Sports Analytics: Predicting the outcome of Football matches

Aneesh Patel

Department of Mathematics

College of Engineering, Design and Physical Sciences, Brunel University

Supervisor: Anne-Sophie Kaloghiros

March 18, 2020

Abstract

Predicting the outcome of football matches is an interesting area of Mathematics and has many useful applications, especially in the betting industry. There are many distributions available such as the Poisson or Negative Binomial distribution that can be used to model the number of goals teams will score in a football match. By estimating the attacking strengths and defensive weaknesses of various teams we can predict the number of goals teams are expected to score against each other based on previous matches. The results of a football match can depend on many factors such as the quality of the teams, the league or tournament the teams are playing in and whether any of the teams have injuries to key players. An interesting factor investigated in this project is the idea of a "home effect" as teams playing at their home stadium are likely to have an advantage.

A popular method used to predict the results of games is Poisson regression since the number of goals scored can be thought to be a Poisson variable as Maher describes in his 1982 paper Modelling Association Football Scores. The results of a subset of games from this seasons Premier League (2019/20) have been predicted using previous data as part of this project. The built-in functions and packages available in R allow for a Poisson Regression model to be easily implemented when compared to other statistical software. Excel has been used to assess the accuracy of predictions. While regression models can be used to predict the outcomes of football matches there will be a lot of variation in the number of goals teams score which cannot be explained by these models such as a red card during the game where one team would have to play with one less player than the other team.

Acknowledgements

I would like to thank my supervisor Anne-Sophie Kaloghiros for her support and guidance throughout this project, friends at university and my parents for their positivity and keeping me motivated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

Modelling and predicting football games is an interesting area of Mathematics and has important applications in the sports industry particularly in betting. The premier league is the highest tier league in England and one of the most popular in the world. There are 20 teams in the league with 380 games being played in total. Each team plays each other twice, once at home and once away. Home games are played at the team's stadium which can give that team an advantage called the "home effect". The home effect could consist of many factors such as having many home supporters at the stadium, or not having to travel elsewhere for the game and being familiar with pitch conditions. The home effect factor is more significant when playing in competitions such as the champions league where the away team may need to travel to another country in Europe. Figure 1.1 shows the percentage of games won by the home team, away team or drawn from the 380 games played in the 2018/19 premier league season. The green section represents the percentage of games won by the home team which almost makes up half of the results highlighting the "home effect".

At the end of the season, the top 4 teams are awarded a place in the champions league and the last 3 teams are relegated to the second division. Staying in the premier league is important for teams competing due to the vast amount of money in the league. The Premier League includes some of the richest football clubs in the world. Deloitte's Football Money League which ranks the wealthiest club teams across all leagues internationally listed seven Premier League clubs in the top 20 for the 2009–10 season, and all 20 clubs were in the top 40 globally by the end of the 2013–14 season, largely as a result of increased broadcasting revenue [1].

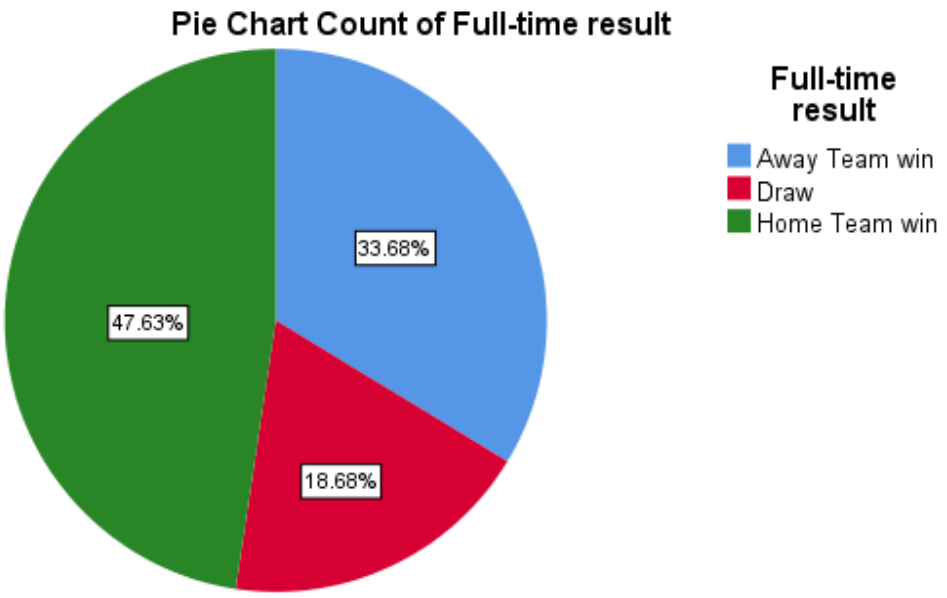

Figure 1.1: Pie chart of results

Betting on football matches is something that has gained a lot of popularity and has become a lucrative business. Many of the major bookmakers such as Sky and William Hill offer a variety of bets such as outright predicting the result, the final score or something more obscure such as the number of corners a team gets. Bets can also be personalised to include a combination of events such as the team to win and a certain player from that team to score the first goal. A popular example is Sky Sports Super 6 where the user must predict 6 correct scores plus the minute of the first goal in any of the six games to win the jackpot of £250,000. The substantial amount of prize money indicates the value of being able to predict football scores. Predicting football scores also has uses in other areas such as coaching and journalism. For my project I will be looking at how the Poisson distribution can be used to model the number of goals scored by the home and away team and will explore this further in chapter 2.

1.2 Data

The majority of analysis I have done is on data from the current 2019/20 Premier League season. I downloaded an excel data set from [2] and simplified it before importing into R. For the data analysis I only required the name of the home and away team and the number of goals each team scored against each other. There is a row in the data set for each game that took place and I have included a screenshot of the first few rows from the data set in appendix A.1. Here FTHG and FTAG represent the number of goals the home and away team scored respectively. As I will use a Poisson distribution to model my data and Poisson regression in order to the predict outcomes of football matches it is required to duplicate each row in my data set since each match is essentially two observations, one for the number of goals the home team scores

and one for the away team therefore needing two rows in the data set instead of one as shown in A.2.

To help me arrange the data into the required form I adapted a piece of code from [3]. I added a column called "Home" displayed in A.3 with number 1 assigned to teams playing at home and 0 for teams playing away. The variables Team, Opponent and Home are used to predict the number of goals each team scores using Poisson regression.

1.3 Objectives

1. Implement and extend Maher's independent Poisson model in R and use estimated parameters to predict the number of goals teams will score against each other. Predict full time scores of a selection games in the second half of this current Premier League season by using data from the first half (190 out of 380) games of this season.

2. Assess how well the model fits the data and the accuracy of the predictions it makes for the selection of games the model is used for. Discuss the strengths and weaknesses of the model and assess the model fit by running statistical tests on R.

3. Suggest any improvements to the model if there was more time to work on the project by assessing the accuracy of predictions and taking into account the limitations of current models that can be used to predict the outcomes of football matches.

1.4 Literature Review

1.4.1 Maher - Modelling association football scores

Maher modelled the home and away goals separately as two independent Poisson distributions [4]. Since the expected number of goals scored is different for each team a Poisson distribution with a variable mean is used. Maher models the number of goals the home and away team will score using four separate parameters. If team i is playing team j at home then the observed score will be (X_{ij}, Y_{ij}).

$X_{i,j} \sim Poisson(\alpha_i \beta_j)$ and $Y_{i,j} \sim Poisson(\gamma_i \delta_j)$. Here the parameter α_i represents the attacking strength of team i at home, β_j represents the defensive weakness of team j away from home, γ_i represents the defensive weakness of team i at home and δ_j represents attacking strength of team j away from home. The α_i parameter is estimated from the observed values for $X_{i,k}$ as k varies across all the games team plays i plays at home and the β_j parameter is estimated from the observed values for $X_{k,j}$ as k varies. Similarly, the γ_i and δ_j parameters are estimated from the observed values of $Y_{i,k}$ and $Y_{k,j}$ respectively. The greater the α_i and δ_j parameters are the

more goals the home and away team are expected to score respectively. Likewise, the smaller the γ_i and β_j parameters are the lower the number of goals the home and away team will concede respectively. Since there are 20 teams in the Premier league, 80 such parameters will have to be estimated in order to model and predict scores. Maher estimated parameters using the method of maximum likelihood.

He determined that the independent Poisson model gives a reasonably good fit to the data. However, the model deviated when predicting the number of low games scoring games (both teams scoring 1 or less goals) and of high scoring games (one or both teams score 3 or more goals). In both cases the model predicted less than the observed number of high or low scoring games and this suggests that independence can not always be assumed. Bi-variate Poisson models were then used to model dependence between scores, especially for low scoring games where both teams may have found it difficult to score due to reasons such as poor pitch condition.

1.4.2 Dixon and Coles - Modelling Association Football Scores and Inefficiencies in the Football Betting Market

Dixon and Coles aimed to exploit potential inefficiencies in the betting market by developing a statistical model which is capable of accurately predicting the probabilities of outcomes of football matches to form the basis of a profitable betting strategy [5]. The probabilities calculated can be compared with bookmakers to find "value bets" where there is a positive expected return. In section 4 of their paper, they present their refinement on Maher's model which at the time was the only paper taking into account the different qualities of teams involved. Dixon and Coles statistical model included various features to help them develop a profitable betting strategy.

A key feature not present in Maher's model was the introduction of a "home effect." In their model $X_{i,j} \sim Poisson(\alpha_i \beta_j \gamma)$ and $Y_{i,j} \sim Poisson(\alpha_j \beta_i)$ where $X_{i,j}$ is the number of goals the home team scores and $Y_{i,j}$ the number of goals the away team scores. For this model, $X_{i,j}$ and $Y_{i,j}$ are independent, α represents the attacking strengths and β represents the defensive weakness of teams. The α and β parameters now only depend on the teams and no longer on whether they are playing home or away. The new parameter $\gamma > 0$, represents the advantage that the team playing at home has. This parameter would be equal to 1 when both teams are playing at a neutral stadium and there is no home team.

Maher has a more general specification than this, allowing for separate home and away, attack and defence parameters for each team. The addition of a home effect parameter as specified in Dixon and Coles paper simplifies Maher's model as there are less parameters needing to

be estimated. Another improvement to Maher's model they mentioned is the incorporation of a time perspective, so that matches played a long time ago have less influence on parameter estimates allowing for dynamic parameters compared to them being static over time. A teams performance is likely to be closer related to performances in recent matches compared to earlier games due to circumstances such as injuries to players or bans and suspensions for players. The betting strategy they mentioned in their paper is to bet on all outcomes where the ratio of model to bookmakers probabilities exceeds a certain level to yield a positive expected return, even allowing for the in-built bias in the bookmakers odds.

Chapter 2

Methodology

2.1 Poisson Distribution

The Poisson distribution is used to model the number of events occurring within a given time interval. A random variable X $\sim$ Poisson (λ) where the parameter λ is the expected value for the number of occurrences of an event in a certain time period. The formula for the Poisson probability mass function is $P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!}$. In a game of football, each time a team has possession there is a chance for the team to attack and score. The probability p that an attack leads to a goal is small and the number of times that a team has possession n in a game is very large. If the attacks are independent of each other and p is constant, the number of goals scored in a game are binomial and since n is large, the Poisson approximation to a Binomial distribution can be applied. Maher suggests using a Poisson distribution with a mean which varies according to the quality of the team to model the number of goals scored by a team in a football match [4].

The Poisson distribution has its mean = variance = λ and skewness of $\frac{1}{\sqrt{\lambda}}$. The maximum likelihood estimator of λ is $\bar{X}$ where $\bar{X}$ is the sample mean. For the 2018/19 Premier League season the average number of goals scored by home teams was 1.57. Using this sample mean as an estimator for λ the probability that a team would score 0 goals is $P(X = 0) = \frac{e^{-1.57}\lambda^0}{0!} = 0.208$. I multiplied this probability by 380 (number of games in a season) to find the expected number of games where the home team would score 0 goals which was 79.1 games, in comparison the observed number of games was 88. I calculated this for r = 0,1,2,...,6 and have shown my results in table 2.1.

Figure 2.1 shows the distribution of the goals scored by home teams in the 2018/19 Premier League season. The mean number of goals is 1.57 and the variance is 1.72.

Number of goals (R)	Actual	Probabilities	Predicted
0	88	0.208	79.1
1	116	0.327	124.1
2	95	0.256	97.4
3	48	0.134	51.0
4	22	0.053	20.0
>=5	11	0.022	8.4

Table 2.1: Probabilities for number of goals scored

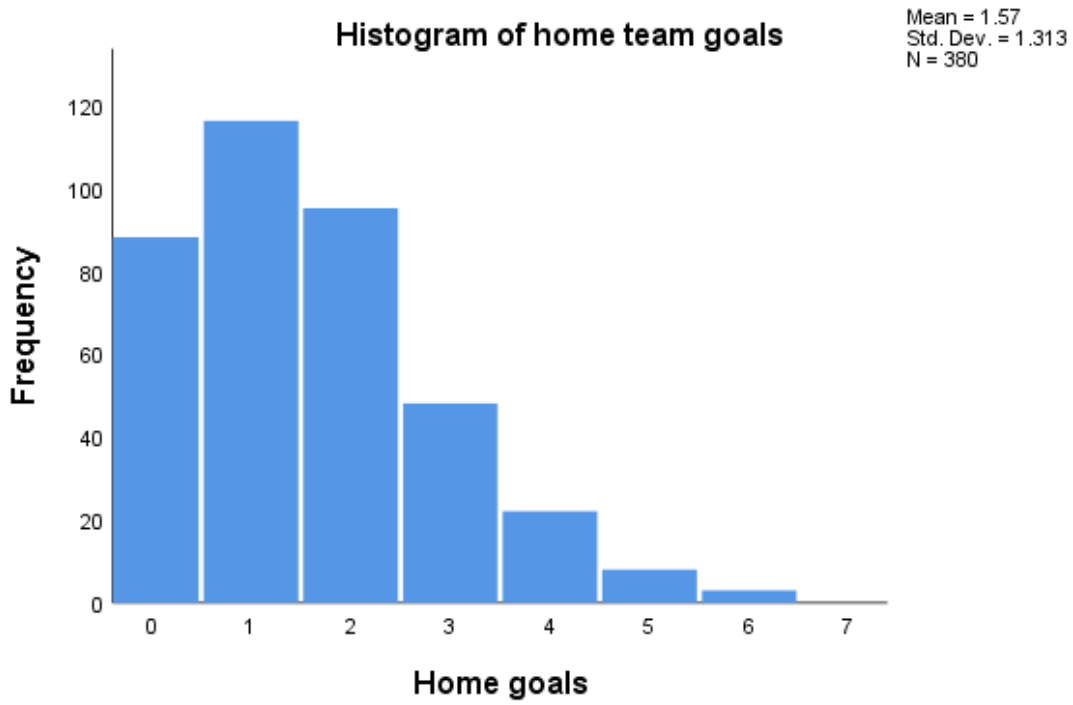

Figure 2.1: Histograms of goals scored by home teams in 2018/19

A graphical representation of a Poisson distribution with a $\lambda = 1.57$ created using Excel is displayed in figure 2.2 to compare with the histogram shown above in figure 2.1. Comparing these histograms we can see that both plots are skewed to the right and the distribution representing the number of goals scored by home teams has a shape similar to a Poisson distribution.

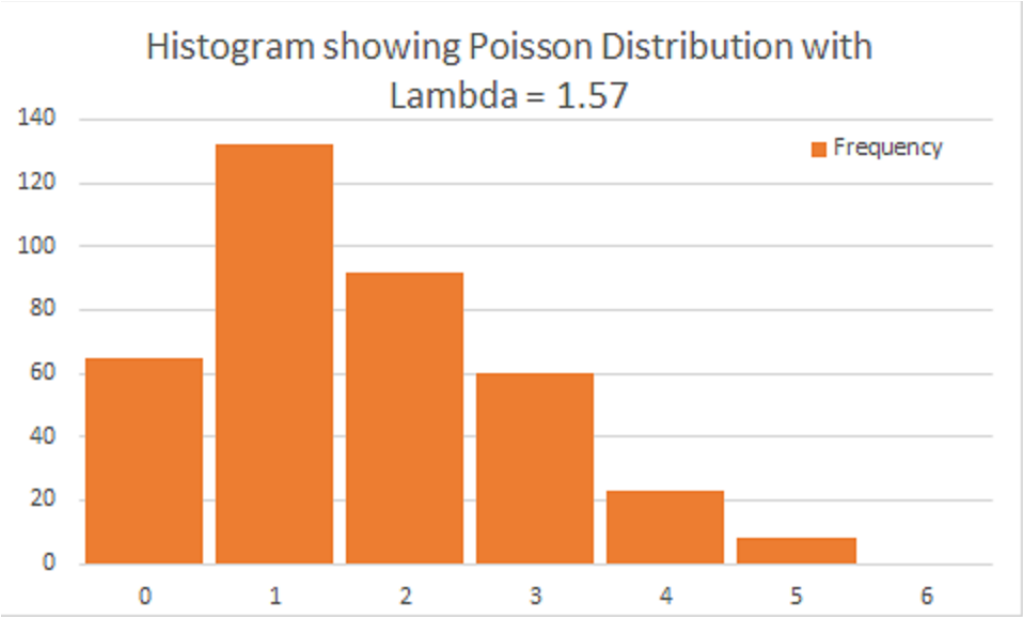

Figure 2.2: Graphical representation of Poisson distribution

2.1.1 Chi-squared goodness of fit test

The Chi-squared goodness of fit test can be used to assess how well the Poisson distribution fits the data set. The term goodness of fit is used to compare the observed sample distribution with the expected probability distribution. The Chi-squared statistic has formula

$$\chi^2 = \sum_{r=0}^{6} \frac{(O_r - E_r)^2}{E_r} \tag{2.1}$$

where O_r is the observed number of times r number of goals were scored by the home team in 380 games for the 2018/19 Premier League season and E_r is the expected number of times which I calculated using the Poisson distribution formula. The χ^2 statistic follows a chi-squared distribution and has k - p degrees of freedom where k is the number of classes required to group the number of goals scored as shown in table 2.1 and p is the number of parameters estimated. In this case there are 6 classes, and 1 parameter (sample mean from data set) has been estimated, therefore the degrees of freedom is 6 - 1 = 5. My hypotheses for this goodness of fit test with α = 0.05 are

$H_0 : X \sim Poisson$

$H_1 : X$ does not follow a Poisson distribution.

Applying the Chi-squared test in R as shown in figure A.4 using the formula shown in equation 2.1 with the values displayed in table 2.1, I get a χ^2 statistic = 2.79 with 5 degrees of freedom and a p-value = 0.7321. This p-value is equal to the P($\chi^2 > 2.79$) = 0.7321 and can be represented by the shaded probability shown in figure 2.3.

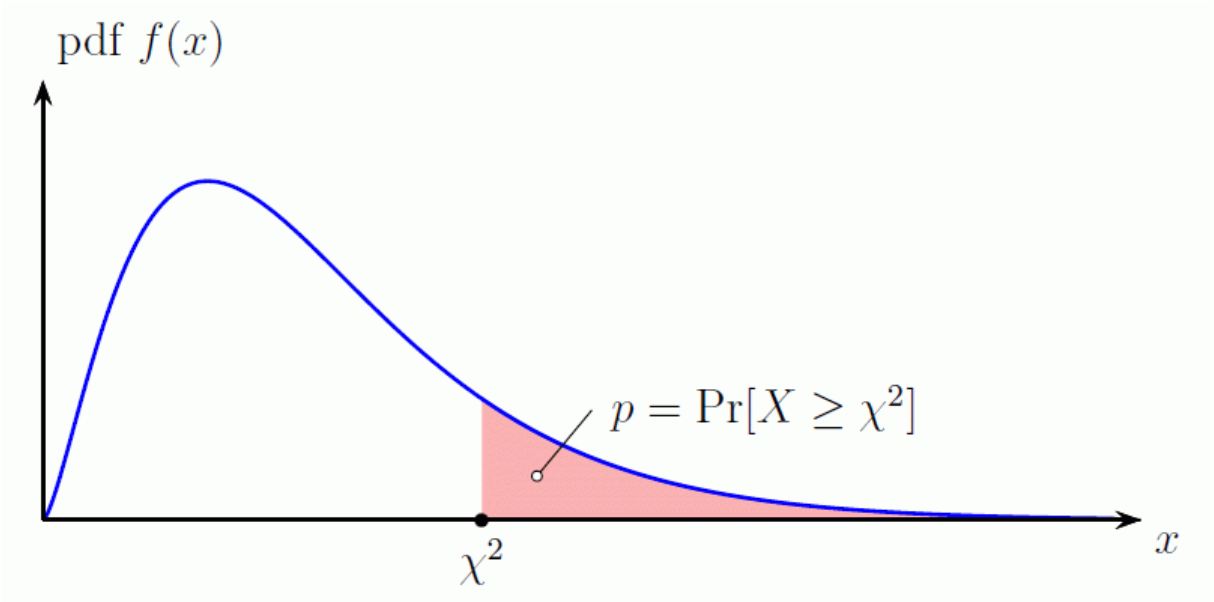

Figure 2.3: Chi-squared distribution

This p-value is insignificant for any level of α. We conclude that there is no real evidence to suggest the data does not follow a Poisson distribution since the p-value > α and the Poisson distribution can be used to model the number of goals scored by football teams. In chapter 3 of my report I will show how Poisson regression can be used to predict the number of goals teams will score using previous data.

2.2 Generalised Linear Models

Generalised Linear Models (GLM's) refer to a class of models where the response variable is assumed to follow an exponential family distribution. Generalised Linear Models are most commonly used to model binary or count data and can be fitted in R using the glm function which is similar to the lm function for fitting linear models. There are a variety of regression models available under the glm function such as Poisson, negative binomial and the zero-inflated model. There are 3 main arguments required to call a glm function in R: the *formula* which is how the response variable is defined in terms of its covariates, the *family* which is the specified distribution of the response variable and *data*, which is the name of the data set R is using to fit the glm. All specified variables must be in the R work space or in the data frame passed to the data argument. There are several methods available for accessing components of the glm object including residuals(), predict(), coef() and summary(). These methods display the required information for the glm in the R console window when the object name I have used to store my model is entered inside the brackets.

2.2.1 Poisson Regression

The generalised linear model I will be using to model my data and predict results is the Poisson regression model. In Poisson regression, the response variable is a count variable with a Poisson distribution and is modelled as a linear function of its covariates. The model I am implementing in R is Maher's independent Poisson model where there are 2 parameters for each team (α, β) and a general home effect parameter γ which Dixon and Coles suggested as an improvement to Maher's model. For my model, the response variable is the number of goals a team is expected to score in a football match which I have shown earlier to be approximately Poisson distributed. The covariates for this model are the team playing, the opponent and home advantage. My model for the expected number of goals scored by the home team is defined as

$$Log(\lambda^H) = intercept + home + team_i + opponent_j. \tag{2.2}$$

The model for the expected number of goals scored by the away team is defined as

$$Log(\lambda^A) = intercept + team_j + opponent_i \tag{2.3}$$

where λ is the expected number of goals, "home" is the impact playing at home has on the number of goals scored, "team" is the attacking strength of a team and "opponent" is the defensive weakness of a team. The logarithm on the left hand side is the link function for the Poisson family and ensures the expected number of goals is never negative. Figure 2.4 shows how the model is defined and called using R.

```
Console ~/
> model1

Call:  glm(formula = Goals ~ Home + Team + Opponent, family = poisson(link = log),
    data = newdata1)
```

Figure 2.4: Poisson Regression model

The summary(model) function in R displays the estimates obtained for my model parameters. The co-coefficients have to be exponentiated before they can be interpreted due to the logarithm on the left hand side in equations 2.2 and 2.3. To find the expected number of goals for a team playing at home (λ^H) we need to compute $exp(intercept + home + team_i + opponent_j)$ and for the away team (λ^A) we need to compute $exp(intercept + team_j + opponent_i)$.

2.3 Maximum Likelihood Estimation

Maximum likelihood estimation is a method of estimating the parameters of a probability distribution by maximising a likelihood function so that the parameters estimates best represent the

data set. The log likelihood function is then differentiated and equated to 0, and we are left with equations that the estimates of the parameters satisfy. The log is taken of the likelihood function since the log is monotonically increasing so maximising the likelihood function is same as maximising the log likelihood function. An iterative technique enables the equations to be solved and the maximum likelihood estimates for these parameters to be found. The probability distribution I am considering is the Poisson distribution.

Maher says the number of goals scored by the home team $X_{ij} \sim Poisson(\alpha_i\ \beta_j)$ where the mean is a combination of the parameters α_i and β_j multiplied together [4]. The parameter α_i represents the attacking strength of team i and β_j represents the defensive weakness of team j. The log likelihood function for the goals scored by the home team is given by

$$logL(\alpha,\beta) = \sum_i \sum_{j\neq i} \Big(log(e^{-\alpha_i\ \beta_j}) + log(\alpha_i\ \beta_j)^{\ x_{ij}} - log(x_{ij}!) \Big) \tag{2.4}$$

$$= \sum_i \sum_{j\neq i} \big(-\alpha_i\ \beta_j + x_{ij}\ log(\alpha_i\ \beta_j) - log(x_{ij}!) \big) \tag{2.5}$$

The estimates for the parameters are the ones that make equation 2.5 maximal given the goals scored in the observed games. Differentiating this equation gives

$$\frac{\partial logL}{\partial \alpha_i} = \left(\sum_{j\neq i} -\beta_j + \frac{x_{ij}}{\alpha_i} \right) \tag{2.6}$$

and equating equation 2.6 to 0 gives us the equations which the maximum likelihood estimates satisfy.

$$\hat{\alpha}_i = \frac{\sum_{j\neq i} x_{ij}}{\sum_{j\neq i} \hat{\beta}_j}\ and\ \hat{\beta}_j = \frac{\sum_{i\neq j} x_{ij}}{\sum_{i\neq j} \hat{\alpha}_i}. \tag{2.7}$$

An iterative technique can be used to solve equations 2.7 for $\hat{\alpha}_i$ and $\hat{\beta}_j$. Maher suggested in his paper to use the Newton-Raphson method to find the maximum likelihood estimates for the model parameters. R calls the number of iterations required to fit a glm Fisher scoring iterations and the number is displayed in the console window as part of the model summary.

Chapter 3

Analysis

3.1 Model Summary

The model summary function on R displays the model parameters estimated by maximum likelihood estimation under the coefficients section and provides information on the model deviance's. These parameters are for Maher's independent Poisson model with the addition of a home effect parameter as described in Dixon and Coles model. I have defined this model as "model1" in R and this is the model I will be using to predict the outcome of football matches. The coefficients will need to be exponentiated before they can be interpreted as the model uses a log link function. The exponentiated coefficients are displayed below in figure 3.1.

```
Console ~/
> exp(coef(model1))
             (Intercept)                     Home          TeamAston Villa          TeamBournemouth
               1.3415475                1.1086917                0.9871353                0.7521113
            TeamBrighton              TeamBurnley              TeamChelsea       TeamCrystal Palace
               0.9016684                0.9116279                1.2960336                0.6673571
             TeamEverton            TeamLeicester            TeamLiverpool             TeamMan City
               0.8264683                1.6529121                1.9342579                2.0363761
          TeamMan United             TeamNewcastle              TeamNorwich     TeamSheffield United
               1.1927257                0.7640685                0.7657047                0.8848404
         TeamSouthampton             TeamTottenham               TeamWatford             TeamWest Ham
               0.9191888                1.3761592                0.4962297                0.8321852
           TeamWolves       OpponentAston Villa      OpponentBournemouth        OpponentBrighton
               1.1203417                1.1323404                0.9211175                0.9047369
        OpponentBurnley          OpponentChelsea   OpponentCrystal Palace        OpponentEverton
               1.0585076                0.9588902                0.7275819                0.9954059
      OpponentLeicester        OpponentLiverpool          OpponentMan City       OpponentMan United
               0.6114759                0.5194368                0.8339864                0.8199636
      OpponentNewcastle          OpponentNorwich OpponentSheffield United      OpponentSouthampton
               0.9399088                1.3303716                0.6399379                1.2394024
      OpponentTottenham          OpponentWatford          OpponentWest Ham           OpponentWolves
               0.9436950                1.1153924                1.2067879                0.8392166
>
```

Figure 3.1: Exponentiated coefficients

A total of 40 parameters have been estimated; the intercept, home effect, and the attacking strengths (team) and defensive weaknesses (opponent) of the all the teams apart from Arsenal. The team and opponent parameters for Arsenal are not shown since this team is taken to be

the reference group (R by default chooses which team is first alphabetically for a categorical covariates control group). The team and opponent parameters for Arsenal act as a reference and are the baseline for other teams to be compared to. The control group can be changed from the default to any of the 19 other teams using the piece of shown in figure 3.2.

```
14  contrasts(newdata1$Team) <- contr.treatment(n=levels(newdata1$Team), base=10)
```

Figure 3.2: Code to change baseline team

This line of code will set the tenth team (in alphabetical order) as the reference team, and the remaining teams attacking strengths and defensive weaknesses would be compared to this team. For my analysis, I have kept arsenal as the reference team. The incidence rate (exponentiated coefficient) for team = "Aston Villa" is 0.987. An interpretation for this rate is the attacking strength for Aston Villa is 0.987 times the attacking strength for the baseline team (team = "Arsenal") or Arsenal's attacking strength is 1 / 0.987 = 1.013 times better than Aston Villa's. Similarly, the defensive weakness (opponent = "Aston Villa") is 1.132 worse than the reference team (opponent = "Arsenal"). The home effect parameter is estimated to be 1.109 which suggests that playing at home increases the number of goals the home team will score by approximately 10%.

3.2 Deviance

Deviance is a measure of goodness of fit of a glm. The null deviance is a measure of how well the response variable is predicted by a model that includes only the intercept (overall mean number of goals) and no covariates. The residual deviance is a measure of how well the response variable is predicted by the model when the predictor variables are included (home effect, team and opponent) and for a Poisson regression model is defined by the formula shown in equation 3.1. The predictor variables improve the model fit if the residual deviance is significantly lower than the null deviance of the model.

3.2.1 Goodness of fit test

To assess the fit of my model, I will be running a deviance goodness of fit test in R. The residual deviance has n-p degrees of freedom where n is the number of observations and p are the number of parameters estimated, therefore the residual deviance has 380 - 40 = 340 degrees of freedom. In comparison, the null deviance has n degrees of freedom (380). R uses the residual deviance of the model as the test statistic to perform the goodness of fit test for the model.

For a Poisson regression model, the deviance is given by

$$D = 2\sum_{i=1}^{n}\left(Y_i\, log(\frac{Y_i}{\hat{\mu}_i}) - (Y_i - \hat{\mu}_i)\right) \tag{3.1}$$

and follows a chi-squared distribution with n-p degrees of freedom where $\hat{\mu}_i$ is the predicted mean for observation i and Y_i are the observed values. The closer the predicted means are to their observed values, the smaller the residual deviance will be and the better the fit of the model is. The hypotheses for this deviance goodness of fit test are

H_0 : The independent Poisson regression model is a good fit for the data set

H_1 : The independent Poisson regression model is not good fit for the data set.

To calculate the p-value for this goodness of fit test, we calculate the probability to the right of the residual deviance value for the chi-squared distribution on 340 degrees of freedom. R calculates χ^2 statistic = 347.1 and the p-value to be 0.383 which is not significant as shown in figure 3.3. There is no significant evidence to suggest the model is a not a good fit for the data set.

```
Console ~/
> # poisson model
> with(model1, cbind(res.deviance = deviance, df = df.residual,
+                    p = pchisq(deviance, df.residual, lower.tail=FALSE)))
     res.deviance  df         p
[1,]     347.1335 340 0.3831594
> |
```

Figure 3.3: Result of deviance goodness of fit test

The residual deviance gives an indication of whether a Poisson regression model fits well, but we should be wary about using the resulting p-value from the goodness of fit test. Pawiton states in his book "In all likelihood" [6], the test is okay for the Poisson regression model where the Poisson means are not too small which is the case for the number of goals scored in a football match (the means range from 0-9 goals). It would be beneficial to look at other tests in R as the deviance goodness of fit test may be not be reliable as suggested by Jonathan Bartlett in his blog on the Deviance goodness of fit test for Poisson regression [7]. The accuracy of predictions the model makes could also help determine if the independent Poisson regression model is a suitable model for the data set.

3.3 Using model to make predictions

Data from the first half of the current Premier League season (190/380) games has been used to fit my model. The model I am using to conduct my analysis has been stored as the object "model1." At the halfway stage of the season each team has played 19/38 games. By making

use of the "predict" function in R, I am going to predict the outcome of games 191-250 for this season, that is games 20-25 out of 38 for each team. During these 60 games, each team will play 6 games each, 3 and home and 3 away from home. How the predict function has been used to predict game number 240 (West Ham Vs Liverpool) is shown below in figure 3.4. The predict function has to be used twice to find the expected number of goals for the home and away team. The predicted number of goals each team will score against each other is then displayed in the console window.

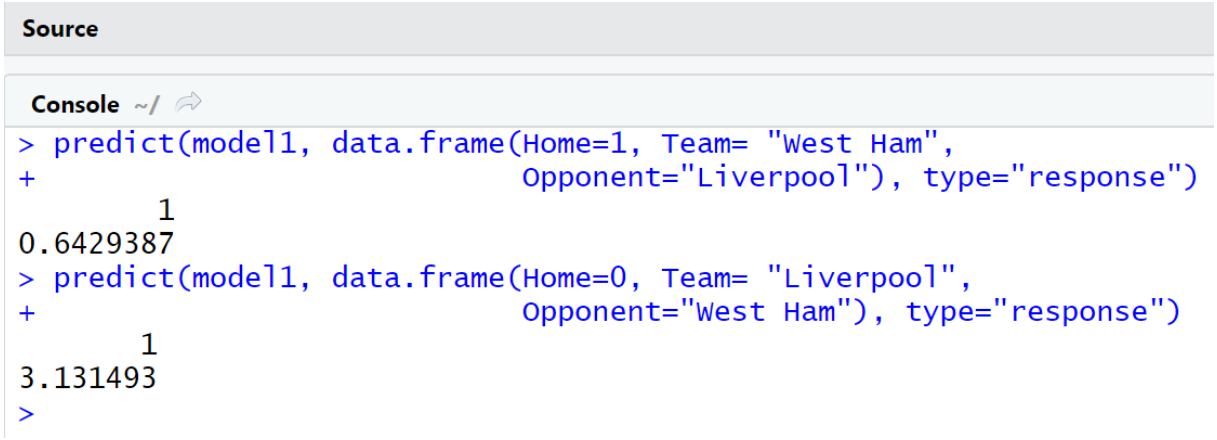

Figure 3.4: Predicting game number 240

Using the expected number of goals for each team, the probabilities for a home win, draw and away win can be calculated using the Poisson distribution function called "dpois" in R. The code to find these probabilities is provided in appendix A.5. The resulting probabilities are shown in figure 3.5 with an away win the most likely outcome.

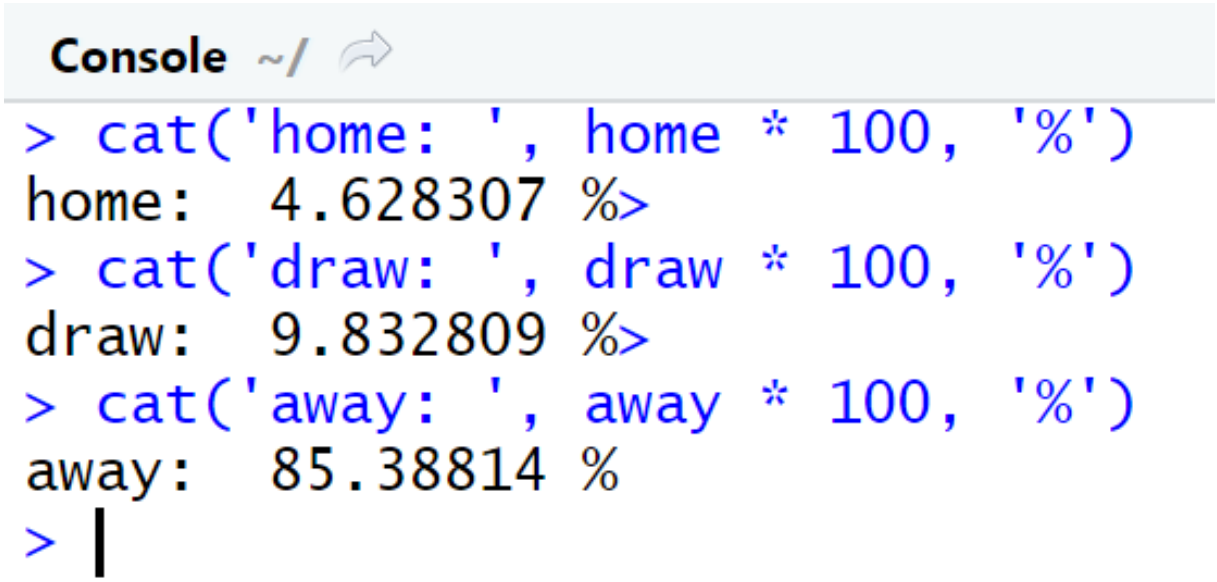

Figure 3.5: Probabilities for the game outcomes

The dpois function in R can also be used to find probabilities for the number of goals each team are expected to score against each other and a plot can be created to model these probabilities as shown in figure 3.6. The code used to generate this plot is provided in appendix A.6.

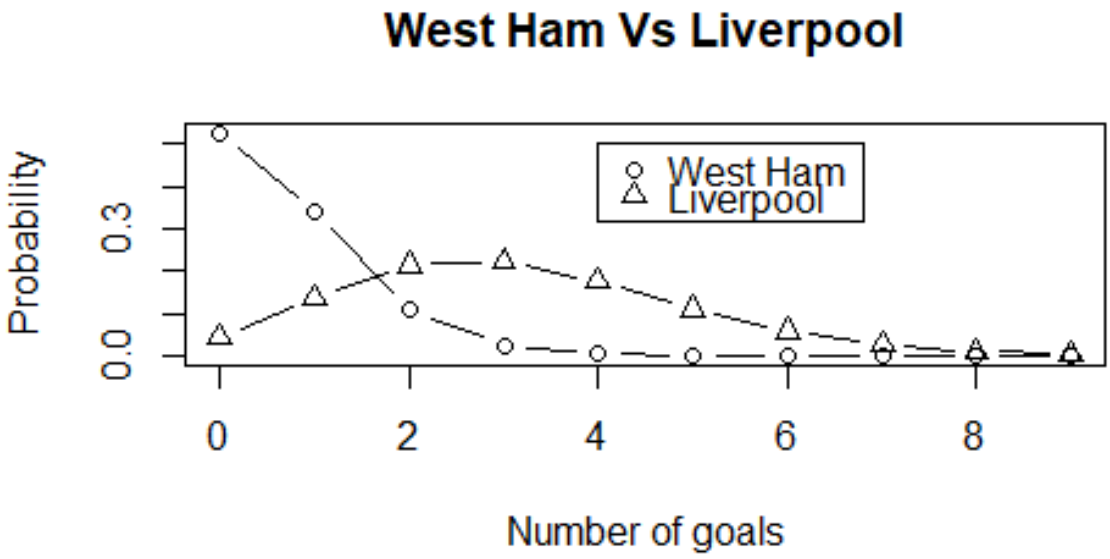

Figure 3.6: Plot showing probabilities for number of goals scored

3.3.1 Skellam distribution

The Skellam distribution is the probability distribution of the difference between two independent Poisson distributed random variable's N1 and N2 with means μ_1 and μ_2 respectively [8]. The probability mass function for the Skellam distribution for difference K = N1 - N2 between two independent Poisson distributed random variables with means μ_1 and μ_2 respectively is given by

$$Pr(K = k) = e^{-(\mu_1+\mu_2)} \left(\frac{\mu_1}{\mu_2}\right)^{\frac{k}{2}} I_k\left(2\sqrt{\mu_1\,\mu_2}\right) \tag{3.2}$$

where I_k is a modified Bessel function [9]. The Skellam distribution has mean μ_1 - μ_2, where μ_1, $\mu_2 \geq 0$ and variance μ_1 + μ_2. The Skellam distribution can be used to model the goal difference where μ_1 is the expected number of goals scored by the home team and μ_2 is the expected number of goals scored for by the away team. The probabilities of the 3 outcomes in football matches, home win, away win, and draw can be represented by looking at the goal difference of home and away goals scored. A difference of exactly 0 representing a draw, a difference greater than 0 representing a home win and a goal difference less than 0 representing an away win.

The "VGAM" package is required to use the Skellam distribution in R. I have plotted a Skellam distribution for game number 240 (West Ham Vs Liverpool) displayed in figure 3.7 which shows that the most probable goal difference is -2 suggesting that Liverpool will win by 2 goals since they are the away team. This is consistent with the plot showing probabilities for the expected number of goals scored for both teams in figure 3.6. The expected goal differences can be used to assess the accuracy of predictions for the games I am predicting. The code to create the Skellam distribution plot is provided in appendix A.7.

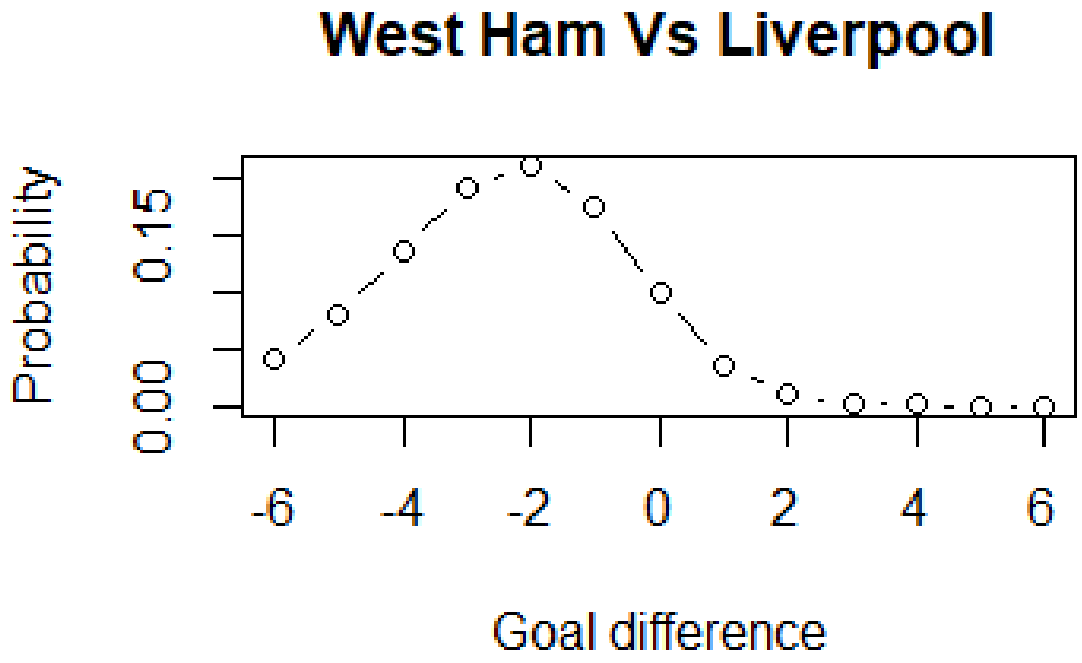

Figure 3.7: Plot showing probabilities of goal differences for West Ham Vs Liverpool

The Skellam distribution is skewed to the left for the game between West Ham and Liverpool since Liverpool are expected to win and they are the away team. I have created another plot where both teams are a similar level i.e. Southampton and Crystal Palace as they were ranked next to each other in the Premier League table at the halfway stage. The plot representing the probabilities for the expected goal difference between Southampton and Crystal Palace in figure 3.8 shows that a draw is the most likely outcome since an expected goal difference of 0 has the highest probability.

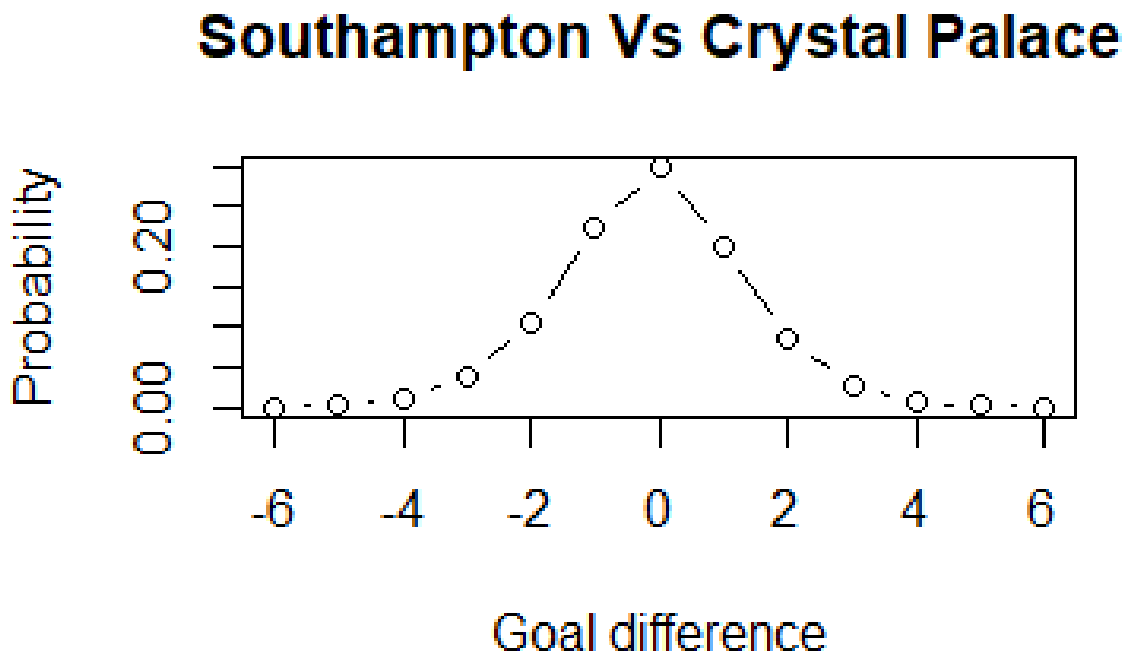

Figure 3.8: Plot showing probabilities of goal differences for Southampton Vs Crystal Palace

For Southampton Vs Crystal Palace, μ_1 = 0.9947 and μ_2 = 1.1096. Here μ_1 and μ_2 are the number of goals Southampton and Crystal Palace are expected to score against each other respectively. The Skellam distribution for this game has a mean equal to 0.9947 - 1.1096 = -0.1149 and variance equal to 0.9947 + 1.1096 = 2.1044. The skewness of the distribution is given by $\frac{\mu_1-\mu_2}{(\mu_1+\mu_2)^{\frac{3}{2}}}$ = -0.03764. The Skellam distribution for this game is almost symmetric since the mean is close to 0.

3.4 Accuracy of predictions

I will be using the goal difference to determine the outcomes of matches and assess the accuracy of predictions the model makes. To compare the predicted results to the actual results, I have created an Excel spreadsheet containing the actual and predicted number of goals for each team and the result for each game.The predicted number of goals for the home and away teams have been rounded to 1 decimal place. There is a screenshot of the data from the first 20 predicted games out of 60 provided in appendix A.8. Here, HG stands for number of goals scored by the home team, AG number of goals for the away team, R for the result of the game and GD for the goal difference (home goals - away goals). A limitation of the model is that it fails to predict any draws since this model does not predict integer numbers for the number of goals teams are expected to score. Despite not predicting any draws, the model managed to predict the outright

result (home or away win) correctly 25/60 times ≈ 41.7%.

Looking back at the history of the Premier League [10] on average, Premier League teams consistently win around 46.2% of games, while the draw occurs around 27.52% of the time and the away team are victorious 26.32% of games. A graph summarising these averages from [10] is displayed below in figure 3.9.

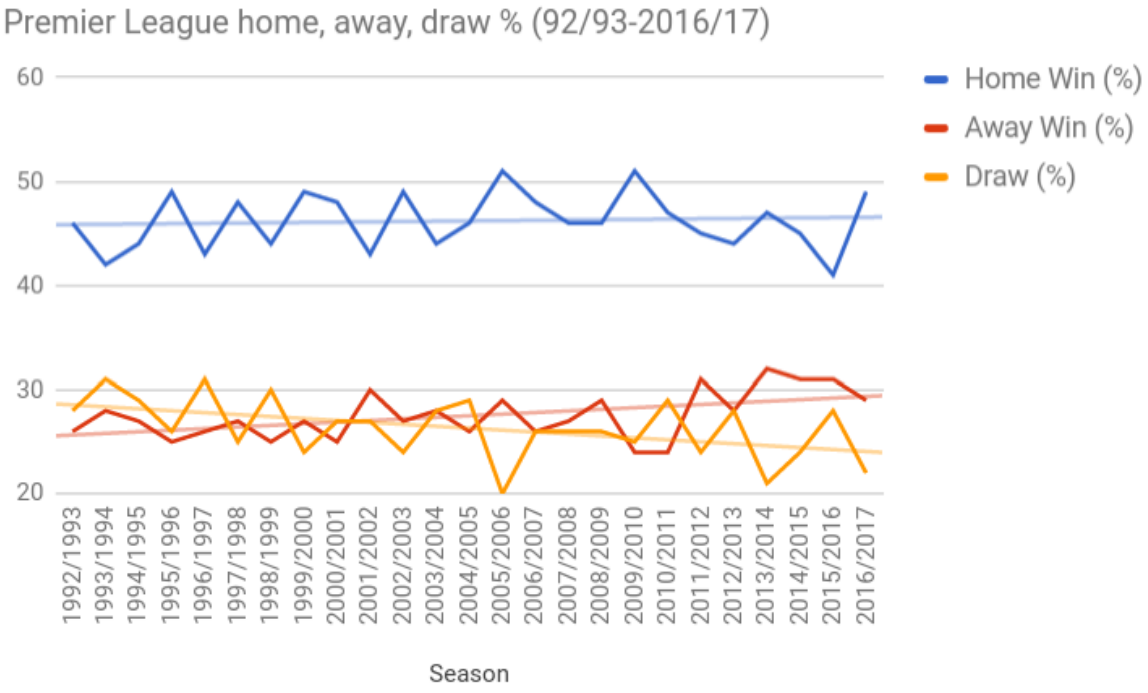

Figure 3.9: Graph showing history of results

To get around this limitation and allow the model to predict draws, I introduced a simple IF statement:

IF GD < -0.1 predict away win,
IF -0.1 ≤ GD ≤ 0.1 predict draw,
IF GD > 0.1 predict home win.

How I implemented this IF statement in Excel is shown below in figure 3.10.

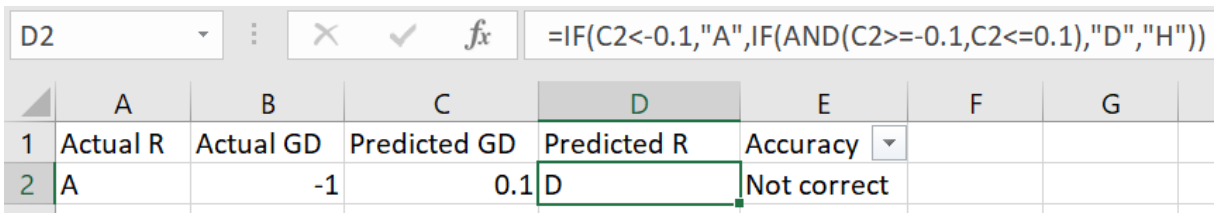

Figure 3.10: IF statement to increase number of draw predictions

After implementing the IF statement and using 0.1 to inflate the number of draws predictions, the model predicts 6 draws and 26/60 results correctly. Table 3.1 shows how using a number greater than 0.1 increases the number of draws predicted and how the accuracy of predictions the model makes improves.

Draw inflation parameter	Number of draws predicted	Accuracy of predictions (%)
0.10	6	26/60 $\approx$ 43.3%
0.15	9	28/60 $\approx$ 46.7%
0.2	11	29/60 $\approx$ 48.3%
0.25	13	29/60 $\approx$ 48.3%
0.30	16	31/60 $\approx$ 51.7%
0.35	19	32/60 $\approx$ 53.3%
0.40	22	33/60 = 55%
0.45	22	33/60 = 55%
0.50	22	33/60 = 55%

Table 3.1: Accuracy of predictions with different draw inflation numbers

We can see that increasing the draw inflation parameter up to 0.40 increases the number of draws predicted and accuracy of predictions. The introduction of this parameter improves the accuracy from 25/60 to 33/60 for correctly predicted games. When 0.3 is chosen as the parameter to inflate the number of draws, the model predicts 16/60 $\approx$ 26.7% of the games to be a draw. This is close to the average number of draws (27.52%) since the Premier League started. In conclusion, 0.3 would be a good choice for the draw inflation parameter as it aids the model to predict a suitable number of draws using data from previous games and currently predicts just over half of the results correctly from 60 games.

3.5 Home effect parameter

The home effect parameter obtained from maximum likelihood estimation for the first half of this current Premier League season was 1.109. A likelihood ratio test can be run the assess the significance of this home effect parameter. The likelihood ratio test is a test for nested GLMs based on the deviance goodness of fit test, in order to see whether or not adding in extra parameters improves the fit of the model. The model including the extra parameter is called the full model, and the model excluding the parameter is called the reduced model, therefore the reduced model is a subset of the full model. The 2 models are defined below:

Reduced model: Goals $\sim$ Team + Opponent

Full model: Goals $\sim$ Home + Team + Opponent

The hypothesis for the likelihood ratio test are: H_0: Home = 0 vs H_1: Home $\neq$ 0.

The test statistic follows a chi-squared distribution with degrees of freedom = difference in the number of parameters between the full and reduced model. When n is large, under H_0 this statistic (χ^2) = Deviance(reduced model) - Deviance(full model). The results of this test are displayed below in figure 3.11.

```
Console ~/
> model2 <- update(model1, . ~ . - Home)
> lrtest(model1,model2)
Likelihood ratio test

Model 1: Goals ~ Home + Team + Opponent
Model 2: Goals ~ Team + Opponent
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  40 -525.48
2  39 -526.18 -1 1.4057     0.2358
>
```

Figure 3.11: Results for Likelihood ratio test

There is no evidence to reject H_0 at any significance level. We are implicitly assuming that the reduced model is a good fit to the data and we are testing whether adding the extra parameter home effect improves the fit of the model. Since we can not conclude that the extra parameter improves fit of the model we should stick to the reduced model.

Chapter 4

Conclusion and Recommendations

4.1 Conclusion

Using data from the first half of the current Premier League season (190/380 games), I was able to implement a model in R using Poisson regression which used maximum likelihood estimation to estimate parameters representing the attacking strength and defensive weaknesses for each team. The mean number of goals each team are expected to score are expressed as a product of these attack and defensive parameters. By making use of the predict function in R, I was able to predict the expected number of goals the home and away team will score against each other and used the goal difference (home goals - away goals) to assess the accuracy of predictions. Altogether, the outcome of 60 games from the current season of the Premier League were predicted (games 191-250 out of 380). A limitation of the model was that it could not predict any draws, since the model does not predict integers for the number of goals teams are expected to score. Introducing an IF statement for the goals difference as described in section 3.4 with 0.3 used as a draw inflation parameter increased the percentage of draw predictions to a percentage similar to the average percentage of draws (27.52% of all results) across several previous seasons of the Premier League and the accuracy of predictions with this inflated number of draws was just over 50%. However, more analysis is required to find the optimal draw inflation parameter which gives the highest accuracy of predictions.

Another weakness of the model is the assumption of independence between the number of goals scored by the home and away team. Maher suggested using a bi-variate model as an improvement to the independent Poisson model to incorporate a dependence in the number of goals scored by the home and away team. However, departure from independence is most significant in the case of low scoring games which Dixon and Coles identified in section 3 of their paper. They stated that the assumption for independence is reasonable except for low

scoring games i.e. 0-0, 1-0, 0-1, 1-1, where teams could be playing for a draw. Dixon and Coles incorporated the lack of independence for low scoring games in their model by introducing another parameter τ which extended Maher's independent Poisson model further. The model I have created so far in R, is an extension to Maher's independent Poisson model, with the inclusion of a home effect parameter (γ) representing the advantage the team playing at their own stadium has. Extending this model further with the addition of a dependence parameter τ as suggested by Dixon and Coles could improve the accuracy predictions, since the model would now take into account some dependency between the number of goals scored by the home and away team, especially for low scoring games.

To determine whether or not this dependence parameter τ would be a useful extension to my current model, I compared the accuracy of predictions for low scoring games to all the games I have predicted. Out of the 60 games I am predicting the result of, 21/60 of those games were low scoring i.e. both teams scored less than 2 goals. Using 0.3 as a draw inflation parameter, the accuracy of predictions was 11/21 $\approx$ 52.4%. In comparison the accuracy of predictions for all games was 31/60 $\approx$ 51.2%. The model had a similar accuracy of predictions for low scoring games (21/60) compared to all the games. Dixon and Coles suggestion to add a dependence parameter for low scoring games is not required since the current model is not less accurate at predicting results for low scoring games.

4.2 Recommendations

If I had 2-3 months more to work on this project, I would fit a model for each of the previous 5 seasons of the Premier League (2014/15 - 2018/19) and predict results for games played during those seasons. Altogether, each team plays 38 games in a season, so I would fit a model for each season using data from the first 30 games each team plays and predict the results for the remaining 8 games each team plays. For the predictions I made for games played for this current season (2019/20), I used data from 19 games to predict results for the next 6 games each team played in the Premier League. Increasing the number of games I am using to fit the model from 19 to 30 could improve the fit of the model since there is more data available to fit the model. By comparing the accuracy of predictions for this season to the previous 5 seasons, I could determine how the number of games used to fit the Poisson regression model impacts the accuracy of predictions the model makes.

The likelihood ratio test I conducted in section 3.5 found that the parameter "home effect" did not improve the fit of the model for the 2019/20 season. However, only 190/380 games were used to fit this model and I would need to conduct this test again using a larger sample of

data (at least 300 games) and do this across multiple previous seasons of the Premier League before concluding. When running the likelihood ratio test again, if I find that the home effect parameter does improve the model fit, I could predict games using a model that does include the home effect parameter (full model) and one without this parameter (reduced model) to determine whether the home effect improves the accuracy of predictions.

So far, I have only analysed data from the Premier League. The home effect could be investigated further by looking at data from the Champions League where away teams usually must travel to another country in Europe to play and there is a greater advantage in being the home team compared to the Premier League. A data set could be created which contains only games Premier League teams were involved in from the Champions League. In each season only 4 Premier League teams compete in the Champions League and there are two stages in the tournament, the group stage and knockout stage. After the group stage, half of the competing teams are eliminated, and the remaining half proceed to the knockout stage [11]. During the group stage, each team plays 6 games, 3 at home and 3 away from home and all the Premier League teams are in different groups. In total, all the Premier League teams competing will play a minimum of 24 games in this tournament combined.

Creating a data set containing all of the group games played by Premier League teams from the previous 5 seasons of the Champions League will give a total of 120 games to fit a model. Using the glm function, I could estimate the home effect parameter for the Champions League which is expected to be greater than the home effect parameter for the Premier League. Using the model which has been fitted, I could predict the results of games Premier League teams have played in this year's Champion's League group stage and assess the accuracy of predictions. This would help me determine whether my model is better for predicting games in the Premier League or Champions League since the "home effect" is expected be a more significant predictor in the Champions League.

Bibliography

[1] Wikepedia. Premier league. https://en.wikipedia.org/wiki/Premier_League. Accessed: 12-11-2019.

[2] Football data. Premier league data sets. http://football-data.co.uk/englandm.php. Accessed: 10-10-2019.

[3] Martin Eastwood. Predicting football using r. http://www.pena.lt/y/2014/11/02/predicting-football-using-r/. Accessed: 18-01-2020.

[4] Michael J Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

[5] Mark J Dixon and Stuart G Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280, 1997.

[6] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford science publications. OUP Oxford, 2013.

[7] Jonathan Bartlett. Deviance goodness of fit test for poisson regression. https://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/. Accessed: 04-02-2020.

[8] Wikepedia. Skellam distribution. https://en.wikipedia.org/wiki/Skellam_distribution. Accessed: 29-01-2020.

[9] Wikepedia. Bessel function. https://en.wikipedia.org/wiki/Bessel_function. Accessed: 29-01-2020.

[10] Smarkets. History of results. https://help.smarkets.com/hc/en-gb/articles/115000647291-Why-you-should-consider-home-advantage-for-football-trading. Accessed: 22-02-2020.

[11] Wikipedia. Format of champions league. https://www.agonasport.com/intlsoccer-championsleague-format. Accessed: 11-03-2020.

Appendix A

Appendix

Figure A.1: Screen capture showing head of data set

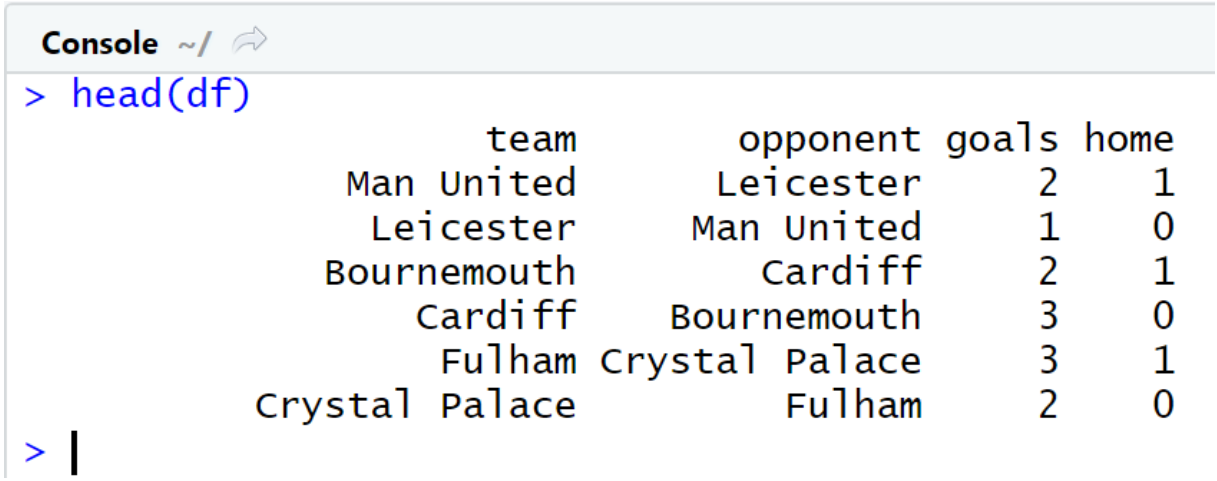

Figure A.2: Screen capture showing head of modified data set

```r

df <- apply(data,1, function(row){
  data.frame(team=c(row['HomeTeam'], row['AwayTeam']),
             opponent=c(row['AwayTeam'], row['HomeTeam']),
             goals=c(row['FTHG'], row['FTAG']),
             home=c(1, 0))
})
df <- do.call(rbind, df)
df[, c(3)] <- sapply(df[, c(3)], as.numeric)
head(df)
```

Figure A.3: Screen capture showing code to modify data set

```
Console ~/
> x <- c(88,116,95,48,22,11)
> p <- c(0.208,0.327,0.256,0.134,0.053,0.022)
> chisq.test(x,p=p)

        Chi-squared test for given probabilities

data:  x
X-squared = 2.7911, df = 5, p-value = 0.7321

>
```

Figure A.4: Screen capture showing output of Chi-squared test

```r
# calculating Probabilities
p <- dpois(0:9,0.6429387) %o% dpois(0:9,3.131493)
rownames(p) <-0:9
colnames(p) <-0:9
print(p)

draw <- sum(diag(p))
away <- sum(p[upper.tri(p)])
home <- sum(p[lower.tri(p)])
cat('home: ', home * 100, '%')

cat('draw: ', draw * 100, '%')

cat('away: ', away * 100, '%')
```

Figure A.5: Code to calculate probabilities of game outcomes

```r
predictHome <- predict(model1, data.frame(Home=1, Team= "West Ham",
                                          Opponent="Liverpool"), type="response")
predictAway <- predict(model1, data.frame(Home=0, Team= "Liverpool",
                                          Opponent="West Ham"), type="response")

# Plot to model probabilities
plotrange <- 0:9
hp <- dpois(plotrange, predictHome)
ap <- dpois(plotrange, predictAway)
plot(plotrange, hp, type="b", ylim=range(hp, ap), main="West Ham Vs Liverpool",
     xlab="Number of goals", ylab="Probability")
points(plotrange, ap, type="b", pch=24)
legend(x=4, y=0.5, legend=c("West Ham", "Liverpool"), pch=c(21, 24))

```

Figure A.6: Code to generate probabilities plot

```r
#skellam distribution

sum(dskellam(-100:-1, predictHome, predictAway)) #Away

sum(dskellam(1:100, predictHome, predictAway)) #Home

sum(dskellam(0, predictHome, predictAway)) #Draw

#Goal difference plot
goalDiffRange <- -6:6
plot(goalDiffRange, dskellam(goalDiffRange, predictHome, predictAway),
     type="b", main="West Ham Vs Liverpool", ylab="Probability",
     xlab="Goal difference")
```

Figure A.7: Code to generate Skellam distribution plot

	A	B	C	D	E	F	G	H	I	J	K
1	HomeTeam	AwayTeam	Actual HG	Actual AG	Actual R	Predicted HG	Predicted AG	Actual GD	Predicted GD	Predicted R	Accuracy
2	Newcastle	Everton	1	2	A	1.1	1.0	-1	0.1	D	Not correct
3	Southampton	Crystal Palace	1	1	D	1.0	1.1	0	-0.1	D	Correct
4	Watford	Aston Villa	3	0	H	0.8	1.5	3	-0.6	A	Not correct
5	Norwich	Tottenham	2	2	D	1.1	2.5	0	-1.4	A	Not correct
6	West Ham	Leicester	1	2	A	0.8	2.7	-1	-1.9	A	Correct
7	Burnley	Man United	0	2	A	1.1	1.7	-2	-0.6	A	Correct
8	Arsenal	Chelsea	1	2	A	1.4	1.7	-1	-0.3	A	Correct
9	Liverpool	Wolves	1	0	H	2.4	0.8	1	1.6	H	Correct
10	Man City	Sheffield United	2	0	H	1.9	1.0	2	0.9	H	Correct
11	Brighton	Chelsea	1	1	D	1.3	1.6	0	-0.3	D	Correct
12	Burnley	Aston Villa	1	2	A	1.5	1.4	-1	0.1	D	Not correct
13	Newcastle	Leicester	0	3	A	0.7	2.1	-3	-1.4	A	Correct
14	Southampton	Tottenham	1	0	H	1.3	2.3	1	-1.0	A	Not correct
15	Watford	Wolves	2	1	H	0.6	1.7	1	-1.1	A	Not correct
16	Man City	Everton	2	1	H	3.0	0.9	1	2.1	H	Correct
17	Norwich	Crystal Palace	1	1	D	0.8	1.2	0	-0.4	A	Not correct
18	West Ham	Bournemouth	4	0	H	1.1	1.2	4	-0.1	D	Not correct
19	Arsenal	Man United	2	0	H	1.2	1.6	2	-0.4	A	Not correct
20	Liverpool	Sheffield United	2	0	H	1.8	0.6	2	1.2	H	Correct
21	Sheffield United	West Ham	1	0	H	1.6	0.7	1	0.9	H	Correct

190 games | 60 games | Predictions

Figure A.8: Screenshot of Excel data set showing the model predictions