STAT 479: Introduction to Bayesian Data Analysis
Ad popularity

October 25, 2021

Overview

Using data from Five Thirty Eight and from data collected from YouTube, this project will investigate what makes Super Bowl ads popular. Specifically, for each ad in the dataset, there are seven categorical predictors: The predictors are

- X_1: whether the ad was funny
- X_2: whether the ad showed the product quickly
- X_3: whether the ad was patriotic
- X_4: whether the ad featured a celebrity
- X_5: whether the ad involved danger
- X_6: whether the ad featured animals
- X_7: whether the ad used sex

The team will also collect the numbers of likes, dislikes, and views each ad received on YouTube.

One can try to predict the rate at which viewers react to each video. To this end, let Z_i be the total number of likes and dislikes received by the i^{th} video and let V_i be the total number of views (in the 10,000's)[1]. One could then model

$$Z_i|\lambda_i \sim \text{Poisson}(V_i \times \lambda_i)$$

[1] So $V_i = 0.3$ means that the video has $0.3 \times 10,000 = 3000$ total views

where λ_i captures the mean number of likes or dislikes per 10,000 viewers.

We are interested in understanding how λ_i varies with respect to each of the categorical predictors. To this end, we can model:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_7 x_{i7}$$

where $x_{i1}, \ldots, x_{i7}$ are the specific values of the predictors $X_1, \ldots, X_7$ for the i^{th} ad.

A critical component of this project will be to elicit reasonable priors for $\beta_1, \ldots, \beta_7$. This can be done, for instance, using prior predictive checks in which you (i) draw parameters from their respective prior distributions, (ii) simulate the dataset, and (iii) compare summary statistics of the simulated dataset to the actual observed dataset. If you have specified a prior which leads to prior predictive summaries that are extremely mis-aligned with the observed summaries, then you should adjust the prior accordingly.

Another critical component of this project is to assess how well the specified model is able to predict engagement for new ads. One way to assess this is to fit the model to $\sim 75\%$ of the observed data and then to look at the predictive performance on the held-out $\sim 25\%$ of the data. Specifically you can look at the posterior predictive distributions of Z_i for the held-out data and look at the mean square prediction error and also whether the 95% posterior predictive intervals contain the actual observed values of Z_i.

You may also think about fitting similar models to the numbers of likes and dislikes and comparing the parameters across models.