STA 290 — Homework 8

Final Project — due Dec. 7 at 5PM

Please include your C code in an appendix.

1. Does pollution kill people? (This dataset is from Ramsey and Schafer (2002), p. 369.) For each of 60 cities, the following data were collected in the early 1960's:

   (a) Mortality rates per 100,000

   (b) Meteorological Variables

      i. Precipitation (mean annual in inches)

      ii. Humidity (annual average percent relative humidity at 1 PM)

      iii. January temp (mean January temperature in degrees Fahrenheit)

      iv. July temp (mean July temp in Fahrenheit)

   (c) Demographic Variables

      i. Over 65 (percent of population 65 years old and older)

      ii. Population per household

      iii. Education (median number of school years completed by persons of age 25 years or more)

      iv. Housing (percent of housing that is sound with all facilities)

      v. Density (population density in persons per square mile of urban area)

      vi. Nonwhite (percent of 1960 population that is nonwhite)

      vii. White collar (percent employment in white-collar occupations)

      viii. Poor (percent of households with annual income under $3,000 in 1960)

   (d) Relative Pollution Potentials

      i. Hydrocarbons

      ii. Nitrogen oxides

      iii. Sulfur dioxide

   The data are available on the course web page, at
   `http://www.stat.duke.edu/courses/Fall00/sta290/datasets/mortality.dat`

   Read in the data (this will not be a simple task), conduct exploratory data analyses, and then model the relationship between mortality and the pollution variables, using the other variables as necessary to control for differences in the data. Find an appropriate model and discuss your findings, being sure to address the question of interest. Note that for this problem, you do not need to do any Bayesian analysis.

2. Returning to the boxing data of HW #5, we now explore a Bayesian method (that of Johnson, 1996) 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 repeatedly 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 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?