Final Exam Question 4 (3 pages)

The English Premier League (EPL) is the premier soccer (football to the rest of the word) competition in England. 20 teams play in the competition. Each team plays every other team both at home and on the road (or, to use English terminology, away), for a total of 38 games per team, and 380 total games. Each season runs from August to May.

So far in the 2016/17 season, there have been 140 games played, with each team playing 14 games. The file epl.csv contains the score data from the 140 games played. The data set contains 6 columns: the first column is the name of the home team, the second column contains an identifier for the home team (a number between 1 and 20, see table below), the third column is the number of goals scored by the home team in that game. Columns 4, 5, and 6, mirror the information in columns 1, 2, and 3, but for the away team.

Team ID	Team
1	Arsenal
2	Burnley
3	Bournemouth
4	Chelsea
5	Crystal Palace
6	Everton
7	Hull City
8	Leicester City
9	Liverpool
10	Manchester City
11	Manchester United
12	Middlesbrough
13	Southampton
14	Stoke City
15	Sunderland
16	Swansea City
17	Tottenham Hotspur
18	Watford
19	West Bromwich Albion
20	West Ham United

We would like to fit a model to the data in order to analyze the quality of each team. Since goals scored can be thought of as counts, a Poisson regression model may be suitable.

Let H_{ij} be the number of goals scored by the home team (team i) in a game between team i and team j played at team i's home. Likewise, let A_{ij} be the number of goals scored by the away team (team j) in the same game.

Now suppose that

$$H_{ij} \sim Poisson(\lambda_{ij})$$
$$A_{ij} \sim Poisson(\theta_{ij})$$

where

$$\log(\lambda_{ij}) = \mu + \alpha_i + \delta_j + \gamma$$
$$\log(\theta_{ij}) = \mu + \alpha_j + \delta_i$$

Interpreting the paramters, μ is a measure of the average goal scoring rate, α_i and α_j are measures of the attacking ability of teams i and j respectively, δ_i and δ_j are measures of the defensive ability of teams i and j respectively, and γ is a measure of 'home-field advantage'. Since the α's and δ's measure each team's attacking and defensive ability relative to the average, we also need the following restrictions:

$$\sum_{k=1}^{20} \alpha_k = 0, \ \sum_{k=1}^{20} \delta_k = 0$$

(a) Load the data from the file, using `read.csv("epl.csv", header=T)`. Write a JAGS model file for the above model, and use in to simulate values from the posterior distribution of the parameters.

(b) Using the posterior means as a judge, which is the best offensive team? Which is the worst offensive team? Which is the best defensive team? Which is the worst defensive team?

(c) Rather than using the posterior means to evaluate the offensive and defensive quality of the teams, we could look at where each team's α and δ values rank compared to all the other teams. In JAGS, the `rank` function takes a vector and returns another vector whose entries are the size ranks of the the elements of the input vector, with 1 indicating that this was the smallest element, 2 indicating the next smallest, and so on.

In your JAGS model, add a line defining a vector of parameters called `rank.d` so that the kth element of rank.d, rank.d[k] is the rank of team k's defensive strength compared to all other teams, with 1 indicating they were the best defensive team (δ_k is the smallest of all δ's), and 20 indicating that they were the worst defensive team (δ_k is the biggest of all δ's). Using the posterior mean ranks for each team, which is the best defensive team? Which is the worst defensive team?

(d) In your JAGS model, add a line defining a vector of parameters called `rank.a` so that the kth element of rank.a, rank.a[k] is the rank of team k's offensive strength compared to all other teams, with 1 indicating they were the best offensive team (α_k is the biggest of all α's), and 20 indicating that they were the worst offensive team (α_k is the smallest of all α's). Using the posterior mean ranks for each team, which is the best offensive team? Which is the worst offensive team?

(e) Yet another way to evaluate the teams, would be to see in what proportion of simulations were the team's parameters ranked in a particular spot. In your JAGS model, add code to define `ranks.d[k,m]` which takes the value 1 is the defensive rank of team k is equal to m, and 0 otherwise (hint, check out the `equals` function in JAGS). This should be done for all teams and all possible ranks. With `ranks.d[k,m]` defined this way, the posterior mean of `ranks.d[k,m]` would be equal to the proportion of simulations in which team k's defensive rank was equal to m. Which team had the highest proportion of rank 1 for the defensive parameter? Which team had the highest proportion of rank 20 for the defensive parameter?

(f) Repeat part (e) for the offensive parameter, defining `ranks.a[k.m]`.

Show all working.