STA2201H Winter 2024 Assignment 3

Due: 11:59pm ET, March 29

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

How to hand in: Submit files via Quercus

1 Abortion in Uganda

This question relates to a survey of Ugandan women who were asked about their abortion experiences. The survey data is in the file `pma`. Also relevant to this question is the `uganda_census` data file, which contains information about age, education, region of residence, and marital status obtained from the Ugandan census.

For this question, we are interested in obtaining estimates of p_a, which is the probability that a woman in age group a has had an abortion, for all age groups $a = 1, \dots A$. In this case we have a total of $A = 6$ age groups (15-19, 20-24, 25-29, 30-34, 35-39, 40-44).

a) Make a plot or plots which compare/s the proportions surveyed women by age, education, marital status and region of residence to the same proportions in census data. Briefly comment on what you observe.

b) Calculate the proportion of survey women in each age group that have had an abortion. 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/region group. There are a total of $G = 3 \times 4 \times 4 = 48$ 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 have had an abortion and $N_{g[a]}$ refers to the size of that particular population group who are aged a in the Ugandan population.

d) Define $y_i = 1$ if respondent i has had an abortion and 0 otherwise. Assume y_i is Bernoulli distributed with probability π_i. Write down a Bayesian logistic hierarchical model for π_i, including education, marital status, age, and region effects, with the age and region effects modeled hierarchically. For the age effect, you can decide on the structure of the model, justified based on what you saw in part a). Note you will need to specify priors on all parameters.

e) Fit the hierarchical model described above in Stan. Create a plot of the estimated age effects.

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

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

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

g) Plot the three different estimates for p_a on the same plot. Comment on what you observe and reflect briefly on the advantages and disadvantages of each estimation approach.

2 Research proposal

The final project for this class involves exploring a research question that you are interested in using a dataset of your choice. For the research proposal, I'm interested in finding out about your topic, and seeing some EDA based on your dataset of choice. Please describe

- your research question(s) of interest, and why they are of interest (if there's an obvious literature, feel free to cite a few papers)
- the dataset you plan to use
- your main dependent variable of interest
- your main independent variables of interest (including control variables)
- an indication of the methods/model you plan to use/run

2.1 Exploratory data analysis

As part of your research proposal please undertake some basic EDA to illustrate the characteristics of your dataset, patterns in the raw data, and to present descriptive statistics related to your data and your research question.

There is no set format, but here are a few pointers of things to look at

- **General characteristics of dataset** and **summary statistics of variables of interest**: for example, how many observations, how were the data collected (is the dataset representative of the population of interest?); you could present a table of summary statistics of main variables, including things like number of observations, mean/median/sd (if a continuous variable), proportions by group, etc…
- **Missing data**: If your dataset does not have any missing observations, then fine to just say this (don't need to do EDA graphs or discuss). If you have missing observations, summarize what is missing, and give a brief discussion about whether or not you think missingness may be a problem (e.g. is there more likely to be missing data for some groups compared to others?)
- **Graphs showing both univariate and bivariate patterns**: likely to be interested in both univariate patterns (e.g., the distribution of continuous variables, proportions for categorical outcomes…) and bivariate patterns (e.g. scatterplots, proportions/boxplots by group, trends over time…).

While you'll probably make a lot of graphs/summaries etc while doing EDA, you don't have to submit everything — just a few key observations. The proposal only needs to be 2-3 pages total (including graphs).

2.2 What to submit

It is expected that you present and write up your findings in Rmd. You should submit:

- your .Rmd/.qmd file; and
- the knitted PDF resulting from your .Rmd/.qmd file file.

If your dataset is reasonably small (and publicly available), then it would be great if you could submit that, too.

Please submit files via Quercus, in a separate document to your assignment.