STA 290 — Homework 9

Final Project

Please include your C code in an appendix.

1. Returning to the boxing data of HW #6, we now explore a Bayesian method (that of Johnson, 1996, yes Val Johnson) for comparing the possible biases of the judges, while still accounting for round effects.

   Under this model, each round i has a true but unobserved outcome, y_i, which measures the extent to which fighter 2 (e.g. Holyfield) beats fighter 1 (e.g. Lewis), which is scaled so that $y_i \sim N(0,1)$ (the scale is arbitrary), and $y_i > 0$ if Holyfield outperformed Lewis, $y_i < 0$ if Lewis beat Holyfield (and the event of a draw has probability zero). However, each judge perceives the true y_i differently, perhaps because they had a different view of the fight (the judges sit on different sides of the ring from each other), perhaps because they have an inherent personal bias in favor of one fighter, perhaps because they blinked during a crucial blow. The round performance observed by judge j in round i is y_{ij}, where the judge will score the round for fighter 1 if and only if $y_{ij} > 0$, i.e., $x_{ij} = I_{\{y_{ij}>0\}}$. We model $y_{ij} \sim N(y_i + b_j, \sigma_j^2)$, where b_j is the bias of judge j in favor of fighter 2 (e.g. Holyfield). We then put a $N(0, \tau^2)$ prior on each b_j. One could put priors on σ_j and/or τ, but for this exercise, we will keep things simpler and just pick numbers for them, in particular, $\sigma_j^2 = 0.04$ for all j and $\tau^2 = 0.1$. Thus the full model is:

$$\begin{aligned} x_{ij}|y_{ij} &= I_{\{y_{ij}>0\}} \\ y_{ij}|b_j, y_i &\sim N(y_i + b_j, 0.04) \\ b_j &\sim N(0, 0.1) \\ y_i &\sim N(0,1) \end{aligned}$$

   (a) Show that the complete conditional for each y_{ij} is a truncated normal, and that the complete conditionals for each y_i and b_j are normal.

   (b) Generate samples from the posterior using Gibbs sampling. An easy way to sample from a truncated normal is to generate samples from an untruncated normal with the same parameters, and take the first sample that is within the truncated region, i.e. if you have a N(0,1) constrained to be greater than 0.5, then you repeated sample N(0,1)'s until you get the first one that is greater than 0.5 and you use that one as your one draw from the truncated distribution.

   (c) Plot your MCMC runs for all y_i's and b_j's and try just a few y_{ij}'s. Discuss your plots.

   (d) Perform any further analysis you think is particularly relevant and describe your findings in the boxing context. Do any judges appear to be significantly different than the others?

2. A survey was taken in one of the large introductory–level statistics courses here at Duke. The following questions were asked (actually, there were more, but those are not included in the data here):

    i. How tall are you, to the nearest inch?
    ii. Are you male or female? (coded as female=1, male=2)
    iii. Compared to the other students at this school, how do you consider your driving skills? (1=better than average, 2=avg, 3=below average)
    iv. How much do you weigh, to the nearest pound?
    v. What is your ideal weight, to the nearest pound?
    vi. How many minutes in an average week do you spend exercising?
    vii. How many books did you read last year for pleasure?
    viii. How many states in the US have you visited?
    ix. What is your resting pulse? (beats/minute)

    Analyze this dataset and report what you find out about this class. You should do both exploratory data analysis as well as fit formal models, i.e., show me that you learned something from this class (or that you already knew it). Note that this is real data, so that some cleaning of the data may be necessary.