Project Name - Basketball Analytics using Statistical Inference

Project Description. (Fitting a Bradley-Terry model)

The GitHub repository NBA Data contains the results of all 1230 NBA games from the 2015-2016 regular season. The 30 teams are encoded numerically from 1 to 30; the key for this encoding is provided in the file *NBA_teams.txt*. Each row of *NBA_record.csv* indicates the home team, away team, and outcome Y for one game, where $Y = 1$ if the home team won and $Y = 0$ otherwise. For parts (a) and (b), you may *not* use an existing software implementation of the Bradley-Terry or logistic regression model; however, you may use any generic optimization or equation-solving routine (or you may implement the Newton-Raphson iterations yourself, if you are brave).

(a) Fit the Bradley-Terry model, with an intercept term α for the home-court advantage, to this data set. What are the 8 teams (in ranked order) with the highest Bradley-Terry scores? How much greater are the log-odds of winning for the home team than for the away team?

One approach to do this in R is to use the generic optimization function optim. To do this, first define a function

```
loglik = function(theta,Home,Away,Y) {
         ...
   }
```

that returns the log-likelihood for the Bradley-Terry model given inputs $\theta = (\alpha, \beta_2, \ldots, \beta_k)$ (constraining $\beta_1 \equiv 0$), Home $= (i_1, \ldots, i_n)$, Away $= (j_1, \ldots, j_n)$, and Y $= (Y_1, \ldots, Y_n)$, where i_m and j_m are the home and away teams for game m. To then read the data file and maximize the log-likelihood:

```
table = read.csv("NBA_record.csv")
result = optim(theta0,loglik,Home=table$Home,Away=table$Away,Y=table$Y,
               method="BFGS",control=list("fnscale"=-1))
```

Here theta0 is any initialization for θ (for example the all 0's vector). The method will use the BFGS algorithm, and "fnscale" $= -1$ indicates that it should perform maximization rather than minimization.

(b) Fit the Bradley-Terry model without an intercept term. (You may do this in R by defining a new function

```
loglik_noalpha = function(theta,Home,Away,Y)
```

where now $\theta = (\beta_2, \ldots, \beta_k)$, and using optim as before.) Evaluate the log-likelihoods at the full model and sub-model MLEs, and conduct a generalized likelihood ratio test of the null hypothesis of no home court advantage, $H_0 : \alpha = 0$. What is the p-value that you obtain for your test?

(c) For the m^{th} game, suppose we define 30 covariates $x_{m,1}, \ldots, x_{m,30}$ in the following way: Let $x_{m,1} = 1$ always. Let $x_{m,i} = 1$ if team i is the home team of this game and $i \neq 1$, and let $x_{m,j} = -1$ if team j is the away team of this game and $j \neq 1$. Let $x_{m,k} = 0$ for all other k. Explain why logistic regression for Y_m using the covariates $x_{m,1}, \ldots, x_{m,30}$ is equivalent to the Bradley-Terry model, where we constrain the Bradley-Terry score of team 1 to be $\beta_1 \equiv 0$. If we were to run this logistic regression, what would be the interpretation of the fitted coefficient for the first covariate $x_{m,1}$? For the 10th covariate $x_{m,10}$?

(d) Fit the logistic regression in part (c) using any standard regression software, and verify that the fitted coefficients match (up to reasonable numerical accuracy) your estimated parameters from part (a).

To do this in R, you may construct a matrix X of size 1230×30 containing the covariates as defined in part (c), and then fit the regression using

```
model = glm.fit(X,table$Y,family=binomial())
coefs = model$coefficients
```

Important References:

1. https://www.r-bloggers.com/2022/02/what-is-the-bradley-terry-model/
2. https://squared2020.com/2017/11/09/bradley-terry-rankings-introduction-to-logistic-regression/
3. https://encyclopediaofmath.org/wiki/Bradley-Terry_model