations: 1789
##  predictors:   4
## ------
##                       Median MAD_SD
## (Intercept)          -61.2   21.4
## height                 3.2    0.3
## factor(male)1        -96.4   33.0
## height:factor(male)1   1.6    0.5
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 28.6    0.5
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
pred2 <- posterior_epred(fit2, newdata = pop2)
poststrat2 <- pred2 %*% pop2$N
median(poststrat2)
```

```
## [1] 154.2513
```

```
mad(poststrat2)
```

```
## [1] 0.7822429
```

The mean was 154.2174 and the MAD sd was 0.7453965

- c:

Repeat (a) and (b), this time performing a regression of log(weight) but still with the goal of estimating average weight in the population, so you will need to exponentiate your predictions in step (ii) before poststratifying.

```
fit3 <- stan_glm(log(weight) ~ height+factor(male), data = data, refresh=0)


print(fit3)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      log(weight) ~ height + factor(male)
##  observations: 1789
##  predictors:   3
## ------
##               Median MAD_SD
## (Intercept)   3.4    0.1
## height        0.0    0.0
## factor(male)1 0.1    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.2    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
pred3 <- exp(posterior_epred(fit3, newdata = pop2))
poststrat3 <- pred3 %*% pop2$N
median(poststrat3)
```

```
## [1] 152.1383
```

```
mad(poststrat3)
```

```
## [1] 0.7118086
```

The mean was 152.1696 and the MAD sd was 0.7160717.