Statistics 143 — Spring 2024 — Assignment 1

Due Monday February 5, 2024

Homework is to be uploaded on Gradescope by 10:00pm on Monday 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. Let $\boldsymbol{X}$ be a $n \times J$ pairing design matrix in which the (i,j)-th element ($i = 1,\ldots,n$, $j = 1,\ldots,J$) is 1 if team j played at home in game i, -1 if team j played away in game i, and 0 otherwise. Show explicitly that the (j,j)-th element of $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ counts the number of games played by team j, and that the (j,k)-th element of $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ (with $j \neq k$) is the number of times team j has played against team k, multiplied by -1.

   Solution:
   Let X_{j*} be the j-th row of X and X_{*j} be the j-th column of X. The (j,j)-th element of X^TX is obtained by $(X^TX)_{jj} = X_{j*}^T \cdot X_{*j}$. Note that $X_{j*}^T = X_{*j}$, so $(X^TX)_{jj} = X_{*j} \cdot X_{*j} = \sum_i (X_{ij})^2$. By defination of X, $(X_{ij})^2 = 1$ if team j played in game i and 0 otherwise. Thus $\sum_i (X_{ij})^2$ equals to the total number of games that team j played.

   Similarly we have

$$\begin{aligned}(X^TX)_{jk} &= X_{j*}^T \cdot X_{*k} \\ &= \textstyle\sum_i X_{ij}X_{ik}\end{aligned}$$

   If team j played against team k in game i, then X_{ij}, X_{ik} will be either 1, -1 or -1, 1, respectively. In either case, we have $X_{ij}X_{ik} = -1$ if both teams played in game i and 0 otherwise. Thus $(X_TX)_{jk}$ equals the number of games that team j,k played against each other, multiplied by -1.

2. For the normal linear model with

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

   with $\boldsymbol{X}$ defined in problem 1, suppose that instead of imposing the linear constraint $\theta_J = -\sum_{j=1}^{J-1}\theta_j$ as in the lecture notes, we assume $\theta_J = 0$ as the "reference" team. Let $\boldsymbol{\theta}_{-1} = (\theta_1,\ldots,\theta_{J-1})$ be the vector of team strengths leaving out θ_J.

   (a) Construct a $J \times (J-1)$ matrix $\boldsymbol{W}$ such that $\boldsymbol{W}\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}$.

$$\boldsymbol{W} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} = [\mathbb{I}_{J-1}\ \mathbf{0}]^T, \text{ where } \mathbb{I}_{J-1} \text{ is a } J-1 \text{ identity matrix.}$$

(b) Let $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{W}$. Describe the characteristics of the matrix $\boldsymbol{X}^*$. In particular, what do the rows of $\boldsymbol{X}^*$ look like depending on whether team J was involved in a game?

If team J is not involved in game i, the i-th row of $\boldsymbol{X}^*$ will consist exactly one 1 **and** -1 and $J-2$ zeros. The location of the 1 and -1 are in the columns corresponding to the teams involved in the game. Otherwise the i-th row of $\boldsymbol{X}^*$ will consist **either** one 1 **or** -1 and and $J-2$ zeros. The 1 or -1 would be in the column of team J's opponent.

(c) If we fit a least-squares regression model with design matrix $\boldsymbol{X}^*$ defined in part (b), how should we carry out estimation of $\boldsymbol{\theta}$, the full vector of team strengths?

As shown in the lecture, we can first obtain the least-square estimation of $\boldsymbol{\theta}_{-1}$ by

$$\hat{\boldsymbol{\theta}}_{-1} = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{y}$$

Then the estimation of $\boldsymbol{\theta}$ is given by $\hat{\boldsymbol{\theta}} = \boldsymbol{W}\hat{\boldsymbol{\theta}}_{-1}$

(d) Suppose we are interested in estimating whether team 1 is better than team J by constructing a confidence interval for $\theta_1 - \theta_J$.

i. Is it true that, for the linear constraint imposed, we can answer the question by simply constructing a confidence interval for θ_1 based on the fitted model?

True. Because we fixed θ_J to be exactly 0, the confidence interval for $\theta_1 - \theta_J$ is equivalent to that of θ_1.

ii. If we had used a different linear constraint, such as $\theta_J = -\sum_{j=1}^{J-1}\theta_j$, would we get a different confidence interval for $\theta_1 - \theta_J$ than the one obtained assuming $\theta_J = 0$?

The result would be **the same**. Because the linear constraint is only made to address identifiability issues (choosing one from infinitely many equally good estimates), the particular choice should have no effect on confidence intervals for the mean score difference $\mathrm{E}(y_{ab}) = \theta_a - \theta_b$. Because $\theta_1 - \theta_J$ is the mean score difference for team 1 playing against team J, the choice of the linear constraint will not impact the confidence interval.

3. The data file `nba_score_2022.csv` in the Data Sets folder on the course Canvas site contains game outcomes for the entire 2022-23 NBA basketball regular season, a total of 1230 games. The data consist of the following variables.

`Date`: Date game was played

`Start..ET.`: Time (ET) the game started

`Visitor`: Visiting team

`PtsV`: Points scored by visiting team

`Home`: Home team

`PtsH`: Points scored by home team

`X`: Whether the game went into overtime (missing if no, OT if yes)

`Attend.`: Number in attendance

`Arena`: Site of game

`PtsDiff`: `PtsH` − `PtsV`

`AtHome`: 1 if played at home team's arena, 0 if played at a neutral site or not the usual home arena

The goal of this problem is to answer questions about team strength at the end of 2022 NBA season.

(a) Fit a normal model for the game score differences across all the games, including a home-field indicator. Summarize the estimates and their standard errors of the team abilities during the 2022 NBA season, and create a table that places the team in rank order according to their estimates. Interpret the magnitude of the strength estimates, as well as the range of the estimates.

Solution:

```
> score = read.csv("nba_score_2022.csv")[1:1230,]
> Teams = unique(score$Home)
> n.team = length(Teams)
> # Construct design matrix
> X = outer(score$Home, Teams, "==") - outer(score$Visitor, Teams, "==")
> W = rbind(diag(n.team-1), rep(-1, n.team-1))
> X_star = X %*% W
> score$X_star = X_star
> ability.hfa.lm = lm(PtsDiff~X_star + AtHome + 0, data=score)
> W_h = rbind(cbind(W,0),c(rep(0,n.team-1),1))
> resid.stderr = summary(ability.hfa.lm)$sigma
> Teams.ability.hfa.est = W_h %*% ability.hfa.lm$coefficients
> V = W_h %*% vcov(ability.hfa.lm) %*% t(W_h)
> dimnames(V) = list(c(Teams,"HFA"), c(Teams,"HFA"))
> Teams.ability.hfa = data.frame( Est=Teams.ability.hfa.est,
+       StdErr=sqrt(diag(V)))
> row.names(Teams.ability.hfa) = c(Teams,"Home Advantage")
> # Ranking teams by strength estimates
> Teams.rank = Teams.ability.hfa[order(-Teams.ability.hfa$Est),]
> print(Teams.rank)
                              Est    StdErr
Boston Celtics          6.3781512 1.3553547
Cleveland Cavaliers     5.2305005 1.3553569
Philadelphia 76ers      4.3721568 1.3553547
Milwaukee Bucks         3.6092731 1.3553525
Memphis Grizzlies       3.6015276 1.3553547
Denver Nuggets          3.0428641 1.3553547
New York Knicks         2.9862360 1.3553503
Home Advantage          2.5362853 0.3619488
Sacramento Kings        2.3026189 1.3553547
Phoenix Suns            2.0796577 1.3553525
New Orleans Pelicans    1.6299934 1.3553546
Golden State Warriors   1.6256222 1.3553614
```

```
Toronto Raptors          1.5906276 1.3553547
Chicago Bulls            1.3363038 1.3553615
Brooklyn Nets            1.0291200 1.3553547
Oklahoma City Thunder    0.9607712 1.3553547
Los Angeles Lakers       0.4281226 1.3553569
Atlanta Hawks            0.3161489 1.3553525
Los Angeles Clippers     0.3084455 1.3553525
Dallas Mavericks        -0.1409781 1.3553547
Miami Heat              -0.1638628 1.3553595
Minnesota Timberwolves  -0.2470232 1.3553608
Utah Jazz               -1.0255983 1.3553547
Washington Wizards      -1.0583443 1.3553569
Orlando Magic           -2.3892885 1.3553525
Indiana Pacers          -2.9059575 1.3553525
Portland Trail Blazers  -3.9929313 1.3553586
Charlotte Hornets       -5.8904603 1.3553547
Houston Rockets         -7.6159462 1.3553525
Detroit Pistons         -7.6995426 1.3553588
San Antonio Spurs       -9.6982081 1.3554616
```

The difference between the strength estimates of two teams is the expected value of the score difference if these two teams play in a game, i.e. $\mathrm{E}(y_1 - y_2) = \hat{\theta}_1 - \hat{\theta}_2$. The magnitude of these estimates positively correlate to the strength of teams. The estimates are all in the range of -10 to 7, meaning that for any two teams play against each other, the score difference is expected to be less than 20 points.

(b) Interpret the estimated home court advantage. What does this value imply for playing on one's home court?

The estimate of the HFA, $\hat{\beta}_h$, is 2.53, meaning that one team is expected to score 2.53 more points if it plays at home relative to a neutral site, and 2(2.53) = 5.06 more points playing relative to the opposing team's site.

(c) One idea is to include information about the arena attendance as a predictor in the model.

i. Is this variable an endogenous or exogenous predictor? Explain.
Exogenous. Because the arena attendance is not relevant to the quality of a team.

ii. Including the arena attendance into the model as it is provided in the data set may not be making the best use of this information. Why might that be? How might you revise the variable so that it would make more sense to include as a potentially good predictor of the game score difference?

One possible problem is that the full capacities of stadiums are different. A large stadium may still have a larger audience than a small stadium even though it is only half-occupied. But in reality it might be the attendance rate that has more impact on the game score difference. Another caveat is that there are a few games

played in a neutral place, in which case we may want to exclude the effect of arena attendance as the audience size should not give either team any additional advantage.

One improvement is to create another derived predictor, `AtHome*AttendRate`, which is defined as `AtHome` $\times \frac{\texttt{Attend}}{\text{Full Capacity of the Arena}}$. Then fit a linear model using the derived variable as the exogenous predictor.

(d) With 95% confidence, is there evidence whether the Los Angeles Lakers was a better or worse team than the Atlanta Hawks?

```
> StdErr.LAL.ATL = sqrt(
+   Teams.ability.hfa["Los Angeles Lakers", "StdErr"] ** 2 +
+   Teams.ability.hfa["Atlanta Hawks", "StdErr"] ** 2 -
+   V["Los Angeles Lakers", "Atlanta Hawks"] * 2)
> Diff.LAL.ATL = Teams.ability.hfa["Los Angeles Lakers", "Est"] -
+   Teams.ability.hfa["Atlanta Hawks", "Est"]
> ConfIntv.LAL.ATL = c(Diff.LAL.ATL - 1.96 * StdErr.LAL.ATL,
+                      Diff.LAL.ATL + 1.96 * StdErr.LAL.ATL)
> print(ConfIntv.LAL.ATL)
[1] -3.734948  3.958895
```

The 95% confidence interval of strength estimate difference, $\hat{\theta}_{\text{LAL}} - \hat{\theta}_{\text{ATL}}$, is (-3.73, 3.96). So there is no evidence that Lakers is a better team than Hawks at the 95% confidence level.

(e) In the first round of the post-season playoffs, the Milwaukee Bucks were paired against the Miami Heat, with the Bucks heavily favored to win the best-of-seven series. Shockingly, the Heat won the series in five games.

i. Determine the approximate normal distribution for the score difference of a game between these two teams when playing at the Fiserv Forum (in Milwaukee). Do the same for a game played at the Kaseya Center (in Miami).

Let y_{MIL}, y_{MIA} be the score difference between Bucks and Heat when played in Milwaukee and Miami, repsectively.

$$\mathbb{E}(y_{\text{MIL}}) = \hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}} + \hat{\beta}_h$$
$$\mathbb{E}(y_{\text{MIA}}) = \hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}} - \hat{\beta}_h$$

The standard error, ε, is given by

$$\begin{aligned}
\text{Var}(y_{\text{MIL,MIA}}) &= \text{Var}(\hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}}) + \text{Var}(\hat{\beta}_h) + 2\text{Cov}(\hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}}, \hat{\beta}_h) + \hat{\sigma}^2 \\
&= 3.733 + 0.131 + 0.003 + 12.67^2 \\
&= 164.3172 = 12.8186^2 \\
\text{Var}(y_{\text{MIA,MIL}}) &= \text{Var}(\hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}}) + \text{Var}(\hat{\beta}_h) - 2\text{Cov}(\hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}}, \hat{\beta}_h) + \hat{\sigma}^2 \\
&= 3.733 + 0.131 - 0.003 + 12.67^2 \\
&= 164.3115 = 12.8184^2
\end{aligned}$$

We thus have

$$y_{MIL} \sim \mathrm{N}(6.31, 12.8186)$$
$$y_{MIA} \sim \mathrm{N}(-1.24, 12.8184)$$

ii. Based on the score difference distributions, what are the probabilities that the Bucks would win a game at the Fiserv Forum, and at the Kaseya Center?

```
mean.scoreDiff.MIL = Teams.ability.hfa['Milwaukee Bucks', 'Est'] -
  Teams.ability.hfa['Miami Heat', 'Est'] +
  Teams.ability.hfa['HFA', 'Est']
mean.scoreDiff.MIA = Teams.ability.hfa['Milwaukee Bucks', 'Est'] -
  Teams.ability.hfa['Miami Heat', 'Est'] -
  Teams.ability.hfa['HFA', 'Est']

var.diff.theta = Teams.rank['Milwaukee Bucks', 'StdErr'] ** 2 +
  Teams.rank['Miami Heat', 'StdErr'] ** 2 -
  V['Milwaukee Bucks','Miami Heat'] * 2
var.MIL = var.diff.theta + Teams.ability.hfa['HFA','StdErr'] ** 2 +
  (V['Milwaukee Bucks','HFA'] -
     V['Miami Heat','HFA']) * 2 + resid.stderr ** 2
var.MIA = var.diff.theta + Teams.ability.hfa['HFA','StdErr'] ** 2 -
  (V['Milwaukee Bucks','HFA'] -
     V['Miami Heat','HFA']) * 2 + resid.stderr ** 2

p.win.MIL = pnorm(0, mean.scoreDiff.MIL,
                  sqrt(var.MIL), lower.tail = FALSE)
p.win.MIA = pnorm(0, mean.scoreDiff.MIA,
                  sqrt(var.MIA), lower.tail = FALSE)
> p.win.MIL
[1] 0.6887
> p.win.MIA
[1] 0.5384
```

iii. The first round playoff series is a best-of-seven affair, with the first team to win 4 games as the series winner. The higher-ranked team plays at home in games 1, 2, 5 and 7, and the lower-ranked team plays at home in games 3, 4 and 6. With the Bucks being the higher-ranked team, calculate the probability that the Bucks should have won the series. *Hints: (1) It is easier to compute the answer if you act as though all seven games are played (even though the series stops after the fourth win by one team) and then compute the probability of winning at least 4 out of 7 games; (2) Consider the distribution of the number of Bucks wins among games 1, 2, 5 and 7 separately from games 3, 4 and 6.*

The probability that Bucks wins is 0.6887 in Milwaukee and 0.5384 in Miami. We may pretend that all 7 games in the series would be played even though the winner was already determined. Let H, A be the number of games that Bucks wins at home and away, respectively. Then we have

$$H \sim \mathrm{Bin}(4, 0.6887)$$

$$A \sim \text{Bin}(3, 0.5384)$$

$$\begin{aligned} P(\text{Bucks Win}) = P(H + A \geq 4) &= P(H = 1) * P(A \geq 3) \\ &+ P(H = 2) * P(A \geq 2) \\ &+ P(H = 3) * P(A \geq 1) \\ &+ P(H = 4) \\ P(\text{Bucks Win}) &= 0.08309583 * (0.1560984) \\ &+ 0.27577121 * (0.4014397 + 0.1560984) \\ &+ 0.40675799 * (0.3441287 + 0.4014397 + 0.1560984) \\ &+ 0.22498550 * 1 = 0.7584697 \end{aligned}$$

Code:

```
> bucks.win.home = dbinom(1:4, 4, p.win.MIL)
> bucks.win.away = dbinom(1:3, 3, p.win.MIA)
> p.bucks.win = bucks.win.home[4]
> for (i in 1:3){
+   p.bucks.win <- p.bucks.win +
                    bucks.win.home[i] * sum(bucks.win.away[(4-i):3])
+ }
> p.bucks.win
[1] 0.7584697
```

(f) There is a literature on game outcome models in which the home court/field advantage differs by team.[1] The most straightforward way in R to fit this model is to start with the no-intercept and no-HFA model (with the usual $\boldsymbol{X}^*$ matrix included in the model), but include the `Home` factor variable as an extra predictor. Essentially, this will act as though there is a separate additive term for each team.

Run this model on the 2022 NBA game data, making sure you address the games that are not played at the "home" teams' home arena (there are five such games).

i. Perform a partial-F test (using the `anova` command) of whether the model with 30 home-court advantage parameters is significantly better than the model with just one home-court advantage parameter. You may need to review your linear model notes to do this!

```
> H = outer(score$Home, Teams, "==") - 0
> H[score$AtHome==0,] = 0
> score$H = H
> ability.hfa.lm2 = lm(PtsDiff~X_star + H + 0, data=score)

> anova(ability.hfa.lm, ability.hfa.lm2)
Analysis of Variance Table

Model 1: PtsDiff ~ X_star + AtHome + 0
```

[1] See, for example, Glickman, M.E., & Stern, H.S. (1998). A State-Space Model for National Football League Scores. *Journal of the American Statistical Association*, 93, 25-35.

```
Model 2: PtsDiff ~ X_star + H + 0
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1200 192540
2   1171 185114 29    7426.1 1.6199 0.0205 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model with 30 HFA parameters has $p-$value 0.0205. At 0.05 significance level, the null hypothesis is rejected, so we may conclude that this model is significantly better than the model with only 1 HFA parameter.

ii. With the fit of this new model, reperform the analysis in part (e)iii to determine the probability that the Bucks should have won the series against the Heat. How has the probability changed when the model assumes that the home-court advantage varies by team?

Repeat the analysis in part (e). Now we have

$$\mathbb{E}(y_{\text{MIL}}) = \hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}} + \hat{\beta}_{\text{MIL}}$$
$$\mathbb{E}(y_{\text{MIA}}) = \hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}} - \hat{\beta}_{\text{MIA}}$$

The standard error, ε, is given by

$$\begin{aligned}
\text{Var}(y_{\text{MIL,MIA}}) &= \text{Var}(\hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}}) + \text{Var}(\hat{\beta}_{\text{MIL}}) + 2\text{Cov}(\hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}}, \hat{\beta}_{\text{MIL}}) + \hat{\sigma}^2 \\
&= 7.743 + 7.879 - 8.121 + 12.57^2 \\
&= 12.8679^2 \\
\text{Var}(y_{\text{MIA,MIL}}) &= \text{Var}(\hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}}) + \text{Var}(\hat{\beta}_{\text{MIA}}) - 2\text{Cov}(\hat{\theta}_{\text{MIL}} - \hat{\theta}_{\text{MIA}}, \hat{\beta}_{\text{MIA}}) + \hat{\sigma}^2 \\
&= 7.743 + 7.880 - 8.123 + 12.57^2 \\
&12.8679^2
\end{aligned}$$

We thus have

$$y'_{\text{MIL}} \sim \mathcal{N}(5.9468, 12.8679)$$
$$y'_{\text{MIA}} \sim \mathcal{N}(-1.7390, 12.8679)$$

The probability that Bucks wins at home and away are thus

```
> p.win.MIL2 = pnorm(0, mean.scoreDiff.MIL2,
                     sqrt(var.MIL), lower.tail = FALSE)
> p.win.MIL2
[1] 0.6780123
> p.win.MIA2 = pnorm(0, mean.scoreDiff.MIA2,
                     sqrt(var.MIA), lower.tail = FALSE)
> p.win.MIA2
[1] 0.5537505
```

With H and A defined the same as in (e)iii,

$$\begin{aligned} P(\text{Bucks Win}) = P(H + A \geq 4) &= P(H = 1) * P(A \geq 3) \\ &+ P(H = 2) * P(A \geq 2) \\ &+ P(H = 3) * P(A \geq 1) \\ &+ P(H = 4) \end{aligned}$$

```
> bucks.win.home2 = dbinom(1:4, 4, p.win.MIL2)
> bucks.win.away2 = dbinom(1:3, 3, p.win.MIA2)
> p.bucks.win2 = bucks.win.home2[4]
> for (i in 1:3){
+   p.bucks.win2 <- p.bucks.win2 +
                    bucks.win.home2[i] * sum(bucks.win.away2[(4-i):3])
+ }
> p.bucks.win2
[1] 0.758403
```

The new model with HFA parameters varying by team estimates a slightly **lower** probability that Bucks won the series.