STA2201H Winter 2022 Assignment 3

Due: 11:59pm ET, April 1

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

How to hand in: Submit files via Quercus

LAST UPDATED: 21 March (to clarify what N_g is in Q1)

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.

2 Sea level rise

This question is related to the `sea_level` dataset, which has measurements of relative sea levels over time for five different US states.[1] Variables of interest are:

- `rsl_m`: the relative sea level (RSL) (reported in meters). This is a measurement relative to the year in which the measurement is taken
- `rsl_se`: standard error of (RSL) measurement
- `indicator`: type of proxy that was used to reconstruct sea level
- `age_ce`: the year that the measurement refers to (CE = common era)

We will be using penalized splines regression to estimate and project sea levels over time.

a) Create a (faceted) plot that shows the available data for each state. Your plots should show the standard errors around the data points and also the different indicators that were used to get measurements. Briefly comment on what you observe.

b) Fit the following second-order penalized splines regression model in Stan over the period 1000-2010:

$$\begin{aligned} y_i &\sim N(\mu_{t[i],s[i]}, s_i^2) \\ \mu_{t,s} &= \sum_{k=1}^{K} B_{k,t}\alpha_{k,s} \\ \alpha_{1,s} &\sim N(0,1) \\ \alpha_{2,s} &\sim N(\alpha_{1,s}, \sigma_\alpha^2) \\ \Delta^2\alpha_{k,s} &\sim (0, \sigma_\alpha^2), \text{ for } k = 3, \ldots, K \\ \sigma_\alpha &\sim N_+(0,1) \end{aligned}$$

where

- y_i is ith observation of relative sea-level with $i = 1, \ldots, N$
- s_i is the standard error of relative sea-level of observation i
- t refers to the year
- s refers to the state
- k refers to the knot position with a total of K knots.
- $B_{t,k}$ refers to the value of a cubic basis spline at knot k for year t

To set your basis splines, you can use the `getsplines` function provided, with knot spacing $I = 50$ years.

Plot your resulting estimates and 95% CIs for the period 1000-2010 overlaying the data (i.e. on your graph from a)). Give some brief commentary on what you observe.

c) Use your model in b) to get projections for relative sea level in the five states up to the year 2100. Add these to your graph from part b). Comment briefly on what you observe.

[1] Thanks to Niamh Cahill for providing these data.

d) Different proxies have different levels of bias in their measurement of sea level. In particular, it can be assumed that the bias of measurements from foraminifera is zero. Adapt your model from b) to add a bias term for other proxies:

$$\begin{aligned}
y_i &\sim N(\mu_{t[i],s[i]} + \gamma_{p[i]}, s_i^2) \\
\mu_{t,s} &= \sum_{k=1}^{K} B_{k,t}\alpha_{k,s} \\
\alpha_{1,s} &\sim N(0,1) \\
\alpha_{2,s} &\sim N(\alpha_{1,s}, \sigma_\alpha^2) \\
\Delta^2\alpha_{k,s} &\sim (0, \sigma_\alpha^2), \text{ for } k = 3, \dots, K \\
\sigma_\alpha &\sim N_+(0,1) \\
\gamma_p &\sim N(0,1) \text{ for } p = 1, 2
\end{aligned}$$

where $p = 1$ if the indicator is peat and $p = 2$ if the indicator is plants. Report the estimates of bias for peat and plants (with 95% CIs). Plot estimates of sea level rise over time from this model and the original model on the same graphs by state and comment on differences.

3 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

3.1 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).

3.2 What to submit

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

- your R Markdown file; and
- the knitted PDF resulting from your R Markdown 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.