STA2201H Winter 2023 Assignment 2

Due: 11:59pm ET, March 4

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

How to hand in: Submit files via Quercus

1 IQ

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 onserved standard deviation of the IQ scores in the town is equal to 15, the observed mean is equal to 113 and the number of observations is equal to 10. Additionally, for Bayesian inference, the following prior will be used:

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

with $\mu_0 = 100$ and $\sigma_{\mu 0} = 15$.

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 estimates. Which estimate has a larger bias? Which estimate 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 Gompertz

Gompertz hazards are of the form

$$\mu_x = \alpha e^{\beta x}$$

for $x \in [0, \infty)$ with $\alpha, \beta > 0$. It is named after Benjamin Gompertz, who suggested a similar form to capture a 'law of human mortality' in 1825.

This question uses data on deaths by age in Sweden over time. The data are in the `sweden` file in the class repo. I grabbed the data from the Human Mortality Database.

We will assume that the deaths we observe in a particular age group are Poisson distributed with a rate equal to the mortality rate multiplied by the population, i.e.

$$D_x \sim \text{Poisson}(\mu_x P_x)$$

where x refers to age. In this question we will be estimating mortality rates using the Gompertz model as described above.

a) Describe, with the aid of a couple of graphs, some key observations of how mortality above age 50 in Sweden has changed over time.
b) Carry out prior predictive checks for α and β, based on populations by age in Sweden in 2020. Summarize what you find and what you decide to be weakly informative priors for these parameters.
c) Fit a model in Stan to estimate α and β for the year 2020. Note that it may be easier to specify the likelihood on the log scale (you can do this in Stan using the `poisson_log` function). Priors should be informed by your prior predictive checks and any other information available. Ensure that the model has converged and other diagnostics are good. Interpret your estimates for α and β.
d) Carry out some posterior predictive checks to assess model performance.

e) Now extend your model to estimate α and β in every year over the interval 1990-2020. Plot the resulting point estimates and 95% credible intervals for your estimates of α and β over time. Comment briefly on what you observe.
f) Life expectancy at age x is defined as

$$\int_x^\omega e^{-\mu_a} da$$

where ω is the oldest age group (you may assume this is age 100). Life expectancy is the expected number of years of life left at age x. The integral can be approximated by summing over discrete age groups. Based on your estimates in the previous question, estimate life expectancy at age 40 (note starting age!) for every year from 1990-2020. Plot your resulting point estimates and 95% credible intervals over time and comment briefly.

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 `read_table`.

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