STAT 578 – Advanced Bayesian Modeling – Fall 2019

Assignment 6

File `illinimensbb.csv`, in comma-separated values (CSV) format, contains 2018–2019 season statistics and roster information for fifteen Illini men's basketball players. The first column contains the jersey number, and the remaining columns contain player name, height (`Ht`, inches), position (`Pos`, `C`=center, `F`=forward, `G`=guard), minutes of playing time (`MIN`), field goals made (`FGM`), field goals attempted (`FGA`), and number of shots blocked (`BLK`).

You will build a logistic regression model for field goals, and a Poisson loglinear regression model for shots blocked, using JAGS and `rjags`.

1. [2 pts] Using `plot(Ht ~ Pos, data= ⋯)`, display box plots of height by position. Is there a relationship between height and position? (Such a relationship might cause substantial posterior correlations between regression coefficients if both height and position are used as explanatory variables.)

2. Let y_i be the number of field goals made by player i out of n_i attempts ($i = 1, \ldots, 15$). Consider the following logistic regression (with implicit intercept) on player position and height:

$$\begin{aligned} y_i \mid p_i &\sim \text{indep. Bin}(n_i, p_i) \\ \text{logit}(p_i) &= \beta_{\text{Pos}(i)} + \beta_{\text{Ht}} H_i \end{aligned}$$

   where

$$\begin{aligned} \text{Pos}(i) &= \text{player } i \text{ position (C, F, G)} \\ H_i &= \text{player } i \text{ height after } \textit{centering} \text{ and } \textit{scaling} \text{ to sample standard dev. 0.5} \end{aligned}$$

   Consider the prior

$$\beta_{\text{C}}, \beta_{\text{F}}, \beta_{\text{G}} \sim \text{iid}\, t_1\big(0, 10^2\big) \qquad \beta_{\text{Ht}} \sim t_1\big(0, 2.5^2\big)$$

   (a) [2 pts] List an appropriate JAGS model. Include nodes for the vector of binomial probabilities p_i and a vector y^{rep} of replicate responses.

   Now run your model using `rjags`. Make sure to use multiple chains with overdispersed starting points, check convergence, and monitor the regression coefficients, probabilities, and replicate responses (after convergence) long enough to obtain effective sample sizes of at least 4000 for each regression coefficient.

   (b) [2 pts] Display the `coda` summary of the results for the monitored regression coefficients.

   (c) [2 pts] With your posterior samples, display scatterplots of (i) β_{C} versus β_{Ht}, (ii) β_{F} versus β_{Ht}, and (iii) β_{G} versus β_{Ht}. Do you see (posterior) correlations?

(d) [2 pts] Consider the modeled probability that Ayo Dosunmu (No. 11) successfully makes an attempted field goal. Plot the (approximate) posterior density of this probability.

(e) [2 pts] Approximate the posterior probability that $\beta_{\mathrm{F}} > \beta_{\mathrm{G}}$ (i.e., that forwards have a higher probability of successfully making an attempted field goal than guards, after adjusting for height). Also, approximate the Bayes factor favoring $\beta_{\mathrm{F}} > \beta_{\mathrm{G}}$ versus $\beta_{\mathrm{F}} < \beta_{\mathrm{G}}$. (Note that, by symmetry, $\beta_{\mathrm{F}} > \beta_{\mathrm{G}}$ and $\beta_{\mathrm{F}} < \beta_{\mathrm{G}}$ have equal prior probability.) What can you say about the data evidence that $\beta_{\mathrm{F}} > \beta_{\mathrm{G}}$?

(f) [2 pts] Use the *chi-square discrepancy* to compute an approximate posterior predictive p-value. Does it indicate any evidence of problems (such as overdispersion)?

(g) Now consider expanding the model to allow for overdispersion, as follows:

$$\operatorname{logit}(p_i) \quad = \quad \beta_{\mathrm{Pos}(i)} \; + \; \beta_{\mathrm{Ht}}\, H_i \; + \; \varepsilon_i$$

with

$$\varepsilon_i \mid \sigma_\varepsilon \;\sim\; \text{iid } \mathrm{N}\big(0, \sigma_\varepsilon^2\big) \qquad\qquad \sigma_\varepsilon \;\sim\; \mathrm{U}(0, 10)$$

and everything else the same as before.

(i) [3 pts] List an appropriately modified JAGS model.

Then run it using `rjags`, with all of the usual steps.

(ii) [1 pt] Plot the (approximate) posterior density of σ_ε.

(iii) [2 pts] Repeat part (e) under this expanded model. Does your conclusion change?

3. Let y_i be the number of shots blocked by player i ($i = 1, \ldots, 15$). Consider the following Poisson loglinear regression (with implicit intercept) on player position and height, using minutes of playing time as a rate (exposure) variable:

$$y_i \mid r_i, t_i \;\sim\; \text{indep. Poisson}(t_i r_i)$$

$$\log(r_i) \quad = \quad \beta_{\mathrm{Pos}(i)} \; + \; \beta_{\mathrm{Ht}}\, H_i^*$$

where

$$\begin{aligned} t_i &= \text{player } i \text{ total minutes of playing time} \\ \mathrm{Pos}(i) &= \text{player } i \text{ position (C, F, G)} \\ H_i^* &= \text{player } i \text{ height after } \textit{standardizing} \\ &\quad\;\text{(centering and scaling to sample standard dev. 1)} \end{aligned}$$

(Note that the scaling of H_i^* is different than that of H_i in the previous part.)

Consider the prior

$$\beta_{\mathrm{C}}, \beta_{\mathrm{F}}, \beta_{\mathrm{G}}, \beta_{\mathrm{Ht}} \;\sim\; \text{iid } \mathrm{N}\big(0, 100^2\big)$$

(a) [2 pts] List an appropriate JAGS model. Include nodes for the vector of Poisson means $\lambda_i = t_i r_i$ and a vector y^{rep} of replicate responses.

Now run your model using `rjags`. Make sure to use multiple chains with overdispersed starting points, check convergence, and monitor the regression coefficients, Poisson means, and replicate responses (after convergence) long enough to obtain effective sample sizes of at least 4000 for each regression coefficient.

(b) [2 pts] Display the `coda` summary of the results for the monitored regression coefficients.

(c) [2 pts] The sampling model implies that

$$e^{\beta_{\text{Ht}}}$$

represents the factor by which the mean rate of blocking shots changes for each increase in height of one standard deviation (here, about 3.5 inches). (Under the model, this factor is the same for all positions.) Form an approximate 95% central posterior credible interval for this factor. According to your interval, does it seem that greater height is associated with a higher rate of blocking shots?

(d) [2 pts] Use the *chi-square discrepancy* to compute an approximate posterior predictive p-value. Does it indicate any evidence of problems?

(e) For each player (i), approximate $\Pr\left(y_i^{\text{rep}} \geq y_i \mid y\right)$, which is a kind of marginal posterior predictive p-value.

(i) [2 pts] Show your R code, and display a table with the player names and their values of this probability.

(ii) [1 pt] Name any players for whom this probability is less than 0.05. (Any such player blocked notably more shots than the model would suggest, for his position and height.)

(iii) [1 pt] Notice that the probability equals 1 for some players. Why is that actually *not* surprising? (Hint: How many shots were actually blocked by those players? How much playing time did they have?)

Total: 32 pts