# STA2201H Winter 2021 Assignment 2

**Due:** 11:59pm ET, March 29 2021

**What to hand in:** .Rmd file and the compiled pdf

**How to hand in:** Submit files via Quercus

Note that at the end of this document there are details about the research proposal.

# 1 Wells

This question uses data looking at the decision of households in Bangladesh to switch drinking water wells in response to their well being marked as unsafe or not. A full description from the Gelman Hill text book (page 87):

*"Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases, with risks estimated to be proportional to exposure. Any locality can include wells with a range of arsenic levels. The bad news is that even if your neighbor's well is safe, it does not mean that yours is safe. However, the corresponding good news is that, if your well has a high arsenic level, you can probably find a safe well nearby to get your water from—if you are willing to walk the distance and your neighbor is willing to share. [In an area of Bangladesh, a research team] measured all the wells and labeled them with their arsenic level as well as a characterization as "safe" (below 0.5 in units of hundreds of micrograms per liter, the Bangladesh standard for arsenic in drinking water) or "unsafe" (above 0.5). People with unsafe wells were encouraged to switch to nearby private or community wells or to new wells of their own construction. A few years later, the researchers returned to find out who had switched wells."*

The outcome of interest is whether or not household $i$ switched wells:

$$y_i = \begin{cases} 1 & \text{if household } i \text{ switched to a new well} \\ 0 & \text{if household } i \text{ continued using its own well.} \end{cases}$$

The data we are using for this question are here: http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat and you can load them in directly using

```
d <- read.table(url("the_url_above"))
```

The variables of interest for this questions are

- `switch`, which is $y_i$ above
- `arsenic`, the level of arsenic of the respondent's well
- `dist`, the distance (in metres) of the closest known safe well

a) Do an exploratory data analysis illustrating the relationship between well-switching, distance and arsenic. Think about different ways of effectively illustrating the relationships given the binary outcome. As usual, a good EDA includes well-thought-out descriptions and analysis of any graphs and tables provided, well-labelled axes, titles etc.

Assume $y_i \sim Bern(p_i)$, where $p_i$ refers to the probability of switching. Consider two candidate models.

- Model 1:

$$\text{logit}\left(p_i\right) = \beta_0 + \beta_1 \cdot \left(d_i - \bar{d}\right) + \beta_2 \cdot \left(a_i - \bar{a}\right) + \beta_3 \cdot \left(d_i - \bar{d}\right)\left(a_i - \bar{a}\right)$$

- Model 2:

$$\begin{aligned}\text{logit}\left(p_i\right) =& \beta_0 + \beta_1 \cdot \left(d_i - \bar{d}\right) + \beta_2 \cdot \left(\log\left(a_i\right) - \overline{\log(a)}\right) \\ &+ \beta_3 \cdot \left(d_i - \bar{d}\right)\left(\log\left(a_i\right) - \overline{\log(a)}\right)\end{aligned}$$

where $d_i$ is distance and $a_i$ is arsenic level.

b) Fit both of these models using Stan. Put $N(0, 1)$ priors on all the $\beta$s. You should generate pointwise log likelihood estimates (to be used in later questions), and also samples from the posterior predictive distribution (unless you'd prefer to do it in R later on). For model 1, interpret each coefficient.

c) Let $t(\boldsymbol{y}) = \sum_{i=1}^{n} 1\left(y_i = 1, a_i < 0.82\right) / \sum_{i=1}^{n} 1\left(a_i < 0.82\right)$ i.e. the proportion of households that switch with arsenic level less than 0.82. Calculate $t(\boldsymbol{y^{rep}})$ for each replicated dataset for each model, plot the resulting histogram for each model and compare to the observed value of $t(\boldsymbol{y})$. Calculate $P\left(t\left(\boldsymbol{y}^{rep}\right) < t(\boldsymbol{y})\right)$ for each model. Interpret your findings.

d) Use the `loo` package to get estimates of the expected log pointwise predictive density for each point, $ELPD_i$. Based on $\sum_i ELPD_i$, which model is preferred?

e) Create a scatter plot of the $ELPD_i$'s for Model 2 versus the $ELPD_i$'s for Model 1. Create another scatter plot of the difference in $ELPD_i$'s between the models versus log arsenic. In both cases, color the dots based on the value of $y_i$. Interpret both plots.

f) Given the outcome in this case is discrete, we can directly interpret the $ELPD_i$s. In particular, what is $\exp(ELPD_i)$?

g) For each model recode the $ELPD_i$'s to get $\hat{y}_i = E\left(Y_i|\boldsymbol{y}_{-i}\right)$. Create a binned residual plot, looking at the average residual $y_i - \hat{y}_i$ by arsenic for Model 1 and by log(arsenic) for Model 2. Split the data such that there are 40 bins. On your plots, the average residual should be shown with a dot for each bin. In addition, add in a line to represent +/- 2 standard errors for each bin. Interpret the plots for both models.

## 2 Maternal mortality

This question relates to estimating the maternal mortality for countries worldwide. A maternal death is defined by the World Health Organization as "the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes". The indicator we are interested in is the (non-AIDS) maternal mortality ratio (MMR) which is defined as the number of non-AIDS maternal deaths divided by the number of live births.

In the data folder of the class repo there are two files relevant to this question. `mmr_data` contains information on, for a range of countries over a range of years:

- Observations of the proportion of non-AIDS deaths that are maternal ($PM^{NA}$)
- Data source, most commonly from Vital Registration systems (VR)
- The Gross Domestic Product (GDP)
- The General Fertility Rate (GFR)
- The average number of skilled attendants at birth (SAB)
- The geographical region of the country
- The total number of women, births, deaths to women of reproductive age (WRA), and the estimated proportion of all WRA deaths that are due to HIV/AIDS

The `mmr_data` file will be used for fitting. Note that data on $PM^{NA}$ is not available for every country.

The `mmr_pred` file contains information on GDP, GFR, SAB, total number of births, deaths and women, and proportion of deaths that are due to HIV/AIDS, for every country at different time points (every five years from mid 1985 to mid 2015). Information in this file is used for producing estimates of MMR for countries without data, and for producing estimates centered at a particular time point.

Consider the following model

$$\begin{aligned} y_i|\eta_{c[i]}^{\text{country}}, \eta_{r[i]}^{\text{region}} &\sim N\left(\beta_0 + \eta_{c[i]}^{\text{country}} + \eta_{r[i]}^{\text{region}} + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}, \sigma_y^2\right) \\ \eta_c^{\text{country}} &\sim N\left(0, \left(\sigma_\eta^{\text{country}}\right)^2\right), \text{ for } c = 1, 2, \ldots, C \\ \eta_r^{\text{region}} &\sim N\left(0, \left(\sigma_\eta^{\text{region}}\right)^2\right), \text{ for } r = 1, 2, \ldots, R \end{aligned}$$

where

- $y_i$ is the $i$th observed $\log PM^{NA}$ in country $c[i]$ in region $r[i]$
- $C$ is total number of countries and $R$ is total number of regions
- $x_{i,1}$ is log(GDP)
- $x_{i,2}$ is log(GFR)
- $x_{i,3}$ is SAB

a) Turn this model into a Bayesian model by specifying appropriate prior distributions for the hyper-parameters and fit the Bayesian model in Stan. Report the full model specification as well as providing the Stan model code.

Hint: I would recommend indexing countries and regions, and calculating $C$ and $R$ based on the full set of countries contained in `mmr_pred`, rather than the subset contained in `mmr_data`. This will mean you will automatically get estimates for $\eta$ for every country and region, even the missing ones, which will help later on.

b) Check the trace plots and effective sample size to check convergence and mixing. Summarize your findings using a few example trace plots and effective sample sizes.
c) Plot (samples of the) prior and posterior distributions for $\beta_0, \sigma_y, \sigma_\eta^{\text{country}}$ and $\sigma_\eta^{\text{region}}$. Interpret the estimates of $\beta_1$ and $\beta_3$.
d) Use the MCMC samples to construct 95% credible intervals for the $PM^{NA}$ for 5-year periods from 1985.5 to 2015.5 for one country with data and one country without any observed $PM^{NA}$ values. Provide point estimates and CIs in a table and a nice plot. Add the observed data to the plot as well (for the country that has it).
e) The non-AIDS MMR is given by

$$\begin{aligned} MMR^{NA} &= \frac{\#\text{ Non-AIDS maternal deaths}}{\#\text{ Births}} \\ &= \frac{\#\text{ Non-AIDS maternal deaths}}{\#\text{ Non-AIDS deaths}} \cdot \frac{\#\text{ Non-AIDS deaths}}{\#\text{ Births}} \\ &= PM^{NA} \cdot \frac{\#\text{ Deaths } * (1 - \text{ prop AIDS})}{\text{Births}} \end{aligned}$$

where deaths and births are to all women of reproductive age in the country-period of interest. Use this formula, your answers from d) and the data in `mmr_pred` to obtain point estimates and CIs for the non-AIDS MMR for the two countries you chose in (d) in the year 2010.5.
f) In the model used so far, we assume that error variance $\sigma_y^2$ is the same for all observations but this is probably not a very realistic assumption. Let's explore if the model fit changes if we would estimate two variance parameters: one for VR data (denoted by $\sigma^2_{\text{VR}}$) and one for non-VR data (denoted by $\sigma^2_{\text{non-VR}}$). Write out the model specification for this extended model, give the Stan model code, and fit the model. Show priors and posteriors for $\sigma_{\text{VR}}$ and $\sigma_{\text{non-VR}}$ and construct a plot with data for a country with VR data, with point estimates and CIs from the models with and without equal variance.

# 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)

## 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...).

## 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.