STA2201H Winter 2023 Assignment 3

Due: 11:59pm ET, March 31

What to hand in: .Rmd/.qmd file and the compiled pdf, and any stan files

How to hand in: Submit files via Quercus

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.

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.

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.

d) Fit the following hierarchical model

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

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.

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.

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.