S&DS 242/542: Homework 11

Due Wednesday, April 23, at 1PM

This assignment will explore an example of multiple logistic regression known as the **Bradley-Terry model** for pairwise comparisons. The model may be described as follows:

Suppose there are p sports teams. A parameter $\beta_j \in \mathbb{R}$ is given to each team $j \in \{1, 2, \ldots, p\}$, representing the strength of that team. These teams play n total games, where each game is played between two teams $j, k \in \{1, 2, \ldots, p\}$, and one team j or k is the winner. The outcomes of the n games are modeled as independent Bernoulli random variables, where the probability p_{jk} that team j wins against team k is given by

$$p_{jk} = \frac{e^{\beta_j}}{e^{\beta_j} + e^{\beta_k}}$$

Note that these probabilities are unchanged if the same constant is added to all $\beta_1, \ldots, \beta_p$, i.e. β_j is only meaningful relative to β_k for the other teams. To address this ambiguity, let us fix by convention $\beta_p = 0$ for the last team, so that $\beta_1, \ldots, \beta_{p-1}$ represent the relative strengths of teams $j = 1, \ldots, p-1$ compared to this last team p.

1. **Reduction to logistic regression.** (a) Explain how this model may be written as a logistic regression model with n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $X = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and $Y \in \{0, 1\}$ with

$$\mathbb{P}[Y = 1 \mid X = x] = p(x), \qquad \log \frac{p(x)}{1 - p(x)} = \beta_1 x_1 + \ldots + \beta_p x_p$$

For each game $i = 1, \ldots, n$, what are the values of the predictors $X_i = (x_{i1}, \ldots, x_{ip})$?

(b) Suppose that each game between teams j and k is played in the home city of either team j or team k, and we wish to extend the Bradley-Terry model to

$$\log \frac{p_{jk}}{1 - p_{jk}} = \beta_j - \beta_k + \beta_0 \cdot \mathbf{1}\{\text{team } j \text{ is the home team}\} - \beta_0 \cdot \mathbf{1}\{\text{team } k \text{ is the home team}\}$$

Thus $\beta_0 \in \mathbb{R}$ represents a home-team advantage. Explain how this may be incorporated into your logistic regression model of part (a), by adding an intercept term and using an appropriate choice of $X_i = (x_{i1}, \ldots, x_{ip})$.

2. **Score equations for maximum likelihood estimation.** Let $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ be the unknown model parameters, where β_0 is the home-team advantage, and we fix $\beta_p = 0$. Letting $\ell_n(\beta)$ be the log-likelihood function for all n games, derive a simple form for the system of score equations

$$0 = \frac{\partial \ell_n(\beta)}{\partial \beta_0}, \quad 0 = \frac{\partial \ell_n(\beta)}{\partial \beta_1}, \quad \ldots, \quad 0 = \frac{\partial \ell_n(\beta)}{\partial \beta_{p-1}}$$

that you would solve to compute the MLEs $\hat{\beta}$ for β. Your equations should be expressed in terms of $\beta_0, \ldots, \beta_p$ (where $\beta_p = 0$), the game outcomes $Y_1, \ldots, Y_n \in \{0, 1\}$, and the teams playing in each game.

3. **Testing the home-team advantage.** Suppose we wish to test

$$H_0 : \beta_0 = 0 \qquad \text{vs.} \qquad H_1 : \beta_0 > 0$$

i.e. whether there is a positive home-team advantage.

(a) Explain how you would carry out an asymptotic level-α test based on the estimated value of $\hat{\beta}_0$, assuming that the model is correctly specified (i.e. the true distribution of the game outcomes $Y_1, \ldots, Y_n$ are correctly described by the Bradley-Terry model).

(b) If there is no home-team effect, then the model should remain the same if we randomly re-assign the home team of each game. Explain how to use this idea to compute a permutation null distribution for $\hat{\beta}_0$, and to carry out a permutation test of H_0 vs. H_1 that is robust against possible misspecification of the Bradley-Terry model.

4. **Fitting the Bradley-Terry model.** The file `NBA_record.csv` contains the results of all 1230 NBA games from the 2015–2016 regular season. The 30 teams are numbered from 1 to 30, according to the file `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.

(a) Fit the Bradley-Terry model with an intercept term β_0 for the home-team advantage, using maximum likelihood. What are the 5 teams (in ranked order) with the highest Bradley-Terry scores? What is the estimated value of β_0?

[If you wish, you may use a generic optimizer to maximize the log-likelihood rather than implementing a Newton-Raphson procedure. In `R`, you may first define a function

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

that returns the log-likelihood for the Bradley-Terry model given inputs $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ (constraining $\beta_p = 0$), Home $= (i_1, \ldots, i_n)$, Away $= (j_1, \ldots, j_n)$, and Y $= (Y_1, \ldots, Y_n)$. Then, you may read the data file and maximize the log-likelihood by

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

Here `beta.init` is any initialization for β (for example the all 0's vector). The MLE $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1})$ is then given by `result$par`.]

(b) Compute a model-based estimate of the standard error of $\hat{\beta}_0$ using the Fisher information matrix, and implement your hypothesis test in Problem 3(a). What is the p-value of your test?

(c) Implement your permutation test in Problem 3(b), using at least 500 random permutations. (This may take a few minutes to compute.)

Plot a histogram of your simulated permutation null distribution for $\hat{\beta}_0$. How closely does this permutation null distribution seem to match the normal null distribution for $\hat{\beta}_0$ that corresponds to your standard error estimate in part (b)?