STATS 3001 / STATS 4104 / STATS 7054
Statistical Modelling III
Assignment 1
2022

DEADLINE:

- Friday 11th March 2021 5pm (Week 2)

QUESTIONS:

1. **AFL home-ground advantage theory**

   In this assignment, we are going to examine if there is a home ground advantage in the AFL.

   We will need to develop a more complicated model that the usual linear regression, and so we cannot just use the built in function `lm()`, we will need to use some coding to fit it.

   In this question, we will develop the correct design matrix X for our model, and in the next question we will apply our method to some real data.

   Set up

   Consider a football competition with six teams, A, B, C, D, E, F. Each team has its own home ground and the following games were played in the first four weeks of the season.

Table 1: Home and away table.

Week	Home_team	Away_team
1	A	D
1	B	E
1	C	F
2	D	B
2	E	C
2	F	A
2	A	E
3	B	F
3	C	D
3	D	C
4	E	A
4	F	B

Let y_{ijk} denote the difference between the points scored by the home team, i, and the points scored by the away team, j, when the two teams played on the k^{th} occasion.

For example

$$y_{141}$$

is the difference in scores between Team A playing at home and Team D playing away in the first week, *i.e.* first row of table.

Consider the linear model

$$M : y_{ijk} = \mu + \alpha_i - \alpha_j + e_{ijk},$$

where it is assumed that the e_{ijk} are uncorrelated and

$$E[e_{ijk}] = 0, \quad Var(e_{ijk}) = \sigma^2.$$

In the equation

$$\alpha_i$$

is the **strength** of the home team, and

$$\alpha_j$$

is the **strength** of the away team.

(a) Write down the model matrix for the given home and away table (Table 1), assuming the data ordering above and without imposing constraints on the parameters $(\mu, \alpha_1, \alpha_2, \ldots, \alpha_6)$.

(b) Show that the columns of X are not linearly independent and explain how this is evident in the formulation

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

Hint: consider adding a constant to the α's

(c) Show that the columns of X are linearly independent if the constraint $\alpha_1 = 0$ is imposed.

(d) Explain why the parameter μ can be considered to be the **home ground advantage**.
Hint: Compare the expected score in favour of Team 1 at a home game and at an away game playing in both cases against Team 2.

(e) Interpret the parameters $\alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6$ and the hypothesis

$$H_0 : \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$$

in context.

(f) Consider the models

$$M_1 : y_{ijk} = \mu + \alpha_i + e_{ijk} \text{ and } M_2 : y_{ijk} = \mu + \alpha_j + e_{ijk}$$

both with the constraint $\alpha_1 = 0$ and let X_1 and X_2 be the corresponding model matrices. If X is the model matrix for the model M in Part (c), then explain the relationship between X and $X_1 - X_2$.

2. **AFL home-ground advantage coding**

So now we use the ideas from Q1 to code this up and find if there is a home ground advantage in the AFL.

The file `AFL2019.csv` contains the results from the 198 AFL games played during the home and away season. The variables recorded are

Variable_Name	Description
Round	Round numbers from 1-23
Location	Venue where match played
Home.Team	The home team
Away.Team	The away team
Home.Score	The total points scored by the home team
Away.Score	The total points scored by the away team

(a) Read the data into R. Obtain a list of the AFL teams in 2019. Which team will be used as reference level if the standard factor coding is used?

(b) Add a new column, `difference`, to the data frame that contains the difference between the home team and away team scores. Include the relevant R code and the first 6 lines of the data frame.

(c) Consider the model, M, introduced in Question 1a. Calculate a matrix, X, containing all of the necessary columns of the model matrix, except the intercept.
Hint 1: Use the result of Question 1f to construct the X matrix. Note also that you can use the X matrix in an `R` model formula, for example,

`y~X`.

Hint 2:: `model.matrix()` is your friend.

(d) Fit the model, M. Obtain residuals vs fitted values plot and also a normal quantile plot of the residuals. Comment on whether the regression assumptions appear reasonable.

(e) Obtain the residuals vs leverage plot of the data and comment on whether there are any influential points. Explain also whether there are any points of high leverage.

(f) Based on this analysis, what is the estimated home team effect? Is the effect statistically significant?

(g) Test the hypothesis

$$H_0 : \alpha_2 = \alpha_3 = \ldots = \alpha_{17} = 0.$$

State the F-statistic, the degrees of freedom and the P-value as well as your final conclusion.

(h) If the Brisbane Lions play at home against Carlton, what is the expected number of points (round to the nearest point) that the Lions will win by?