STA2201H Take-home exam

Due: 11:59pm, 14 April 2020

What to hand in: .Rmd file, the compiled pdf and Stan code

How to hand in: Submit files via Quercus

1 Spending behavior

The Canadian company KOHO provides financial services, such as spending accounts and a Visa card, through an online-only app. They have come to you with a modeling problem. They are interested in understanding how spending behavior has changed in light of the recent social distancing measures put in place in Toronto. In particular, among other things, KOHO are interested in the proportion of purchases made online, how this has changed since social distancing, and how changes differ by population subgroup (age and income level).

KOHO have given you data on every transaction made by their customers in Toronto from March 1 to April 6 2020. In addition to knowing the date and monetary amount of each transaction, you also know the broad category of type of purchase, the location of the purchase, and whether it was made online or not (1 if online and 0 otherwise). You also know the age group and income level of each customer. Assume there are four age groups (18-29, 30-44, 45-64, 65+) and four income groups (<$30,000; $30,000-$59,999; $60,000-$99,999; $100,000+)

Let's assume social distancing started on March 16, when UofT went online.

a) Introduce notation and specify a fully Bayesian model that could be used to understand the change in the proportion of online purchases since social distancing and differences by age group and income. Note I don't think there is only one right model set-up here, there are many interesting alternatives (both in terms of structure of covariates and what form of the outcome is modeled). That said, given the nature of the data, I would encourage you to formulate some sort of hierarchical model. You will need to specify and define notation for the outcome of interest and covariates, all with appropriate indexing, and specify the likelihood, group-level models, and priors. Explain how you would assess whether the average change in the proportion of online purchases is higher for 18-29 year olds in the second income bracket ($30,000-$59,999) versus 45-64 year olds in the same income bracket.

b) Based on the posterior samples on your model parameters, how would you estimate the expected change in the proportion of online purchases for people aged 18-29 in the first income bracket and construct a 95% credible interval? How would you predict the change in the proportion of online purchases and construct a 95% prediction interval for a group of 30 people aged 18-29 in the first income bracket? Show working with formulas and pseudo code, where appropriate.

Now let's focus on the period since social distancing started (i.e. March 16-April 6). In addition, to simplify things, let's consider the effect of age group only. Define y_i to be the number of online purchases for individual i, and n_i to be the total number of purchases for individual i.

KOHO would like to use their data to estimate the proportion of purchases made online for the whole of Toronto. The types of people who use KOHO are not really representative of the broader Toronto; in general they tend to be relatively young. However, all is not lost: from the last census we have population counts by age group in Toronto. So we could post-stratify to get a more representative estimate, i.e.

$$\hat{\pi}^{\text{ps}} = \frac{\sum_G \hat{\pi}_g N_g}{N}$$

where $\hat{\pi}^{\text{ps}}$ is the estimated proportion of purchases made online, g refers to a particular age group (i.e. there are $G = 4$ total age groups), $\hat{\pi}_g$ is the estimated proportion for a particular group, N_g is the number of people in a particular age group based on the census, and N is the total population based on the census.

c) Assume that differences in the propensity to make online purchases is fully captured by differences in age. Let π^* be the true proportion of purchases that are made online in Toronto during March 16-April 6. If $\hat{\pi}_g$ is taken to be the observed proportions from the data (i.e. the MLE), show that $\hat{\pi}^{\text{ps}}$ is an unbiased estimator of π^*.

d) Now instead of using raw proportions in the data, you estimate $\hat{\pi}_g^{\text{mr}}$ with a hierarchical model, allowing for a varying intercept α_g by age group, with α_g modeled hierarchically as a draw from a Normal distribution with mean μ_α and variance σ_α^2.

The estimate of the Toronto proportion is

$$\hat{\pi}^{\text{mrp}} = \frac{\sum_G \hat{\pi}_g^{\text{mr}} N_g}{N}$$

Find an expression for $E(\hat{\pi}_g^{\text{mr}}|\mathbf{y})$. You may assume that the n's are large enough such that the Normal approximation to the Binomial is fine.

e) Calculate the bias of $\hat{\pi}^{\text{mrp}}$ given the true proportion π^*. Given it's greater than zero, why would we prefer this estimator over the unbiased $\hat{\pi}^{\text{ps}}$? Discuss.

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 ith 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 η for every country and region, even the missing ones, which will help later on.

E.g. to get full list of country iso codes and regions, could do something like:

```
country_region_list <- mmr_pred %>%
  group_by(iso) %>%
  slice(1) %>%
  arrange(iso) %>%
  select(iso, region)

iso.c <- country_region_list$iso # the iso country of each country
C <- length(iso.c) # number of countries

region.c <- country_region_list$region # the region that country c belongs to (name)
regions <- unique(region.c) # a list of all unique regions
R <- length(regions) # number of regions

# the region index that country c belongs to
r.c <- as.numeric(factor(region.c, levels = regions))
```

Then to get the relevant indexes for each observation i:

```
 # the country of the ith observation
c.i <- as.numeric(factor(mmr_data$iso, levels = iso.c))
r.i <- r.c[c.i] # the region of the ith observation
```

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 β_1 and β_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 σ_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 σ^2_{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 σ_{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 Airbnb

In this question you will be exploring what factors are associated with nightly rates of accommodation listed on Airbnb in Toronto. In the data folder of the class repo there is a file called `airbnb`. This contains variables describing Airbnb listings in Toronto as of 7 December 2019. I downloaded this from the Inside Airbnb website: http://insideairbnb.com/get-the-data.html. I restricted the dataset in the repo to only contain a selection of all variables available, but other than that made no changes. The goal is to model `price` (or log(price) might be more appropriate).

a) Carry out EDA on this data set. Note that this should include checking the data for missing values, data quality, etc, as well as a descriptive analysis of the data (keeping in mind the modeling goal). As a start, you'll notice that the `price` column is a character, which isn't helpful. Here's some code to get it into a number:

```
airbnb <- airbnb %>%
  mutate(price = str_remove(price, "\\$"),
         price = str_remove(price, ","),
         price = as.integer(price)
         )
```

Note that for the next questions it is okay with me to remove some of the troublesome observations, e.g. I removed observations with missing values for variables like `review_scores_rating`. Just make sure you document what you're removing.

b) Propose two candidate Bayesian models for log(price) and fit the two models in Stan (note: you will need to calculate the log-likelihood for the next question). Make sure the model converged and mixing is fine by checking model diagnostics.

c) Using LOO-CV, determine which of your models is preferred.

d) For the preferred model, discuss the results with the help of good explanations and graphs. It's up to you what you highlight, but just note that discussing values of coefficients from `summary(mod)` is not enough.

e) Leave 20% of the data out at random (this is your test set). Rerun your preferred model on the remaining 80% of the data (this is the training set). Use the coefficient estimates to estimate the nightly rate for the each of the observations in the test set. The root mean squared error is:

$$RMSE = \sqrt{\frac{\sum_{i-1}^{N}(\hat{y}_i - y_i)^2}{N}}$$

where y_i is the observed log(price) and $\hat{y}_i$ is the estimated log(price). Calculate the RMSE for the entire test set and also by room type. Briefly comment.

4 Short questions that are unrelated to each other

a) Show that if survival times are exponentially distributed, that the gamma distribution is the conjugate prior for the unknown hazard.

b) Specify the likelihood function if you only observe $\bar{y} = 1/2(y_1 + y_2)$ where $y_i \sim N(\mu, \sigma^2)$ and $Cor(y_1, y_2) = \rho \neq 0$.