Statistical Modelling III
Assignment 1

Gia Bao Hoang - a1814824

Semester 1 2023

Q1

a)

The model matrix:

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & -1 \\ 1 & 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

b)

Let $v_0, v_1, v_2, v_3, v_4, v_5, v_6$ be the columns of X and $c_1, c_2, c_3, c_4, c_5, c_6$ be constants.

We have:

$$c_1v_1 + c_2v_2 + c_3v_3 + c_1v_4 + c_5v_5 + c_6v_6$$

If $c_1 = c_2 = c_3 = c_4 = c_5 = c_6 = 1$, then:

$c_1v_1 + c_2v_2 + c_3v_3 + c_1v_4 + c_5v_5 + c_6v_6 = v_1 + v_2 + v_3 + v_4 + v_5 + v_6$

$$= \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -1 \\ 0 \\ 0 \\ 1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$\therefore$ The columns of X are not linearly independent.

In addition, this can also be shown through the formula.

Let $v_1 = (\mu, \alpha_1, \alpha_2, ..., \alpha_6)$. With Xv_1 have:

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i - \alpha_j + e_{ijk} \\ &= \mu + \alpha_i - \alpha_j + e_{ijk} + 1 - 1 \\ &= \mu + (\alpha_i + 1) - (\alpha_j + 1) + e_{ijk} \end{aligned}$$

And with Xv_2 where $v_2 = (\mu, \alpha_1 + 1, \alpha_2 + 1, ..., \alpha_6 + 1)$, we can produce $\mu + (\alpha_i + 1) - (\alpha_j + 1) + e_{ijk}$

$\therefore Xv_1 = Xv_2$

$\therefore$ The columns of X are not linearly independent.

c)

If $\alpha_1 = 0$, we do not estimate α_1 anymore. Hence, we can remove the second column of X. The new design matrix X:

$$X = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 & 0 & -1 \\ 1 & 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 1 \end{bmatrix} \therefore rref(X) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Based on the $rref(X)$, there is a pivot in every columns. Hence, the columns of the new X is linear independence.

d)

We will look at cases where Team 1 plays against Team 2. The following is the difference in scores between 2 teams. The first one is when Team 1 is the home team, while the second one is when Team 2 is the home team:

$$\begin{aligned} y_{12k} &= \mu + \alpha_1 - \alpha_2 + e_{12k} \\ \therefore E[y_{12k}] &= E[\mu + \alpha_1 - \alpha_2 + e_{12k}] \\ &= E[\mu] + E[\alpha_1] - E[\alpha_2] + E[e_{12k}] \\ &= \mu + E[\alpha_1] - E[\alpha_2] \\ y_{21k} &= \mu + \alpha_2 - \alpha_1 + e_{21k} \\ \therefore E[y_{21k}] &= E[\mu + \alpha_2 - \alpha_1 + e_{21k}] \\ &= E[\mu] + E[\alpha_2] - E[\alpha_1] + E[e_{21k}] \\ &= \mu + E[\alpha_2] - E[\alpha_1] \end{aligned}$$

In both cases, we can see that the strength of the teams does not change regardless whether they are home or away team. And in both case, the parameter μ "assists" the strength of the home team. As a result, the parameter μ can be considered to be the `home ground advantage`.

e)

- $\alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6$ are respectively the strength of team B,C,D,E,F with relative to team A (α_1).
- In context, the null hypothesis states that there is no difference in the strength of team B, C, D, E, F and the strength of team A.

f)

With the constraint $\alpha_1 = 0$, we can remove the columns for the parameter α_1. We have:

$$X_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} X_2 = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \therefore X_1 - X_2 = \begin{bmatrix} 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The only difference between X and $X_1 - X_2$ is the first column, or the intercept column. While in X, the first column is full of value 1, the first column in $X_1 - X_2$ is full of value 0.

Q2

Load the packages

```
pacman::p_load(tidyverse)
```

a)

- Read data in.

```
df <- read.csv("AFL2019.csv")
```

- Record Home.Team and Away.Team as factors

```
df <- df %>%
  mutate(
    Home.Team=as.factor(Home.Team),
    Away.Team=as.factor(Away.Team)
  )
```

- List of AFL teams in 2019

```
AFL_teams_2019 <- unique(df$Home.Team)
AFL_teams_2019
```

```
##  [1] Carlton           Collingwood       Melbourne         Adelaide Crows
##  [5] Western Bulldogs  Brisbane Lions    St Kilda          GWS Giants
##  [9] Fremantle         Richmond          Sydney Swans      Essendon
## [13] Port Adelaide     Geelong Cats      West Coast Eagles North Melbourne
## [17] Hawthorn          Gold Coast Suns
## 18 Levels: Adelaide Crows Brisbane Lions Carlton Collingwood ... Western Bulldogs
```

- The reference level will be the first level of the factor in alphabetical order. In this case, there are 18 levels and team Adelaide Crows will be used as reference level if the standard factor coding is used.

b)

- Add new column `difference`.

```
new_df <- df %>%
  mutate(difference = Home.Score - Away.Score)
head(new_df)
```

```
##   Round       Location        Home.Team         Away.Team Home.Score Away.Score
## 1     1            MCG          Carlton          Richmond         64         97
## 2     1            MCG      Collingwood      Geelong Cats         65         72
## 3     1            MCG        Melbourne     Port Adelaide         61         87
## 4     1  Adelaide Oval   Adelaide Crows          Hawthorn         55         87
## 5     1 Marvel Stadium Western Bulldogs      Sydney Swans         82         65
## 6     1          Gabba   Brisbane Lions West Coast Eagles        102         58
##   difference
## 1        -33
## 2         -7
## 3        -26
## 4        -32
## 5         17
## 6         44
```

c)

The model matrix X will be constructed based on question 1f. First 2 matrices X1 and X2 will be constructed. Then X will be (X1-X2) exclude the intercept.

```
X1 <- model.matrix(difference ~ Home.Team, data=new_df)
X2 <- model.matrix(difference ~ Away.Team, data=new_df)
X <- (X1-X2)[,-1]
```

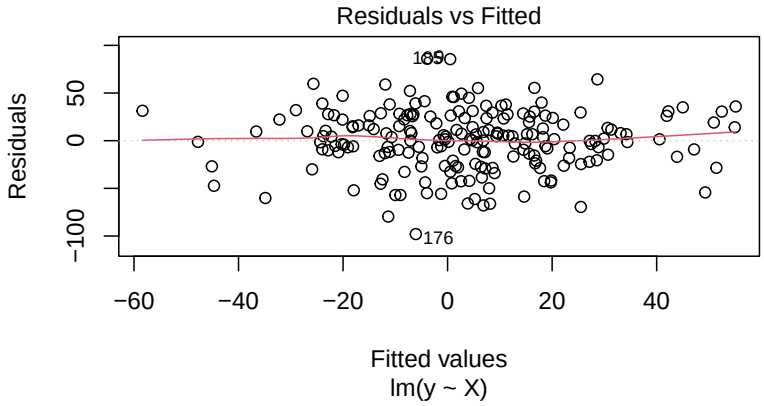
d)

- Fit the model M

```
y <- new_df$difference
M <- lm(y ~ X)
```

- Plot the residuals vs fitted values plot

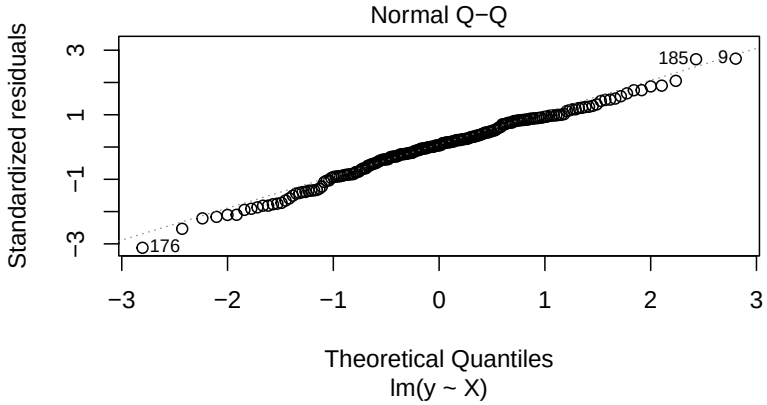
```
par(mar = c(9, 5, 2, 3))
plot(M, which=1)
mtext("This plot shows the relationship between the strength of home and away
      teams and the difference in score. The red line represents the mean score
      difference value.", side = 1, line = 8)
```



This plot shows the relationship between the strength of home and away teams and the difference in score. The red line represents the mean score difference value.

- Plot the normal quantile plot

```
par(mar = c(9, 5, 2, 3))
plot(M, which=2)
mtext("This plot shows a normal quantile plot of 198 random samples from a
      standard normal distribution. The red line represents the expected
      quantiles if the data were normally distributed.", side = 1, line = 8)
```



This plot shows a normal quantile plot of 198 random samples from a standard normal distribution. The red line represents the expected quantiles if the data were normally distributed.

Regression assumptions

Linearity: The residuals are roughly randomly scattered about the zero line in the residuals versus fitted values plot, apart from slight curvature near the endpoints. Hence, the linearity assumption is close to reasonable.

Homoscedasticity: The spread about the zero line appears roughly constant in the residuals versus fitted values plot. Hence, the assumption of constant variance is reasonable.

Normality: There is some departure from normality in both tails of the distribution of residuals. However, the majority of the data is close to normally distributed. Hence, normality assumption is reasonable.

Independence: The plots can not verify this assumption.

e)

The estimated home team effect is `3.682`. Since the p-value is `0.1174` (> 0.05), the effect is not statistically significant. The estimated home team effect is the intercept in the model. We can look at the model summary.

```
summary(M)
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.927 -18.233   1.647  23.677  85.908
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.682      2.340   1.573   0.1174
## XHome.TeamBrisbane Lions       12.892      9.738   1.324   0.1872
## XHome.TeamCarlton             -14.758      9.738  -1.516   0.1314
## XHome.TeamCollingwood          12.703      9.748   1.303   0.1942
## XHome.TeamEssendon             -2.705      9.755  -0.277   0.7818
## XHome.TeamFremantle            -5.852      9.729  -0.602   0.5482
## XHome.TeamGeelong Cats         23.498      9.541   2.463   0.0147 *
## XHome.TeamGold Coast Suns     -38.552      9.532  -4.044 7.77e-05 ***
## XHome.TeamGWS Giants           10.227      9.747   1.049   0.2955
## XHome.TeamHawthorn              8.704      9.738   0.894   0.3726
## XHome.TeamMelbourne           -17.828      9.738  -1.831   0.0688 .
## XHome.TeamNorth Melbourne       1.738      9.739   0.178   0.8586
## XHome.TeamPort Adelaide         4.924      9.540   0.516   0.6064
## XHome.TeamRichmond              9.823      9.747   1.008   0.3149
## XHome.TeamSt Kilda            -16.710      9.531  -1.753   0.0813 .
## XHome.TeamSydney Swans         -1.692      9.747  -0.174   0.8624
## XHome.TeamWest Coast Eagles     9.190      9.541   0.963   0.3367
## XHome.TeamWestern Bulldogs      7.049      9.745   0.723   0.4704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.93 on 180 degrees of freedom
## Multiple R-squared:  0.2947, Adjusted R-squared:  0.2281
## F-statistic: 4.424 on 17 and 180 DF,  p-value: 1.337e-07
```

```
broom::tidy(M)
```

```
## # A tibble: 18 x 5
##    term                        estimate std.error statistic   p.value
##    <chr>                          <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                     3.68      2.34     1.57  0.117
##  2 XHome.TeamBrisbane Lions       12.9       9.74     1.32  0.187
##  3 XHome.TeamCarlton             -14.8       9.74    -1.52  0.131
##  4 XHome.TeamCollingwood          12.7       9.75     1.30  0.194
##  5 XHome.TeamEssendon             -2.71      9.76    -0.277 0.782
##  6 XHome.TeamFremantle            -5.85      9.73    -0.602 0.548
##  7 XHome.TeamGeelong Cats         23.5       9.54     2.46  0.0147
##  8 XHome.TeamGold Coast Suns     -38.6       9.53    -4.04  0.0000777
##  9 XHome.TeamGWS Giants           10.2       9.75     1.05  0.295
## 10 XHome.TeamHawthorn              8.70      9.74     0.894 0.373
## 11 XHome.TeamMelbourne           -17.8       9.74    -1.83  0.0688
## 12 XHome.TeamNorth Melbourne       1.74      9.74     0.178 0.859
## 13 XHome.TeamPort Adelaide         4.92      9.54     0.516 0.606
## 14 XHome.TeamRichmond              9.82      9.75     1.01  0.315
## 15 XHome.TeamSt Kilda            -16.7       9.53    -1.75  0.0813
## 16 XHome.TeamSydney Swans         -1.69      9.75    -0.174 0.862
## 17 XHome.TeamWest Coast Eagles     9.19      9.54     0.963 0.337
## 18 XHome.TeamWestern Bulldogs      7.05      9.75     0.723 0.470
```

f)

We can look for these values in the model summary. The F-statistic is `4.424`, with`17` numerator degrees of freedom and `180` denominator degrees of freedom. The p-value is `1.337e-07`. Since the p-value < 0.05, there is sufficient evident to reject the null hypothesis. In context, there is at least one team with a different strength from team Adelaide Crows.

```
summary(M)
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.927 -18.233   1.647  23.677  85.908
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.682      2.340   1.573   0.1174
## XHome.TeamBrisbane Lions      12.892      9.738   1.324   0.1872
## XHome.TeamCarlton            -14.758      9.738  -1.516   0.1314
## XHome.TeamCollingwood         12.703      9.748   1.303   0.1942
## XHome.TeamEssendon            -2.705      9.755  -0.277   0.7818
## XHome.TeamFremantle           -5.852      9.729  -0.602   0.5482
## XHome.TeamGeelong Cats        23.498      9.541   2.463   0.0147 *
## XHome.TeamGold Coast Suns    -38.552      9.532  -4.044 7.77e-05 ***
## XHome.TeamGWS Giants          10.227      9.747   1.049   0.2955
```

```
## XHome.TeamHawthorn             8.704      9.738   0.894   0.3726
## XHome.TeamMelbourne          -17.828      9.738  -1.831   0.0688 .
## XHome.TeamNorth Melbourne      1.738      9.739   0.178   0.8586
## XHome.TeamPort Adelaide        4.924      9.540   0.516   0.6064
## XHome.TeamRichmond             9.823      9.747   1.008   0.3149
## XHome.TeamSt Kilda           -16.710      9.531  -1.753   0.0813 .
## XHome.TeamSydney Swans        -1.692      9.747  -0.174   0.8624
## XHome.TeamWest Coast Eagles    9.190      9.541   0.963   0.3367
## XHome.TeamWestern Bulldogs     7.049      9.745   0.723   0.4704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.93 on 180 degrees of freedom
## Multiple R-squared:  0.2947, Adjusted R-squared:  0.2281
## F-statistic: 4.424 on 17 and 180 DF,  p-value: 1.337e-07
```

```
broom::glance(M)
```

```
## # A tibble: 1 x 12
##   r.squ~1 adj.r~2 sigma stati~3 p.value    df logLik   AIC   BIC devia~4 df.re~5
##     <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <int>
## 1   0.295   0.228  32.9    4.42 1.34e-7    17  -963. 1965. 2027. 195207.     180
## # ... with 1 more variable: nobs <int>, and abbreviated variable names
## #   1: r.squared, 2: adj.r.squared, 3: statistic, 4: deviance, 5: df.residual
```

g)

- The estimated home team effect is `3.682`, the estimated strength of the Brisbane Lions is `12.892` while the estimated strength of the Carlton is `-14.758`. We can predict the expected difference in score by substituting the values into the model M.

```
y0 <- 3.682 + 12.892 - (-14.758)
y0
```

```
## [1] 31.332
```

- If the Brisbane Lions play at home against Carlton, the Lions will win by roughly 32 points