Data. The results of all the current 2023-24 Premier League season of football games is available from the website http://www.football-data.co.uk/englandm.php in CSV format. A copy (downloaded 15:00 4th Feb 2024) is also on the Ultra site in case of website issues.

Download the data, load it into R and examine the data available. There are a large number of covariates for each match, including the teams, scoreline and various statistics about shots, fouls, offsides, etc. The full explanation of all variables is at http://www.football-data.co.uk/notes.txt.

Model. We would like to fit a simple model which predicts the probability of a home win. We will assign to team i a 'strength' variable, β_i. Then, if team i plays at home and team j plays away, we model the odds on a home win as:

$$\frac{\pi}{1-\pi} = e^{\beta_i - \beta_j}$$

This is called a Bradley-Terry model and is widely used for pairwise comparisons such as sporting encounters.

We can treat this as a logistic regression and use all the tools of generalised linear models as follows. Let $y = 1$ if the home team wins and $y = 0$ if they do not win (lose or draw). Let there be a predictor x_i for each team, with:

$$x_i = \begin{cases} 0 & \text{if team } i \text{ was not playing in the game} \\ +1 & \text{if team } i \text{ was playing at home} \\ -1 & \text{if team } i \text{ was playing away} \end{cases} \tag{1}$$

Then, the GLM with:

- linear predictor: $\eta = \beta_1 x_1 + \cdots + \beta_p x_p$;
- response function: $h(\eta) = \frac{e^\eta}{1+e^\eta}$;
- distributional assumption: $Y \mid x, \beta \sim \text{Bernoulli}(\pi(x))$, where $\pi(x) = h(\eta(x))$ and the further assumption that the outcome of each match is conditionally independent of the others given x, β;

corresponds to the standard Bradley-Terry model. After that lengthy build-up, we turn to modelling this in R.

Tasks.

(a) The first challenge is to construct a design matrix representing the setup in equation (1). Remember R automatically creates dummy variables for factors, so we can use it to do the work for us. Read the help file for the `model.matrix` function, then use it to create two matrices of indicators for home (`X_home`) and away (`X_away`) teams.

Combine both matrices to create a single matrix `X` with entries per equation (1). Be sure to turn this into an R data frame so that the variables are named.

Finally, since we use indicator variables for factors in the model, we need to drop one team to avoid a singular solution; we drop the first team alphabetically, which is Arsenal.

(b) Create the vector y holding the home team win status (1 or 0). Note, home team win is 1, so both draw and away team win are 0.

Combine this with `X` into a single data frame called `matchdata`. This is now a data frame we can use to fit the model we want.

(c) Fit the Bradley-Terry model via a logistic regression using the `glm` function in R and examine the model.

Sort the coefficients and compare it to the league table https://www.premierleague.com/tables. Is there complete agreement? (Remember, Arsenal implicitly has coefficient 0). Can you think what might explain any differences?

(d) What do you expect the dispersion parameter ϕ to be under a logistic regression model?

In any GLM, the deviance has expected value equal to the degrees of freedom, $n-p$. Use this fact to estimate the dispersion for the model here. Is it overdispersed, about right, or underdispersed?

(e) It is often said that there is a 'home team advantage'. How can we model this?

Fit this model. Does it suggest a home team advantage is detectable in this dataset?

(f) Construct a plot which shows the 95% confidence region for the strengths of Man City and Liverpool jointly, using the $(1-\alpha)$ Hessian CR method.

(g) At the end of a season, the bottom three teams in the Premier League table are relegated (i.e., dropped down into the lower division called the Championship). As of 4th February, that is Everton, Burnley and Sheffield United. We are interested in testing the hypothesis that the three teams currently at the bottom of the league are equally weak: that is, might $\beta_{\text{Everton}} = \beta_{\text{Burnley}} = \beta_{\text{SheffieldUnited}}$?

Fit the model where these three teams have the same coefficient and apply a likelihood ratio test against the model from (e). Recall that we just saw in the last lecture that the deviance is $2\phi(\ell_{\text{sat}} - \ell(\hat{\beta}))$, so you do not need to compute the log-likelihood manually (recall, what is ϕ in logistic regression?)

(h) On 17th February, Chelsea will play away from home against Man City. What is the probability of Chelsea winning this match? (Note: **do not** place any wagers based on this model which is far too simple to beat bookie's odds!!)

(i) Typically, if we want to predict with a model like this, we cannot assess the accuracy of the model on the same data it was fitted on. Hence, we typically use just some of the data for 'training' the model and hold some aside for 'testing' in order to ascertain an unbiased estimate of accuracy. The accuracy is unbiased because the model has never 'seen' the testing data.

Fit the model using the first 150 matches. Then, for the unused data points, produce a table showing the number of actual home/away wins against the predicted number of home/away wins. Then, compute the percentage accuracy as:

$$\text{Accuracy} = \frac{(\#\text{ correctly predicted home wins}) + (\#\text{ correctly predicted away wins})}{\text{Total number of games}} \times 100$$

(j) If you have any time left, experiment with creating other models using the same data (perhaps create a rescaled Binomial model for the number of goals, rather than a pure binary home/away win model, or repeating the split of data into traning and test sets using a value other than 150).