Statistics 143 — Spring 2024 — Assignment 2

Due Tuesday February 20, 2024

Homework is to be uploaded on Gradescope by 10:00pm on Tuesday evening.

Please make sure on your assignment you indicate clearly other students with whom you collaborated, as well as any assistance you received from generative AI tools.

Written assignment

1. For the Maher Poisson model for game scores, suppose we want to estimate whether team i is better than team j. Assume for this problem that teams are playing at a neutral site, so that there is no home-field advantage.

   (a) Why is $\delta_{ij} = \log \lambda_{ij} - \log \lambda_{ji}$, as defined in the lecture notes, a reasonable parameter to assess the difference in team strengths? What values of δ indicate that team i is better?

   As defined in the lecture, we model $Y_{ijk} \sim \text{Poisson}(\lambda_{ij})$. λ_{ij} represents the scoring rate of team i against team j. If $\lambda_{ij} > \lambda_{ji}$, then team i is expected to have higher score than team j, which likely indicates that team i has higher strength than j, and vice versa. Therefore, it makes sense to use $\delta_{ij} = \log \lambda_{ij} - \log \lambda_{ji}$ to assess the difference in team strengths. Given that logarithm is a monotonic function, a positive δ_{ij} implies $\lambda_{ij} > \lambda_{ji}$, thus team i better.

   (b) Express δ_{ij} in terms of α_i, β_i, α_j and β_j.

   $\delta_{ij} = \alpha_i + \beta_i - \alpha_j - \beta_j$

   (c) Suppose you fit the Maher model to obtain the maximum likelihood estimates of the α_i and β_i, as well as the estimated covariance matrix of these estimates.

      i. Explain why the *offense* parameters, the α_i, are expected to be estimated higher than the *defense* parameters, the β_j, in games like hockey and soccer?

      In general, the average score of hockey or soccer games would be larger than 1. The Maher model estimates the expected points that team i gets against team j by

      $$\mathbb{E}(p_{ij}) = \lambda_{ij} = \exp(\alpha_i - \beta_j)$$

      For hockey or soccer games, the expected score of each team is likely to be larger than 1. Thus

      $$\exp(\alpha_i - \beta_j) > 1$$
      $$\alpha_i - \beta_j > 0$$

      ii. Describe a procedure to form a 95% confidence interval for δ_{ij}. (you will use this result in problem 2)

      As shown in 1b, we have $\delta_{ij} = \alpha_i + \beta_i - \alpha_j - \beta_j$.
      Therefore we can estimate δ_{ij} by

      $$\hat{\delta}_{ij} = \hat{\alpha}_i + \hat{\beta}_i - \hat{\alpha}_j - \hat{\beta}_j$$

And the variance of $\hat{\delta}_{ij}$ is

$$\begin{aligned}
\text{Var}(\hat{\delta}_{ij}) &= \text{Var}(\hat{\alpha}_i + \hat{\beta}_i - \hat{\alpha}_j - \hat{\beta}_j) \\
&= \text{Var}(\hat{\alpha}_i) + \text{Var}(\hat{\beta}_i) + \text{Var}(\hat{\alpha}_j) + \text{Var}(\hat{\beta}_j) \\
&+ 2(\text{Cov}(\hat{\alpha}_i, \hat{\beta}_i) - \text{Cov}(\hat{\alpha}_i, \hat{\alpha}_j) - \text{Cov}(\hat{\alpha}_i, \hat{\beta}_j) \\
&- \quad \text{Cov}(\hat{\beta}_i, \hat{\alpha}_j) - \text{Cov}(\hat{\beta}_i, \hat{\beta}_j) \\
&+ \quad \text{Cov}(\hat{\alpha}_j, \hat{\beta}_j))
\end{aligned}$$

2. The data file `england_premier_score_2022.csv` in the Data Sets folder on the course Canvas site contains game results for the 2022-23 season of England Premier League (EPL) games. The data set consist of 380 games. The following variables will be relevant for this problem.

   `Home`: Home team

   `scoreHome`: Home team's number of goals

   `Away`: Visiting team

   `scoreAway`: Visiting team's number of goals

   `atHome`: 1 if played at home team's arena, 0 if played at a neutral site or not the usual home arena (always 1 for this data set)

   The data set consists of 380 games. The goal of this problem is to answer questions about team strength during the 2022 EPL season.

   (a) Fit the Maher Poisson model without a home-field effect for EPL game scores, ensuring that the sum of offense and defense parameters sum to 0. Based on ranking the estimated ability parameters, who are the best three teams based on their offense, and who are the best three teams based on their defense?

   Solution:

```
> score = read.csv("england_premier_score_2022.csv")
> Teams = unique(score$Home)
> n_teams <-length(Teams)
> # Construct design matrix
> X_home = cbind(outer(score$Home, Teams, "=="),
                 -outer(score$Away, Teams, "=="))
> X_away = cbind(outer(score$Away, Teams, "=="),
                 -outer(score$Home, Teams, "=="))
> X = rbind(X_home, X_away)
> W = rbind(diag(2*n_teams-1), rep(-1, 2*n_teams-1))
> X_star = X %*% W
> games = data.frame(score=c(score$scoreHome, score$scoreAway))
> games$X_star = X_star
> # Fit linear model
```

```
> ability.pois = glm(score~X_star+0, family='poisson', data=games)
> # Get ability estimate
> Teams.ability.est =  W %*% ability.pois$coefficients
> V = W %*% summary(ability.pois)$cov.unscaled %*% t(W)
> Teams.ability.stderr = sqrt(diag(V))
> Teams.ability = data.frame(
    Est=Teams.ability.est,
    StdErr=Teams.ability.stderr)
> rowname = c(
    paste(Teams,"_att",sep=''),
    paste(Teams,"_def",sep='')
)
> row.names(Teams.ability) = rowname
> dimnames(V) <- list(rowname, rowname)
> # Rank by attack and defence strength
> Teams.ability.att = Teams.ability[1:n_teams, ]
> Teams.ability.att = Teams.ability.att[order(-Teams.ability.att$Est),]
> Teams.ability.def = Teams.ability[(n_teams+1):(2*n_teams), ]
> Teams.ability.def = Teams.ability.def[order(-Teams.ability.def$Est),]
> Teams.ability.att[1:3,]
                          Est    StdErr
Manchester City_att 0.7227061 0.1020975
Arsenal_att         0.6664924 0.1054191
Liverpool_att       0.5100466 0.1139469
> Teams.ability.def[1:3,]
                           Est    StdErr
Newcastle Utd_def   0.31805035 0.1705993
Manchester City_def 0.29254913 0.1706030
Manchester Utd_def  0.06244722 0.1496953
```

(b) Based on the results in the last problem, construct a confidence interval for $\delta_{UTD,MC}$ as defined in part 1(a), for measuring the relative strengths of Manchester United (Utd) and Manchester City playing on a neutral field. Interpret the confidence interval.

$\delta_{UTD,MC}$ is estimated to be $\hat{\delta}_{UTD,MC} = \hat{\alpha}_{UTD} + \hat{\beta}_{UTD} + \hat{\alpha}_{MC} + \hat{\beta}_{MC} = -0.70445$. Its variance is given by

$$\begin{aligned}
\text{Var}(\hat{\delta}_{UTD,MC}) &= \text{Var}(\hat{\alpha}_{UTD}) + \text{Var}(\hat{\beta}_{UTD}) + \text{Var}(\hat{\alpha}_{MC}) + \text{Var}(\hat{\beta}_{MC}) \\
&\quad + 2(\text{Cov}(\hat{\alpha}_{UTD}, \hat{\beta}_{UTD}) - \text{Cov}(\hat{\alpha}_{UTD}, \hat{\alpha}_{MC}) - \text{Cov}(\hat{\alpha}_{UTD}, \hat{\beta}_{MC}) \\
&\quad - \quad \text{Cov}(\hat{\beta}_{UTD}, \hat{\alpha}_{MC}) - \text{Cov}(\hat{\beta}_{UTD}, \hat{\beta}_{MC}) \\
&\quad + \quad \text{Cov}(\hat{\alpha}_{MC}, \hat{\beta}_{MC})) \\
&= 0.12921^2 + 0.14969^2 + 0.10210^2 + 0.17060^2 \\
&\quad + 2(-0.001205 + 0.00041645 + 0.00039622 \\
&\quad + \quad 7.4805 \times 10^{-5} + 0.0010948 - 0.0012286) \\
&= 0.077731
\end{aligned}$$

The 95% confidence interval for the difference in scoring rates between Manchester

United and Manchester City is $(-1.251, -0.158)$. This implies that, with 95% confidence, Manchester City is a better team.

Code:

```
> var.delta.utd.mc = V['Manchester Utd_def','Manchester Utd_def'] +
+   V['Manchester Utd_att','Manchester Utd_att'] +
+   V['Manchester City_att','Manchester City_att'] +
+   V['Manchester City_def','Manchester City_def'] +
+   2 * (V['Manchester Utd_att','Manchester Utd_def'] -
+          V['Manchester Utd_att', 'Manchester City_att'] -
+          V['Manchester Utd_att', 'Manchester City_def'] -
+          V['Manchester Utd_def', 'Manchester City_att'] -
+          V['Manchester Utd_def', 'Manchester City_def'] +
+          V['Manchester City_att', 'Manchester City_def'])
> delta.utd.mc = Teams.ability.att['Manchester Utd_att', 'Est'] +
+   Teams.ability.def['Manchester Utd_def', 'Est'] -
+   Teams.ability.att['Manchester City_att', 'Est'] -
+   Teams.ability.def['Manchester City_def', 'Est']
> c(delta.utd.mc-1.96*sqrt(var.delta.utd.mc),
+   delta.utd.mc+1.96*sqrt(var.delta.utd.mc))
[1] -1.250907 -0.158001
```

(c) From the estimated strength parameters, estimate the probabilities of game score differences from Manchester City winning by 6 goals to Manchester United winning by 6 goals. What is the probability Manchester City wins the game? Ties the game? Loses the game?

```
library(skellam)
goal.diff= data.frame("score.diffenence"=-6:6,
  "probability"=dskellam(-6:6,
    exp(Teams.ability.att["Manchester Utd_att", "Est"]-
          Teams.ability.def["Manchester City_def", "Est"]),
    exp(Teams.ability.att["Manchester City_att", "Est"]-
          Teams.ability.def["Manchester Utd_def", "Est"]), log=FALSE))
> p.mc.win = sum(goal.diff[goal.diff$score.diffenence<0,'probability'])
> p.tie = goal.diff[goal.diff$score.diffenence==0,'probability']
> p.mc.lose = sum(goal.diff[goal.diff$score.diffenence>0,'probability'])
> goal.diff
   score.diffenence  probability
1                -6 5.250403e-03
2                -5 1.697306e-02
3                -4 4.644708e-02
4                -3 1.043912e-01
5                -2 1.847847e-01
6                -1 2.425717e-01
7                 0 2.166946e-01
8                 1 1.199222e-01
9                 2 4.516324e-02
```

```
10                3 1.261370e-02
11                4 2.774575e-03
12                5 5.012540e-04
13                6 7.665665e-05
> p.mc.win
[1] 0.6004182
> p.tie
[1] 0.2166946
> p.mc.lose
[1] 0.1810517
```

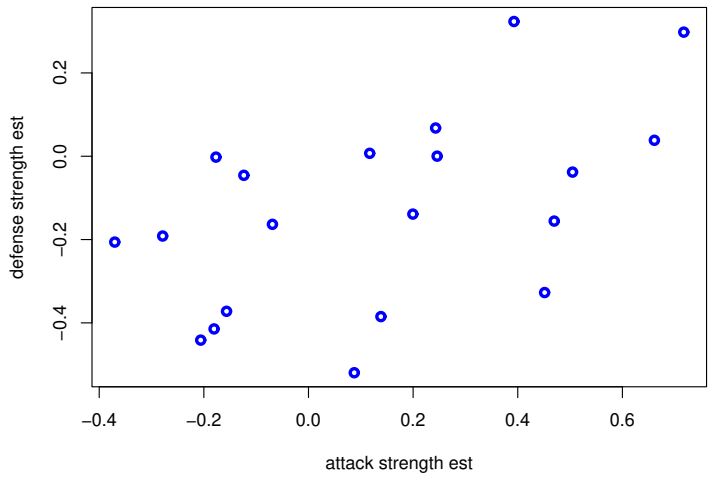
(d) Now refit the above model, but include a home-field advantage parameter. Based on the fit, how much more of a multiplicative increase in the mean goals scored is playing at home versus playing at an opposing team's home field? What do you make of this home field advantage?

```
> games$hfa = c(score$atHome, -score$atHome)
> ability.hfa.pois = glm(score~X_star+hfa+0, family='poisson', data=games)
> W_h = rbind(cbind(W,0),c(rep(0,2*n_teams-1),1))  # block matrix
> Teams.ability.hfa.est = W_h %*% ability.hfa.pois$coefficients
> Teams.ability.hfa = data.frame( Est=Teams.ability.hfa.est)
> row.names(Teams.ability.hfa) = c(rowname, "HFA")
> exp(2 * Teams.ability.hfa['HFA','Est'])
[1] 1.341254
```

A team is expected to score 1.341x times more points if it plays at home field.

(e) Based on the model in part (d), how related are the offense and defense estimated strengths? Construct a scatter plot of the teams' strength estimates and summarize the relationship.

```
> cor(Teams.ability.hfa[1:n_teams,'Est'],
+     Teams.ability.hfa[(n_teams+1):(n_teams*2),'Est'])
[1] 0.5194067
> plot(Teams.ability.hfa[1:n_teams,'Est'],
+      Teams.ability.hfa[(n_teams+1):(n_teams*2),'Est'],
+      col="blue",lwd=3,
+      xlab="attack strength est",ylab="defense strength est")
```



From the scatter plot, we observe that points are mostly distributed along the diagonal from the lower-left corner to upper-right corner, indicating a positive and weakly linear correlation between the estimated offense strength and defense strength of teams. In general, a team with high offense strength is also likely to have high defense ability. The Pearson correlation of the estimated attack and defense strength is 0.5194, which also indicates a positive correlation.

(f) Fit the Dixon and Coles model on the EPL game outcomes, with a home-field advantage parameter. You will need to install the `goalmodel` package to do so – see the R code accompanying the Poisson lecture notes. Based on the Dixon and Coles model, what are the top three teams in terms of their offensive strength? What about the top three teams in terms of their defensive strength? How do these compare to the Maher model results?

```
> library(goalmodel)
> ability.hfa.dc = goalmodel::goalmodel(goals1=score$scoreHome,
                                        goals2=score$scoreAway,
                                        team1=score$Home,
                                        team2=score$Away,
                                        hfa = TRUE, dc=TRUE)
>Teams.ability.hfa.dc = data.frame(
  attack=ability.hfa.dc$parameters$attack,
  defense=ability.hfa.dc$parameters$defense)
> Teams.ability.hfa.dc[order(-Teams.ability.hfa.dc$attack),][1:3,]
                  attack    defense
Manchester City 0.5873571 0.42535363
Arsenal         0.5288894 0.16995852
Liverpool       0.3686030 0.09774777
> Teams.ability.hfa.dc[order(-Teams.ability.hfa.dc$defense),][1:3,]
                  attack   defense
Newcastle Utd   0.2482002 0.4756757
Manchester City 0.5873571 0.4253536
Manchester Utd  0.1088974 0.1965154
```

The Dixon-Coles model predicts the exact same top 3 teams as Maher model, by both offense and defense strength.

(g) To address whether there is any practical difference in predicting games between the Maher versus Dixon and Coles models, determine the predicted goals scores (for both teams) for all the 2022-23 season EPL games from both models (with the home-field advantage). The predicted winner of each game within a model is the one that has the higher predicted goals scored. In how many of these games is the winner different between the Maher versus Dixon and Coles models? What do you make of this result?

```
pred.pois = predict(ability.hfa.pois, type='response')
pred.dc = predict_expg(ability.hfa.dc,
                       team1=score$Home,
                       team2=score$Away,
                       return_df = TRUE)
pred.home.win = data.frame('poisson'=pred.pois[1:nrow(score)]>
                             pred.pois[(nrow(score)+1):nrow(games)],
                           'dc'=pred.dc$expg1>pred.dc$expg2)
sum(pred.home.win[,1] != pred.home.win[, 2])
[1] 4
```

Out of 380 games, there are only 4 games ($\sim$1%) in which Dixon and Coles model predicts different winner from Maher model. We see that these two models are arguably consistent in predicting winners.

3. As we have seen in lecture, the Thurstone-Mosteller model can be derived by assuming competitor i with strength parameter θ_i has a performance that is randomly generated from a $\text{N}(\theta, \frac{1}{2})$ distribution. Thurstone also considered, in his original 1927 paper,[1] a model in which competitor i's performance was generated from a $\text{N}(\theta_i, \sigma_i^2)$ distribution, where θ_i and σ_i^2 are parameter that would both be estimated during model-fitting.

(a) Suppose competitors i and j have performance distributions that come from $\text{N}(\theta_i, \sigma_i^2)$ and $\text{N}(\theta_j, \sigma_j^2)$, respectively. What is the probability i defeats j conditional on the model parameters?

Let z_i, z_j be the score of team i and j in one game. This model assumes

$$z_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$$
$$z_j \sim \mathcal{N}(\theta_j, \sigma_j^2)$$

So $z_i - z_j \sim \mathcal{N}(\theta_i - \theta_j, \sigma_i^2 + \sigma_j^2)$. Let $y_i = 1$ if team i wins and 0 otherwise

$$P(y_i = 1) = P(z_i - z_j > 0) = P(Z > \frac{0 - (\theta_i - \theta_j)}{\sqrt{\sigma_i^2 + \sigma_j^2}})$$
$$= 1 - P(Z < \frac{0 - (\theta_i - \theta_j)}{\sqrt{\sigma_i^2 + \sigma_j^2}}) = \Phi(\frac{\theta_i - \theta_j}{\sqrt{\sigma_i^2 + \sigma_j^2}})$$

(b) Consider a round-robin tournament of J competitors in which each competitor plays every other competitor exactly once. For the above Thurstonian model, what is the

[1] Thurstone, L. L. (1927). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4), 384-400.

likelihood function of $(\theta_1, \ldots, \theta_J, \sigma_1^2, \ldots, \sigma_J^2)$?

Let y_{ij} be the outcome of the game between team i and j, $y_{ij} = 1$ if team i wins the game and 0 if team j wins. Then

$$\ell(\boldsymbol{\theta}, \boldsymbol{\sigma}^2) = \prod_{i=1}^{J-1} \prod_{j=i+1}^{J} \left(\Phi(\frac{\theta_i - \theta_j}{\sqrt{\sigma_i^2 + \sigma_j^2}})^{y_{ij}} \left(1 - \Phi(\frac{\theta_i - \theta_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}) \right)^{(1-y_{ij})} \right)$$

(c) As usual, assume a linear constraint on the θ_j, such as $\sum_{j=1}^{J} \theta_j = 0$. For the Thurstonian model in this problem, is this constraint sufficient to estimate the model parameters uniquely, even competitors play each other multiple times and assuming the Ford (1957) condition is satisfied? Why or why not?

No, we cannot determine an unique estimation of model parameters. As shown in part (b), the likelihood of model parameters is given by $\Phi^{y_{ij}}(1-\Phi)^{(1-y_{ij})}$. Suppose $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\sigma}}^2$ are MLE of the parameters, then simultaneously scale $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\sigma}}$ by any non-negative constant c will give the same likelihood.

$$\begin{aligned} L(c\hat{\boldsymbol{\theta}}, c^2\hat{\boldsymbol{\sigma}}^2) &= \Phi(\frac{c\hat{\theta}_i - c\hat{\theta}_j}{\sqrt{c^2\hat{\sigma}_i^2 + c^2\hat{\sigma}_j^2}})^{y_{ij}} \left(1 - \Phi(\frac{c\hat{\theta}_i - c\hat{\theta}_j}{\sqrt{c^2\hat{\sigma}_i^2 + c^2\hat{\sigma}_j^2}}) \right)^{(1-y_{ij})} \\ &= \Phi(\frac{c(\hat{\theta}_i - \hat{\theta}_j)}{c\sqrt{\hat{\sigma}_i^2 + \hat{\sigma}_j^2}})^{y_{ij}} \left(1 - \Phi(\frac{c(\hat{\theta}_i - \hat{\theta}_j)}{c\sqrt{\hat{\sigma}_i^2 + \hat{\sigma}_j^2}}) \right)^{(1-y_{ij})} = L(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\sigma}}^2) \end{aligned}$$

(d) Consider players A, B and C with performance distributions $\mathrm{N}(\theta_A, \sigma_A^2)$, $\mathrm{N}(\theta_B, \sigma_B^2)$ and $\mathrm{N}(\theta_C, \sigma_C^2)$, respectively. Does the Thurstonian model for this problem always satisfy weak stochastic transitivity, that is, $p_{ab} \geq \frac{1}{2}$ and $p_{bc} \geq \frac{1}{2}$ implying $p_{ac} \geq \frac{1}{2}$?

Yes, this model always satisfies weak stochastic transitivity. By part (a), $p_{ab} = \Phi(\frac{\theta_a - \theta_b}{\sqrt{\sigma_a^2 + \sigma_b^2}})$,

$$\begin{aligned} p_{ab} \geq \frac{1}{2} &\implies \frac{\theta_a - \theta_b}{\sqrt{\sigma_a^2 + \sigma_b^2}} \geq 0 \implies \theta_a \geq \theta_b \\ p_{bc} \geq \frac{1}{2} &\implies \frac{\theta_b - \theta_c}{\sqrt{\sigma_b^2 + \sigma_c^2}} \geq 0 \implies \theta_b \geq \theta_c \end{aligned}$$

Then we have

$$\begin{aligned} & \theta_a \geq \theta_c \\ \implies & \frac{\theta_a - \theta_c}{\sqrt{\sigma_a^2 + \sigma_c^2}} \geq 0 \\ \implies & p_{ac} = \Phi(\frac{\theta_a - \theta_c}{\sqrt{\sigma_a^2 + \sigma_c^2}}) \geq \frac{1}{2} \end{aligned}$$

4. This problem involves a reexamination of the NBA data from the first problem set. For this problem, the game outcome of interest which team won. As a reminder, the data file `nba_score_2022.csv` can be downloaded from the Data Sets folder on the course Canvas site. Please refer to the previous problem set for the details of the data.

   (a) Create a binary vector that indicates whether the team labeled as the "home" team won the game. Fit a Thurstone-Mosteller model including a home-court indicator to these data, with the constraint that the sum of the strength parameters is zero. Display the team strength estimates along with the standard errors in order of the strengths (from highest to lowest). How does this ranking compare to the version you produced using the normal model in the first problem set?

```
> score = read.csv("
nba_score_2022.csv")[1:1230,]
> score["homeWin"] = ifelse(score$PtsDiff>0, 1,0)
> Teams = unique(score$Home)
> n_teams <-length(Teams)
> # Construct design matrix
> X = outer(score$Home, Teams, "==") - outer(score$Visitor, Teams, "==")
> W = rbind(diag(n_teams-1), rep(-1, n_teams-1))
> X_star = X %*% W
> score$X_star = X_star
> # Fit probit model
> ability.probit.hfa = glm(homeWin~X_star+AtHome+0,
                           family=binomial(link="probit"), data=score)
> # Add HFA
> W_h = rbind(cbind(W,0),c(rep(0,n_teams-1),1))  # block matrix
> # Ability estimate probit
> Teams.ability.probit = data.frame(
                            "Est"=W_h %*% ability.probit.hfa$coefficients)
> V.probit = W_h %*% summary(ability.probit.hfa)$cov.unscaled %*% t(W_h)
> Teams.ability.probit["Stderr"] = sqrt(diag(V.probit))
> row.names(Teams.ability.probit) <- c(Teams,"HFA")
> dimnames(V.probit) = list(c(Teams,"HFA"), c(Teams,"HFA"))
> Teams.ability.probit = Teams.ability.probit[
                            order(-Teams.ability.probit$Est),]
> Teams.ability.probit
                              Est     Stderr
Milwaukee Bucks         0.566832104 0.14561605
Boston Celtics          0.535691965 0.14500580
Philadelphia 76ers      0.434073967 0.14215161
Denver Nuggets          0.368301219 0.14106592
Cleveland Cavaliers     0.334832070 0.14068924
Memphis Grizzlies       0.318049213 0.14048394
HFA                     0.224371786 0.03756713
New York Knicks         0.204998294 0.13887170
Sacramento Kings        0.204187612 0.13887117
Brooklyn Nets           0.145374103 0.13829630
Phoenix Suns            0.134979967 0.13800543
```

```
Miami Heat              0.105592370 0.13822908
Los Angeles Clippers    0.087578977 0.13802819
Golden State Warriors   0.085663909 0.13809475
Los Angeles Lakers      0.056140608 0.13769127
Toronto Raptors         0.023844138 0.13817015
New Orleans Pelicans    0.020847655 0.13782106
Minnesota Timberwolves  0.010965258 0.13795020
Atlanta Hawks           0.006548705 0.13798833
Chicago Bulls          -0.014527848 0.13833982
Oklahoma City Thunder  -0.032396646 0.13780957
Dallas Mavericks       -0.109323032 0.13811525
Utah Jazz              -0.137714623 0.13810301
Washington Wizards     -0.172940708 0.13903321
Indiana Pacers         -0.177176514 0.13900952
Orlando Magic          -0.213682027 0.13925323
Portland Trail Blazers -0.261758986 0.13951088
Charlotte Hornets      -0.444428091 0.14316437
San Antonio Spurs      -0.626247059 0.14760841
Houston Rockets        -0.637312577 0.14801910
Detroit Pistons        -0.816994021 0.15613968
```

This ranking agrees with the ranking predicted by normal models *to some extent*. Teams ranked high by the normal model are also strong teams by the Thurstone-Mosteller model, and vice versa. For example, both models predict the same best 5 (Bucks , Celtics, 76ers, Nuggets, Cavaliers) and worst 5 teams (Trail Blazers, Hornets, Spurs, Rockets , Pistons). The exact order of the ranking, however, may be different (e.g. the normal model estimates Celtics to be the best team whereas Thurstone-Mosteller model says Bucks.)

(b) Interpret the estimated home-court advantage parameter. If team A has a 0.55 probability of defeating team B on a neutral court, what is the probability of team A defeating team B on the team A's home court assuming the home-court estimated parameter you obtained?

```
> Teams.ability.probit['HFA',"Est"]
[1] 0.2243718
> d.theta = qnorm(0.55)
> pnorm(d.theta+Teams.ability.probit['HFA',"Est"])
[1] 0.6368431
```

The estimated HFA shows that a team would have a 0.224 boost in its strength estimate if it plays at home court. Recall that the Thurstone-Mosteller model estimates

$$\Pr(y_i = 1) = \Phi(\theta_a - \theta_b + \beta_h)$$

If team A has a 0.55 probability of defeating team B on a neutral court, we can compute the difference of estimated strength of team A and B (`d.theta` in the code) by $\theta_a - \theta_b = \Phi^{-1}(0.55)$. Then adding the home field advantage term to strength of team A to compute

$$\Pr(y_i = 1) = \Phi(\Phi^{-1}(0.55) + \beta_h)$$

We see that Team A would have a 0.64 probability of defeating team B on the its home court.

(c) With 95% confidence, is there evidence whether the Los Angeles Lakers was a better or worse team than the Atlanta Hawks? How does this compare to the result from the first homework? What do you make of the difference?

```
> # Compare Lakers and Hawk with 95% confidence interval
> StdErr.LAL.ATL=sqrt(
+   Teams.ability.probit["Los Angeles Lakers", "StdErr"]^2+
+   Teams.ability.probit["Atlanta Hawks", "StdErr"]^2 -
+   2*V.probit["Los Angeles Lakers", "Atlanta Hawks"])
> Diff.LAL.ATL = Teams.ability.probit["Los Angeles Lakers", "Est"] -
+   Teams.ability.probit["Atlanta Hawks", "Est"]
> ConfIntv.LAL.ATL = c(Diff.LAL.ATL - 1.96 * StdErr.LAL.ATL,
+                      Diff.LAL.ATL + 1.96 * StdErr.LAL.ATL)
> ConfIntv.LAL.ATL
[1] -0.3414829  0.4406667
```

There is no strong evidence at the 95% confidence level that Lakers is a better or worse team than Hawks, consistent with HW1. This is expected since the binary model discards more information than the normal model, so it may not be able to make more accurate estimation of team strengths.

(d) Now fit a Bradley-Terry model with a home-court advantage parameter to the data. How do the team rankings you obtain from the Bradley-Terry fit compare to the Thurstone-Mosteller rankings?

```
> ability.logit.hfa = glm(homeWin~X_star+AtHome+0,
                          family=binomial, data=score)
> Teams.ability.logit = data.frame(
                        "Est"=W_h %*% ability.logit.hfa$coefficients)
> V.logit = W_h %*% summary(ability.logit.hfa)$cov.unscaled %*% t(W_h)
> Teams.ability.logit["StdErr"] = sqrt(diag(V.logit))
> row.names(Teams.ability.logit) <- c(Teams,"HFA")
> dimnames(V.logit) = list(c(Teams,"HFA"), c(Teams,"HFA"))
> Teams.ability.logit[order(-Teams.ability.logit$Est),]
                             Est     StdErr
Milwaukee Bucks         0.93974508 0.24398104
Boston Celtics          0.88068517 0.24214369
Philadelphia 76ers      0.72533418 0.23539868
Denver Nuggets          0.60854251 0.23246215
Cleveland Cavaliers     0.53985369 0.23104789
Memphis Grizzlies       0.49980997 0.23025736
HFA                     0.37006206 0.06191402
Sacramento Kings        0.34124983 0.22681367
New York Knicks         0.33863857 0.22674818
Brooklyn Nets           0.24020321 0.22548171
Phoenix Suns            0.22195459 0.22459602
Miami Heat              0.17765364 0.22536753
Golden State Warriors   0.14637390 0.22489909
Los Angeles Clippers    0.13800258 0.22465883
Los Angeles Lakers      0.09405362 0.22388511
```

```
New Orleans Pelicans    0.03458784 0.22421745
Toronto Raptors         0.03350955 0.22517357
Minnesota Timberwolves  0.03324408 0.22461337
Atlanta Hawks           0.01360455 0.22469647
Chicago Bulls          -0.03067615 0.22568399
Oklahoma City Thunder  -0.05376257 0.22424105
Dallas Mavericks       -0.17167391 0.22507041
Utah Jazz              -0.21016852 0.22493240
Washington Wizards     -0.28550920 0.22756424
Indiana Pacers         -0.28677997 0.22739944
Orlando Magic          -0.35230538 0.22821624
Portland Trail Blazers -0.42754202 0.22886319
Charlotte Hornets      -0.73945854 0.23856395
San Antonio Spurs      -1.03954908 0.25047833
Houston Rockets        -1.04600731 0.25071519
Detroit Pistons        -1.36361393 0.27225704
```

The Bradley-Terry model ranking is basically identical to the T-H model ranking, with only a few swap between neighboring teams (e.g. Warriors and Clippers swapped position.)

(e) Paired comparison models can assist the estimation of the number of games a team wins by the end of a season from game outcomes mid-way through the season. The file "nba-2023.csv" consists of game outcomes of the current NBA season up through January XX, 2024, along with the match-ups through the rest of the season (until April 14, 2024) where the game outcomes are missing.

- Read in the file into a data frame called `nba23`. Create a binary indicator variable for whether the home team won, which takes the value `NA` for the unplayed games.
- Fit the Bradley-Terry model with a home-court advantage to these data. You can include the entire data set with the missing outcomes – R will just ignore those games in the model fit.
- Using the `predict` function with the argument `type="response"` applied to the `nba23` data frame, create a vector of predicted probabilities for the home team across all games. These are also the estimated mean outcomes for each game.
- Create a new vector whose elements are the observed game binary outcomes for the games that have been played, and the estimated mean outcomes for the unplayed games. Call this `y`. This can be accomplished by first assigning `y` to be the vector of game outcomes with `NA`'s for the unplayed games, and then assigning the subset of values of `y` that are missing to be the estimated mean outcomes.
- Assuming that `X` is the pairing design matrix, and `Teams` are the team names, perform the following commands in R:

```
ystar = 2*y - 1 # ystar ranges from -1 to 1
Xty = as.vector(t(X) %*% ystar)
wins.vec = (Xty + 82)/2
names(wins.vec) = Teams
```

In the third line of code, 82 is twice the number of games played by each team. Explain why the above code produces the expected number of total wins per team in the vector `win.vec`.

Sort the vector `wins.vec` in order from highest to lowest, and interpret the results. We will check mid-April the accuracy of these predictions.

```
> nba23 = read.csv("nba-2023.csv")
> nba23['homeWin'] = ifelse(nba23$HomeScore > nba23$AwayScore, 1, 0)
> X23 = outer(nba23$Home.Team, Teams, "==") -
        outer(nba23$Away.Team, Teams, "==")
> nba23['X_star'] = X23 %*% W
> ability23.logit.hfa = glm(homeWin~X_star+atHome+0,
                            family=binomial, data=nba23)
> pred23 = predict(ability23.logit.hfa, newdata=nba23, type='response')
> y = nba23$homeWin
> y[is.na(y)] = pred23[is.na(y)]
> ystar = 2*y - 1 # ystar ranges from -1 to 1
> Xty = as.vector(t(X23) %*% ystar)
> wins.vec = (Xty + 82)/2
> names(wins.vec) = Teams
> wins.vec = wins.vec[order(-wins.vec)]
> as.matrix(wins.vec)
                           [,1]
Boston Celtics         65.21681
Oklahoma City Thunder  59.00034
Minnesota Timberwolves 58.87382
Denver Nuggets         55.69406
Philadelphia 76ers     53.56908
Milwaukee Bucks        53.23439
Cleveland Cavaliers    51.86334
New York Knicks        51.85546
New Orleans Pelicans   47.97199
Indiana Pacers         47.16933
Sacramento Kings       46.57605
Dallas Mavericks       45.37861
Orlando Magic          45.36332
Phoenix Suns           44.66581
Miami Heat             43.47410
Utah Jazz              41.08831
Los Angeles Clippers   41.00000
Los Angeles Lakers     40.83654
Houston Rockets        39.44713
Golden State Warriors  38.10797
Chicago Bulls          36.99234
Memphis Grizzlies      34.88259
Brooklyn Nets          32.29717
Atlanta Hawks          32.28164
Toronto Raptors        29.93946
Portland Trail Blazers 23.70927
Charlotte Hornets      17.96835
San Antonio Spurs      15.83071
```

```
Washington Wizards     12.84394
Detroit Pistons         8.98001
```

Interpretation 1

The code given above computes $X^T y^*$, which will be a vector of length J, assuming J is the number of teams. The j-th row of X^T would be the match schedule of the team j, which can only take value 1 (play home), 0 (not play) or -1 (play away). Then the j-th element of $X^T y^*$ is $\sum_i (X^T)_{ji} \cdot y_i^*$. The i-th element of y^* is the expected outcome of game i, re-scaled to [-1, 1]. If positive, y_i^* is proportional to the expected games that home team wins, and negative y_i^* is proportional to the expected games that visitor team wins. Then $\sum_i (X^T)_{ji} \cdot y_i^*$ gives the sum of all games that team j attends times the scaled expected outcome of that game. Then the third line of the code convert it back to the expected number of games that team j wins in the season. We add 82 because for the outcome of each game, we subtract 1 from y_i to get y_i^*, and there are 82 games in total for each team.

Interpretation 2

An alternative way of interpreting this code is

$$X^T y = \texttt{ExpectedWin-ExpectedLose}$$

Thus,

$$\begin{aligned}(X^T y + 82)/2 &= (\texttt{ExpectedWin-ExpectedLose} + 82)/2 \\ &= (\texttt{ExpectedWin} - (82 - \texttt{ExpectedWin}) + 82)/2 \\ &= 2 * \texttt{ExpectedWin}/2 = \texttt{ExpectedWin}\end{aligned}$$

The result shows that Celtics is expected to have the best record in 2023 season. The model predicts that Celtics, Timberwolves, 76ers, Bucks, Cavaliers, Knicks, Pacers, and Heat are the teams in the Eastern playoff; Thunder, Nuggets, Pelicans, Kings, Mavericks, Magic, Suns, and Jazz can enter Western playoff.