STA2201H Winter 2020 Assignment 3

Due: 11:59pm, 15 March 2020

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

How to hand in: Submit files via Quercus.

1 IQ

Scoring on IQ tests is designed to produce a normal distribution with a mean of 100 and a standard deviation of 15 when applied to the general population. Now suppose we are to sample n individuals from a particular town and then estimate μ, the town-specific mean IQ score, based on the sample of size n. Let Y_i denote the IQ score for the ith person in the town of interest, and assume

$$Y_1, Y_2, \ldots, Y_n | \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$$

For this question, will assume that the standard deviation of the IQ scores in the town is equal to 15, then mean is equal to 113 and the number of observations is equal to 10. Additionally, for Bayesian inference, the following prior will be used:

$$\mu \sim N\left(\mu_0, \sigma^2_{\mu 0}\right)$$

with $\mu_0 = 100$ and $\sigma^2_{\mu 0} = 15$.

a) Write down the posterior distribution of μ based on the information above. Give the Bayesian point estimate and a 95% credible interval of μ, $\hat{\mu}_{Bayes} = E(\mu|\boldsymbol{y})$.

We will now compare the sampling properties of the Bayes estimator to the sample mean, which is the ML estimator.

b) Suppose that (unknown to us) the true mean IQ score is μ^*. To evaluate how close an estimator is to the truth, we might want to use the mean squared error (MSE) $\text{MSE}\left[\hat{\mu}|\mu^*\right] = E\left[\left(\hat{\mu} - \mu^*\right)^2 |\mu^*\right]$. Show the MSE is equal to the variance of the estimator plus the bias of the estimator squared, i.e.

$$\text{MSE}\left[\hat{\mu}|\mu^*\right] = \text{Var}\left[\hat{\mu}|\mu^*\right] + \text{Bias}\left(\hat{\mu}|\mu^*\right)^2$$

c) Suppose that the true mean IQ score is 112. Calculate the bias, variance and MSE of the Bayes and ML estimators. Which estimator has a larger bias? Which estimator has a larger MSE?

d) Write down the sampling distributions for the ML and Bayes estimates, again assuming $\mu^* = 112$ and $\sigma = 15$. Plot the two distributions on the one graph. Summarize your understanding of the differences in bias, variance and MSE of the two estimators by describing how these differences relate to differences in the sampling distributions as plotted. To further illustrate the point, obtain the Bayes and ML MSEs for increasing sample sizes and plot the ratio (Bayes MSE)/(ML MSE) against sample size.

2 Gibbs Sampling

a) Suppose the parameter vector of interest θ has been divided into d components $\theta = (\theta_1, \ldots, \theta_d)$. In Gibbs Sampling, each θ_j^s at iteration s is sampled from the conditional distribution given all the other components of θ, i.e. $p\left(\theta_j | \theta_{-j}^{s-1}, y\right)$, where θ_{-j}^{s-1} is all the components of θ except for j at their current values:

$$\theta_{-j}^{s-1} = \left(\theta_1^s, \ldots, \theta_{j-1}^s, \theta_{j+1}^{s-1}, \ldots, \theta_d^{s-1}\right)$$

Write down an expression for the proposal distribution $J(\theta^* | \theta^{s-1})$ for the Gibbs sampler and show that Gibbs sampling is a special case of the Metropolis Hastings algorithm with $r = 1$.

b) This question relates to the IQ example above. Let's make things a bit more realistic and assume the observed standard deviation is 13. Assume the observed sample mean is still 113.

We will use the following priors to estimate both μ and σ in a Bayesian model:

$$\mu \sim N\left(\mu_0, \sigma_{\mu 0}^2\right)$$
$$1/\sigma^2 \sim \text{Gamma}\left(v_0/2, v_0/2 \cdot \sigma_0^2\right)$$

Let's set $\mu_0 = 100$, $\sigma_{\mu 0} = \sigma_0 = 15$ and $v_0 = 1$.

Use Gibbs sampling in R to obtain posterior samples for μ and for σ. Notes:

- you don't need to derive the full conditionals if you don't want to, can just use the expressions in lecture notes
- use sample mean and precision for initial values
- obtain 1000 samples
- code should be well commented so it's clear what is going on

Output required:

- trace plots for μ and for σ
- histograms of posterior samples for μ and for σ
- point estimate and 95% CI for μ and for σ

3 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:

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

- Model 2:

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

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 βs. 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 $\hat{y}_i = E\left(Y_i = 1 | \boldsymbol{y}_{-i}\right)$. 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.