Assignment_1

Zhao Ming Soh a1751699

```
pacman::p_load(tinytex, cowplot, gridExtra)
```

Question 1

1a.

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

1b.

$$\text{Row reduced echelon of X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The row reduced echelon of matrix X does not have pivots in every single column, this tells us that the constants are not equal to the trivial solution of 0 which means that the column matrix of X are not linearly independent.

1c.

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

$$\text{Row reduced echelon X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The row reduced echelon of X with $\alpha_1 = 0$ has pivots in every columns except for α_1, this tells us that the columns of matrix X are linearly independent as the only way of solving the matrix is by setting the constants equal to 0 which is the trivial solution that underpins the definition of a linearly independent matrix.

1d.

$$\begin{aligned} E[y_{ijk}] &= E[\mu + \alpha_i - \alpha_j + \epsilon_{ijk}] \\ &= E[\mu] + E[\alpha_i] - E[\alpha_j] + E[\epsilon_{ijk}] \\ &= u + \alpha_i - \alpha_j \qquad\qquad \text{given that } E[\epsilon_{ijk}] = 0 \end{aligned}$$

Expected score of team 1 as home team vs team 2 as away team:

$$E[y_{12k}] = \mu + \alpha_1 - \alpha_2$$

Expected score of team 1 as away team vs team 2 as home team:

$$E[y_{21k}] = \mu + \alpha_2 - \alpha_1$$

The parameter μ can be considered as the home ground advantage because the error term, $E[\epsilon_{ijk}]$ is equal to 0, this means that there is no difference in the team strength of home team against away team. Therefore, the μ is the only term that is left which can then be interpreted as the expected home advantage of home team against the away team.

1e.

The null hypothesis, H_o is interpreted as the team strength of $B = \alpha_2, C = \alpha_3, D = \alpha_4, E = \alpha_5, F = \alpha_6, D = \alpha_7$ are equal to each other with respect to $A = \alpha_1$.

1f.

$$X_1 \text{ is the design matrix of the home team}$$

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

$$X_2 \text{ is the design matrix of the away team}$$

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

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

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

Matrix X and the resulting matrix of $X_1 - X_2$ are exactly the same except for the intercept term, μ which is cancelled out in the resulting matrix of $X_1 - X_2$. This means that the home team effect is no longer presence in the resulting matrix of $X_1 - X_2$.

Question 2

Setup

```
pacman::p_load(tidyverse)
AFL2019 <- read_csv("AFL2019.csv")
head(AFL2019)
```

2a.

```
sapply(AFL2019, class)
```

```
##       Round    Location   Home.Team   Away.Team  Home.Score  Away.Score
##   "numeric" "character" "character" "character"   "numeric"   "numeric"
```

```
AFL2019$Home.Team <- as.factor(AFL2019$Home.Team)
AFL2019$Away.Team <- as.factor(AFL2019$Away.Team)
summary(AFL2019$Home.Team)
```

```
##    Adelaide Crows    Brisbane Lions           Carlton       Collingwood
##                11                11                11                11
##          Essendon        Fremantle      Geelong Cats   Gold Coast Suns
##                11                11                11                11
##        GWS Giants          Hawthorn         Melbourne  North Melbourne
##                11                11                11                11
##     Port Adelaide          Richmond          St Kilda     Sydney Swans
##                11                11                11                11
## West Coast Eagles  Western Bulldogs
##                11                11
```

```
summary(AFL2019$Away.Team)
```

```
##    Adelaide Crows    Brisbane Lions           Carlton       Collingwood
##                11                11                11                11
##          Essendon         Fremantle      Geelong Cats   Gold Coast Suns
##                11                11                11                11
##        GWS Giants          Hawthorn         Melbourne   North Melbourne
##                11                11                11                11
##     Port Adelaide          Richmond          St Kilda     Sydney Swans
##                11                11                11                11
## West Coast Eagles  Western Bulldogs
##                11                11
```

The standard reference level for the both teams is Adelaide Crows.

2b.

```
AFL2019 <- AFL2019 %>%
  mutate(Difference = Home.Score - Away.Score)
head(AFL2019$Difference)
```

```
## [1] -33  -7 -26 -32  17  44
```

2c.

```
X_1 <- model.matrix(~ Home.Team, data = AFL2019)
X_2 <- model.matrix(~ Away.Team, data = AFL2019)
X <- (X_1 - X_2)
X <- subset(X, select = -c(1))
```

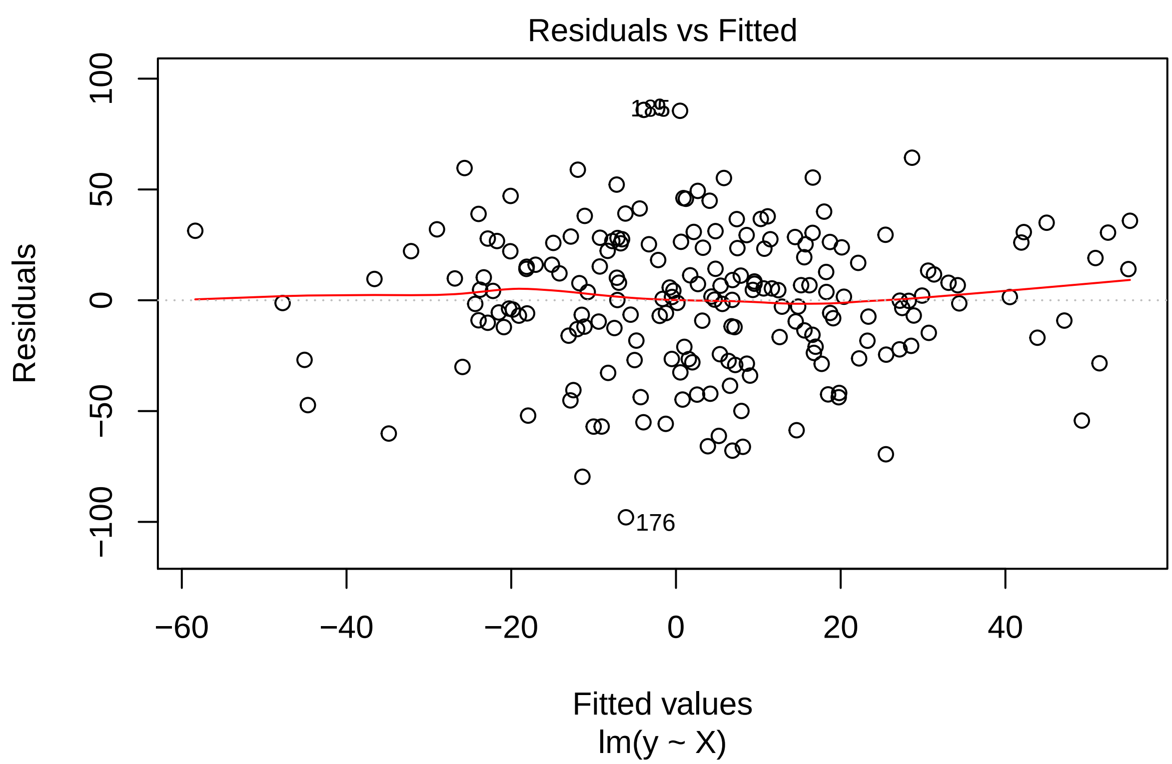
2d.

```
y <- AFL2019$Difference
AFL2019_lm <- lm(y ~ X)
summary(AFL2019_lm)
```

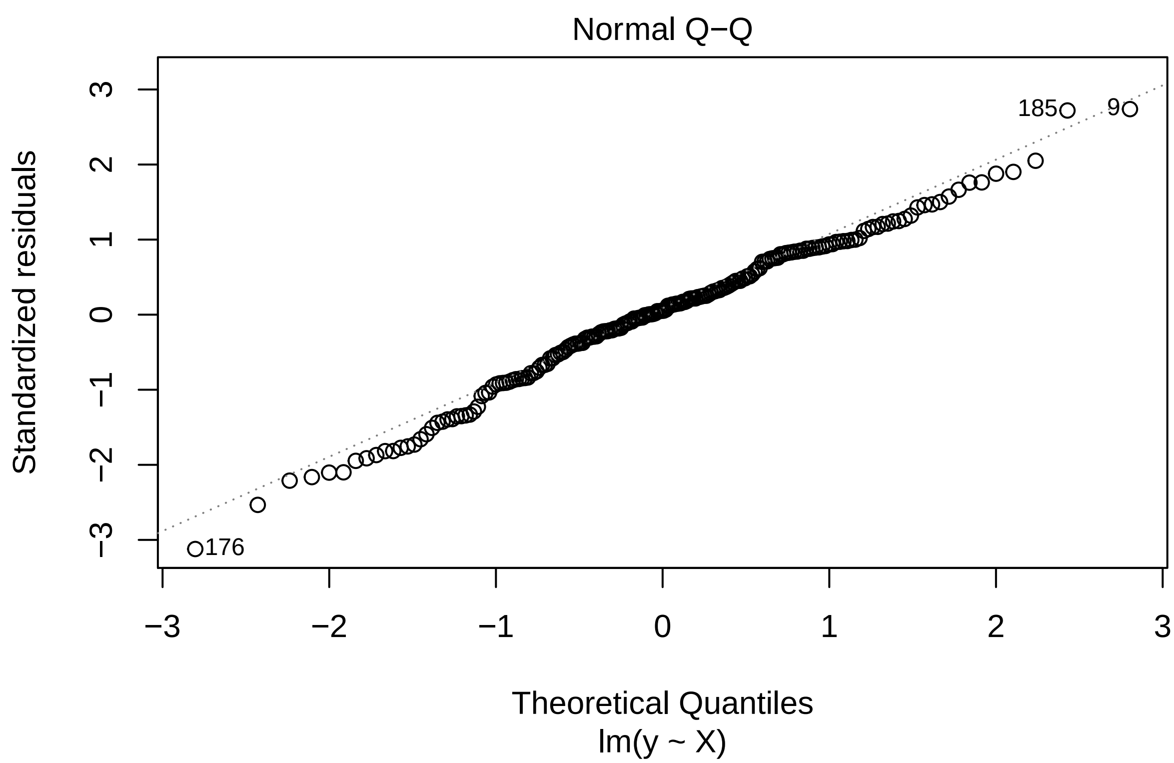
```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -97.927 -18.233   1.647  23.677  85.908
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
```

```
## (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
```

```r
p_1 <- plot(AFL2019_lm, which = 1)
```



```r
p_2 <- plot(AFL2019_lm, which = 2)
```



Linearity :
The assumption of linearity seems to be reasonable as the points are scattered randomly above and below the red line and the red line is approximately horizontal around the zero line.

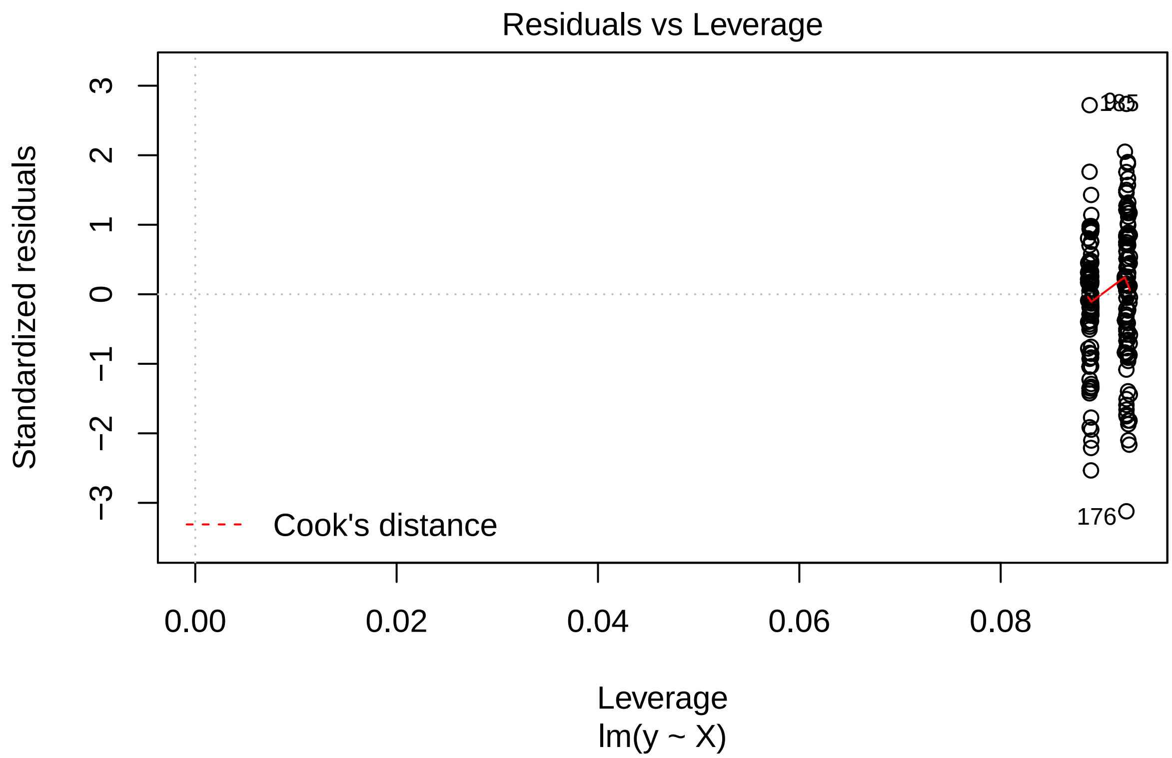
Normality :
The assumption of normality seems to also be reasonable because the majority of the data points are scattered around the middle of the line apart from the deviation of a couple of data points on each end of the line.

Homoscedasticity :
The assumption of homoscedasticity seems to be reasonable because the spread of the data points around the zero line is approximately constant.

Independence :
Becuase there is not enough information given in terms of how the data are collected, we cannot make any conclusion about the assumption of independence for the data.

2e.

```r
plot(AFL2019_lm, which=5)
```



2f.

```r
estimated_coefficient <- coef(AFL2019_lm)
estimated_coefficient
summary(AFL2019_lm)
```

The estimated home team effect is essentially the intercept term which is equal 3.682 when all predictors are equal to 0. Given that the p-value of the estimated home team effect is 0.1174 which is greater than the 0.05 significance level, this tells us that the home team effect is not statistically significant.

2g.

```r
summary(AFL2019_lm)
```

```
##
## 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
```

The F-statistic for the model is 4.424 with degree of freedoms of 17 and 180. Given that the p-value is $1.337 \times 10^{-}7$ is smaller than the significance level of 0.05, we can conclude that the home ground advantage in our model is statistically significant that there is indeed a difference in the points scored between teams playing in their home ground vs the away ground.

2h.

Using the equation :

$$y_{ijk} = \mu + \alpha_i - \alpha_j + \epsilon_{ijk}$$

```r
expected_num_of_points <- ceiling(3.682 + 12.892 - (-14.758))
expected_num_of_points
```

```
## [1] 32
```

$$\begin{aligned} y &= 3.682 + 12.892 - (-14.758) \\ &= 31.332 && \text{round to the nearest point} \\ &= 32 \end{aligned}$$

The expected number of points that the Brisbane Lions will win by is 32.