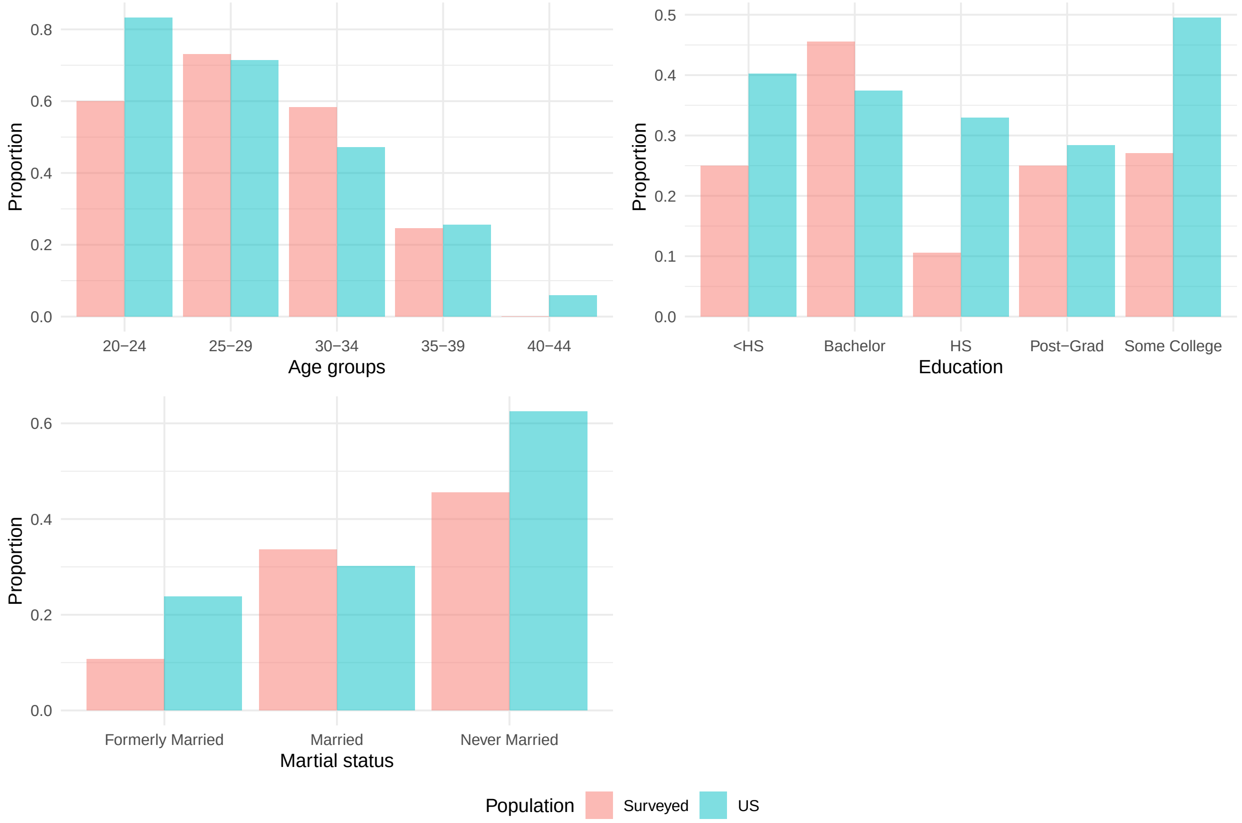
STA2201 Winter 2023 Assignment 3

Quynh Vu

2023-03-20

1. Fertility intentions: This question relates to a 2016 survey of US women who were asked about their future fertility intentions. The survey data is in the file `intentions_survey`. Also relevant to this question is the `us_pops` data file, which contains the number of women in the US in 2016 by age group, education and marital status. For this question, we are interested in obtaining estimates of p_a, which is the probability that a woman in age group a wants to have children in future, for all age groups $a = 1, \dots A$. In this case we have a total of $A = 5$ age groups (20-24, 25-29, 30-34, 35-39, 40-44).

a) Make a plot which compares the proportions surveyed women by age, education, and marital status to the same proportions in the overall US population. Briefly comment on what you observe.



The surveyed population doesn't seem to be representative of the US population (e.g., according to the survey, American women aged 25-29 or those with a bachelor's degree are the most willing to have kids in the future while in reality, it is women aged 20-24 or those who have earned some college experience.) Classifying across marital status yields more accurate estimates in the sense of a lower surveyed proportion corresponding to a lower true one, and vice versa.

b) Calculate the proportion of survey women in each age group that want to have children. We will refer to this set of estimates as $\hat{p}_a^{\text{raw}}$ for each age group a.

```
## # A tibble: 5 x 2
##   ‘Age groups‘ p_raw_hat
##   <fct>            <dbl>
## 1 40-44            0
## 2 30-34            0.583
## 3 35-39            0.246
## 4 25-29            0.731
## 5 20-24            0.6
```

c) Calculate the post-stratified estimates

$$\hat{p}_a^{\text{ps}} = \frac{\sum_{g=1}^{G} \hat{p}_{g[a]}^{\text{raw}} \times N_{g[a]}}{\sum_{g=1}^{G} N_{g[a]}}$$

where g refers to a particular education/marital status group (e.g. people who are married and have less than a high school degree). There are a total of $G = 5 \times 3 = 15$ groups within each age group. Note that $\hat{p}_{g[a]}^{\text{raw}}$ refers to the observed proportion of women in group g who are aged a who want more children and $N_{g[a]}$ refers to the size of that particular population group who are aged a in the US population.

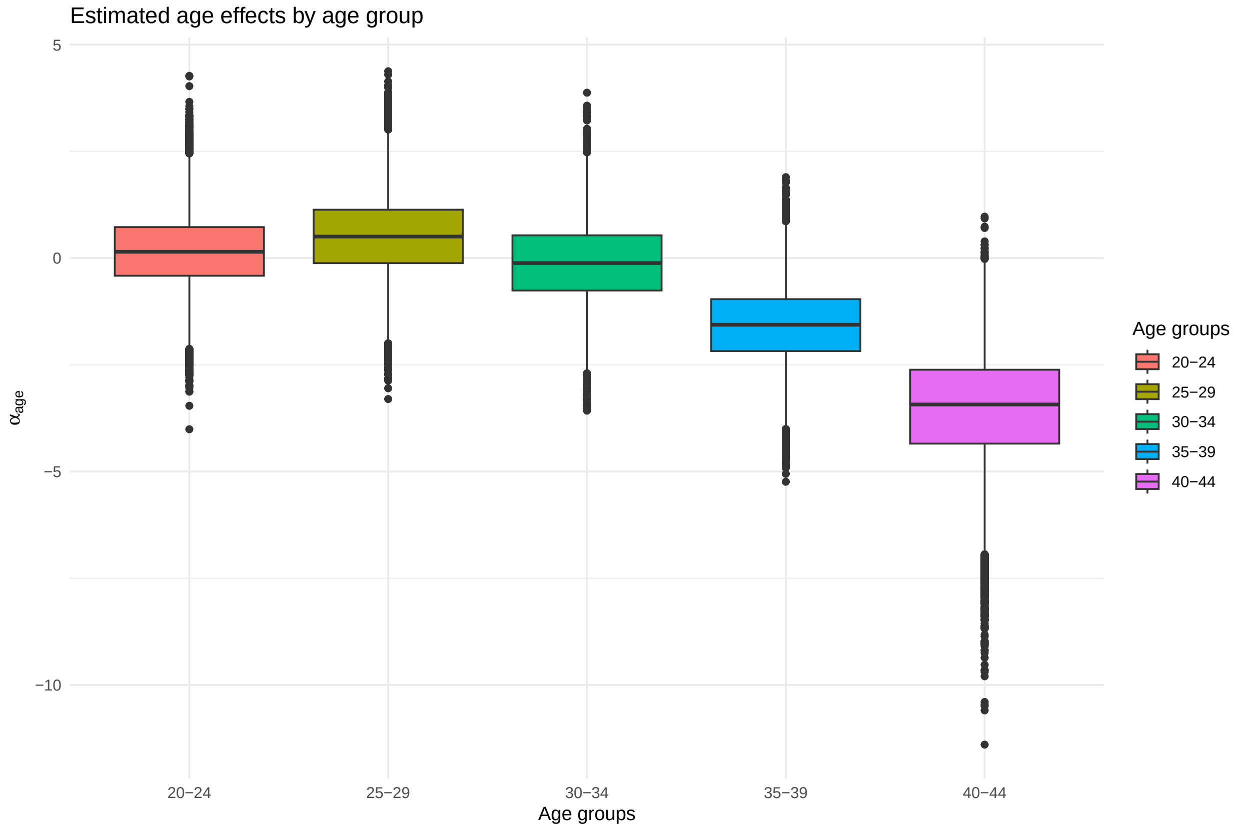
```
## # A tibble: 5 x 2
## # Groups:   age_gp [5]
##   age_gp  p_ps
##   <fct>  <dbl>
## 1 40-44  0
## 2 30-34  0.521
## 3 35-39  0.231
## 4 25-29  0.499
## 5 20-24  0.464
```

d) Fit the following hierarchical model

$$\begin{aligned}
y_i | \pi_i &\sim \text{Bern}\left(\pi_i\right) \\
\pi_i &= \text{logit}^{-1}\left(\beta_0 + \beta_1 \text{formerly married}_i + \beta_2 \text{married}_i + \alpha_{j[i]}^{\text{age}} + \alpha_{k[i]}^{\text{edu}}\right) \\
\alpha_j^{\text{age}} &\sim \text{N}\left(\alpha_{j-1}^{\text{age}}, \sigma_{\text{age}}^2\right), \text{ for } j = 2, \ldots, 5 \\
\alpha_k^{\text{edu}} &\sim \text{N}\left(0, \sigma_{\text{edu}}^2\right), \text{ for } k = 1, \ldots, 5
\end{aligned}$$

where $y_i = 1$ if respondent i wants more children and 0 otherwise, and the formerly married$_i$ and married$_i$ variables are indicator variables. Note you will need to specify priors on $\beta_0, \beta_1, \beta_2, \alpha_1^{\text{age}}$ and the variance parameters. Create a plot of the estimated age effects.

```
##                      mean     se_mean        sd        2.5%        25%
## beta[1]       0.080360573 0.009732231 0.8261906 -1.56042763 -0.4657865
## beta[2]      -0.971089606 0.004224775 0.5661160 -2.10565543 -1.3498871
## beta[3]      -0.007097997 0.002744109 0.3700727 -0.72061028 -0.2609302
## alpha_age[1]  0.152123735 0.010184831 0.8876350 -1.62313129 -0.4137400
## alpha_age[2]  0.516842602 0.011563496 0.9419063 -1.31408086 -0.1184749
## alpha_age[3] -0.109375583 0.011875270 0.9659533 -1.99339644 -0.7594877
## alpha_age[4] -1.572606303 0.011778647 0.9237957 -3.42740130 -2.1792918
## alpha_age[5] -3.535091773 0.014791132 1.3398159 -6.46065594 -4.3456147
## alpha_edu[1] -0.106881579 0.005628717 0.5844569 -1.30264581 -0.4536264
## alpha_edu[2] -0.660030686 0.006134934 0.6360220 -2.08434012 -1.0314348
## alpha_edu[3] -0.124352736 0.005598657 0.5044033 -1.10909535 -0.4372040
## alpha_edu[4]  0.950207968 0.006103078 0.5117508  0.05588385  0.6053990
## alpha_edu[5]  0.017836208 0.005873450 0.5884281 -1.15591802 -0.3398266
## sigma_age     1.312367727 0.003748317 0.4405820  0.62692590  0.9909640
## sigma_edu     0.829959374 0.004279955 0.3771912  0.27875432  0.5636449
##                      50%        75%      97.5%     n_eff      Rhat
## beta[1]       0.09559097  0.6245379  1.7198542  7206.687 1.0007811
## beta[2]      -0.95856031 -0.5816777  0.1030799 17955.757 0.9998258
## beta[3]      -0.01007426  0.2428636  0.7266364 18187.435 1.0001336
## alpha_age[1]  0.14711360  0.7260830  1.9368958  7595.586 1.0008653
## alpha_age[2]  0.50498016  1.1333721  2.4251294  6634.941 1.0009025
## alpha_age[3] -0.11687081  0.5331439  1.7996940  6616.453 1.0010590
## alpha_age[4] -1.56233138 -0.9623226  0.2245777  6151.217 1.0008870
## alpha_age[5] -3.42965647 -2.6156680 -1.2204559  8205.167 1.0005858
## alpha_edu[1] -0.10533407  0.2440877  1.0779456 10781.672 1.0000548
## alpha_edu[2] -0.60005217 -0.2287558  0.4437747 10747.923 1.0001211
## alpha_edu[3] -0.13036825  0.1729549  0.9177589  8116.861 1.0001135
## alpha_edu[4]  0.90869059  1.2567874  2.0684636  7031.034 1.0001378
## alpha_edu[5]  0.00187248  0.3650979  1.2508252 10036.914 1.0001337
## sigma_age     1.25640934  1.5661877  2.3365227 13815.950 1.0000918
## sigma_edu     0.76618398  1.0248699  1.7541575  7766.844 1.0006770
```



e) Calculate the multilevel-regression-with-post-stratification (MRP) estimates

$$\hat{p}_a^{\text{MRP}} = \frac{\sum_{g=1}^{G} \hat{p}_{g[a]}^{\text{MR}} \times N_{g[a]}}{\sum_{g=1}^{G} N_{g[a]}}$$

where $\hat{p}_{g[a]}^{\text{MR}}$ is the proportion of women in group g who are aged a who want more children estimated from your model in d). Report the median estimate of each $\hat{p}_a^{MRP}$ as well as the 95% CIs.

```
## # A tibble: 5 x 4
##   age_gp  p_MRP        LB    UB
##   <fct>   <dbl>     <dbl> <dbl>
## 1 40-44  0.0390 0.0000923 0.887
## 2 30-34  0.517  0.00894   0.994
## 3 35-39  0.186  0.00167   0.966
## 4 25-29  0.652  0.0220    0.995
## 5 20-24  0.580  0.0191    0.992
```

f) The true proportions of women wanting more children by age group are listed in `fertility_intentions_true`. Report the absolute difference by age group for each of the estimates $\hat{p}_a^{\text{raw}}$, $\hat{p}_a^{\text{ps}}$ and $\hat{p}_a^{\text{MRP}}$, as well as the mean absolute difference across all age groups. Comment on what you observe based on the relative performance of each of the estimation approaches.

When considering each age group separately, the Bayesian model yields the least accurate estimates of the proportion of women who want kids in a given age group and the true proportions overall, whereas the post-stratified estimates are the most accurate, except for the group of women aged 35-39. The p_{raw} estimates seem unstable. However, the raw estimates are closest to and the post-stratified estimate deviating the most from the true value on average across age groups, which makes sense since the Bureau data is representative of the population.

```
##   age_gp  p_true     p_raw      p_ps      p_MRP abs_diff_p_raw abs_diff_p_ps
## 1  40-44 0.05854 0.0000000 0.0000000 0.03902973     0.05854000    0.05854000
## 2  30-34 0.47106 0.5833333 0.5214560 0.51695600     0.11227333    0.05039599
## 3  35-39 0.25615 0.2459016 0.2312132 0.18595827     0.01024836    0.02493678
## 4  25-29 0.71420 0.7307692 0.4994255 0.65160296     0.01656923    0.21477448
## 5  20-24 0.83305 0.6000000 0.4637717 0.57962843     0.23305000    0.36927834
##   abs_diff_p_MRP
## 1     0.01951027
## 2     0.04589600
## 3     0.07019173
## 4     0.06259704
## 5     0.25342157
```

The mean absolute difference across all age groups:

```
##   mean_abs_diff_p_raw mean_abs_diff_p_ps mean_abs_diff_p_MRP
## 1          0.08613618          0.1435851          0.09032332
```