Modelling Football Scores in the Dutch Eredivisie

Lotte Bransen 3839869
under supervision of R. Fernandez

July 3rd 2014

Abstract

In this article I will develop a model to predict the scores of individual football matches in the Dutch Eredivisie. By using data from over 1500 recent football matches from the Dutch competition I will determine the attacking and defensive strengths of each of the teams in the competition. I will also implement the fact that the home playing team has some kind of home advantage and therefore the attacking and defensive strengths for all teams will be different when playing at home then when playing away. Finally, I will use my model to predict the outcomes of the matches of previous season to see whether my model is a good fit of reality.

1 Introduction

Since the start of the football competitions in Europe, people have been trying to build models to predict the scores of football matches. In 1982 M.J. Maher published an article in which he used the fact that the number of goals scored by a team can be seen as a Poisson variable. In my model for the Dutch competition I will also assume that the number of goals scored by a team is Poisson distributed.
But why can we assume that the number of goals scored by a team is Poisson distributed? The number of goals scored in a match is given by an integer, as you can not score half a goal. Because of that fact we know that the variable that describes the number of goals scored in a match should be discretely distributed. Next to that, in football it is all about possession: if you do not have the ball you cannot score a goal as Johan Cruijff once said. Each time a team possesses the ball it has the opportunity to attack and this attack

could result in a goal with probability p, where p is small.
Now, if p is constant and the attacks are independent which we can assume is the case in football, then the number of goals scored will be binomial distributed. In this case we are dealing with a Bernouilli trial with a probability of p that an attack is resulting in a success, in this situation thus a goal. The probability mass density function of a Binomial distributed variable X is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

In the case of the number of goals in a match we may take a look at the Binomial distribution with n trials (in this context: attacks) and a probability of $\frac{\lambda}{n}$ that such an attack results in a goal, where λ is the average number of goals scored in a match. In a match you can start an attack in every splitsecond, therefore we can see n as very large and this makes $\frac{\lambda}{n}$ very small, as desired. When we now fill in $p = \frac{\lambda}{n}$ in the probability density function of the Binomial distribution and let n go to infinity we get:

$$P(X_n = k) = \binom{n}{k} \cdot (\frac{\lambda}{n})^k \cdot (1 - \frac{\lambda}{n})^{n-k}$$

$$= \frac{n!}{(n-k)!k!} \cdot \frac{\lambda^k}{n^k} \cdot (1 - \frac{\lambda}{n})^n (1 - \frac{\lambda}{n})^{-k}$$

$$= \frac{n!}{(n-k)!n^k} \cdot \frac{\lambda^k}{k!} \cdot (1 - \frac{\lambda}{n})^n (1 - \frac{\lambda}{n})^{-k}$$

As $n \to \infty$ we get $(1 - \frac{\lambda}{n})^{-k} \to 1$ and

$$\frac{n!}{(n-k)!n^k} = \frac{1}{n^k} \frac{n!}{(n-k)!} = \frac{1}{n^k}(n \cdot (n-1) \cdots (n-k+1)$$

$$= 1 \cdot (1 - \frac{1}{n}) \cdots (1 - \frac{k-1}{n}) \to 1$$

There is one term left, which is the term $(1 - \frac{\lambda}{n})^n$.
The limit $\lim_{x\to\infty}(1 + \frac{1}{x})^x = e^1$ is well-known. By using substitution we get the desired limit: $\lim_{x\to\infty}(1 - \frac{\lambda}{x})^x = e^{-\lambda}$.
So, eventually, we get:

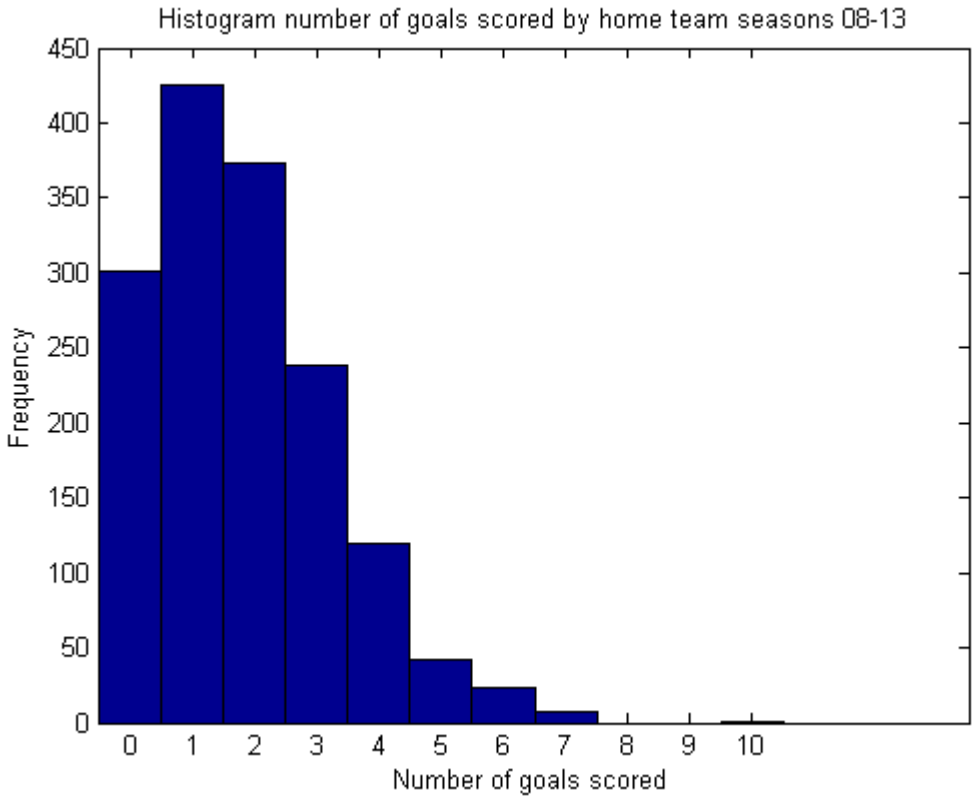
$$P(X_n = k) \to \frac{\lambda^k}{k!} e^{-\lambda}$$

as $n \to \infty$. And this is exactly the probability mass density function of a Poisson distributed random variable X:

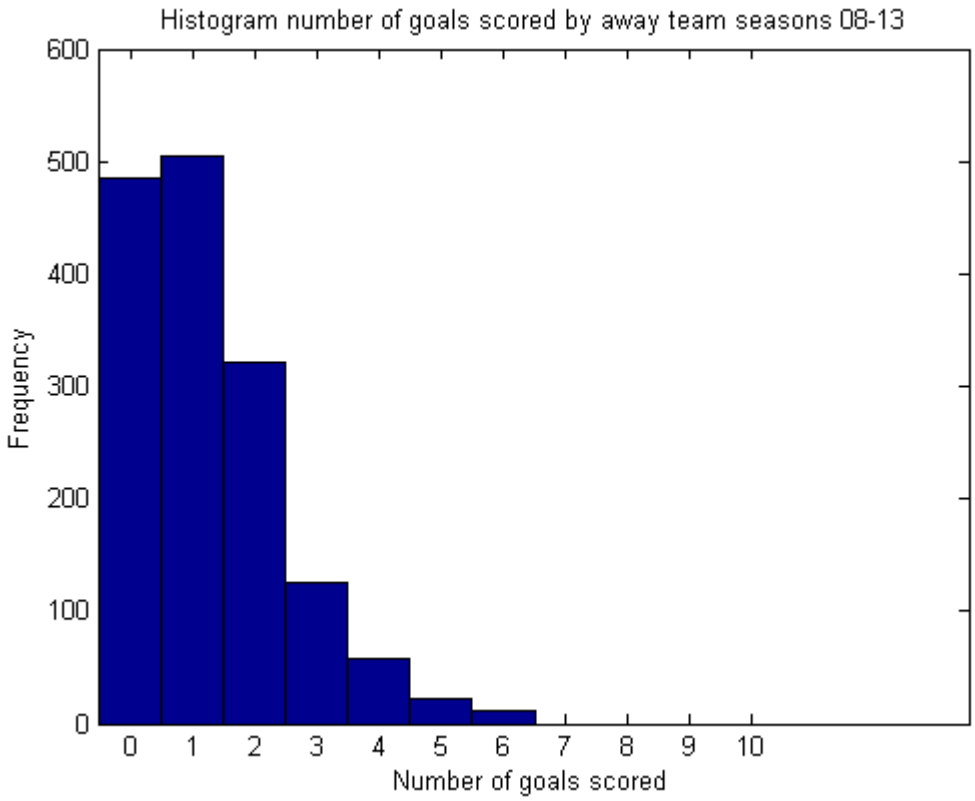
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

with k the number of goals scored and λ the expected number of goals scored in the given time interval.

In these two histograms the number of home goals respectively away goals by the teams over the five seasons show a clear Poisson distribution:



homeall.png



awayall.png

2 The model

As explained in the previous section we can assume that the number of goals scored in a match is Poisson distributed. When we now look at the match in which team i is playing team j, where team i is the home team and thus team j is playing away, we have two stochastic variables to be observed: the number of goals scored by team i and the number of goals scored by team j. The observed score is denoted by (x_{ij}, y_{ij}) and we call the two stochastic variables X_{ij} and Y_{ij}, for the number of goals scored by home team i, respectively, the number of goals scored by away team j. Now we assume that these two stochastic variables are independent and both Poisson distributed, where X_{ij} is Poisson distributed with mean $\alpha_i\beta_j$ and Y_{ij} Poisson distributed with mean $\gamma_i\delta_j$. α_i represents the strength of team i's attack when playing at home, β_j represents the weakness of team j's defence when playing away, γ_i represents the weakness of team i's defence when playing at home and δ_j represents the attacking strength of team j when playing away. As I am building a model for the Dutch Eredivisie, there are 18 teams for which these parameters need to be estimated. There are four parameters to

be estimated for each team, so this makes a total of 72 parameters to be estimated. To estimate these parameters I need data from previous matches played by the teams. To estimate these parameters I need to find a way to determine the attacking and defensive strength's of the teams in the Dutch Eredivisie.

3 The parameters

To estimate the parameters of the model explained earlier, I make use of the maximum likelihood estimation method. Maximum likelihood estimation is a method in which you choose that value of the estimate that makes the observed outcome most probable, you maximize the likelihood. For each match.. To estimate the maximum likelihood estimates I make use of the likelihood function. Since both the stochastic variables are independent in the estimation of the α and β we will only need to use the x, and on the other hand the estimation of the γ and δ will only depend on the y. So, if we look at the home team's scores, the likelihood function is:

$$L(\alpha, \beta) = \prod_i \prod_{j \neq i} \left(\frac{e^{-\alpha_i \beta_j})(\alpha_i \beta_j)^{x_{ij}}}{x_{ij}!}\right)$$

To find the maximum of this function we need to take the derivative with respect to α_i and the derivative with respect to β_j and set them equal to zero. But it is quite difficult to take the derivatives of this function, so to make it ourselves a lot easier we take a look at the log-likelihood function. This function will give the same maximum as the likelihood function as the logarithmic function is continuous and increasing on $\mathbb{R}_{>0}$. This gives:

$$\log(L(\alpha, \beta)) = \prod_i \prod_{j \neq i} \log\left(\frac{e^{-\alpha_i \beta_j} (\alpha_i \beta_j)^{x_{ij}}}{x_{ij}!}\right)$$

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

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

Now, the derivative with respect to α_i is:

$$\frac{\partial \log L}{\partial \alpha_i} = \sum_{j \neq i} (-\beta_j + \frac{x_{ij}}{\alpha_i \beta_j} \cdot \beta_j) = \sum_{j \neq i} (-\beta_j + \frac{x_{ij}}{\alpha_i})$$

Setting this equal to zero gives:

$$-\sum_{j\neq i}\beta_j + \frac{\sum_{j\neq i} x_{ij}}{\hat{\alpha_i}} = 0$$

$$\frac{\sum_{j\neq i} x_{ij}}{\hat{\alpha_i}} = \sum_{j\neq i}\beta_j$$

$$\hat{\alpha_i} = \frac{\sum_{j\neq i} x_{ij}}{\sum_{j\neq i}\beta_j}$$

And in the same way we get $\hat{\beta_j} = \frac{\sum i \neq j x_{ij}}{\sum_{i\neq j}\alpha_i}$. And by using y_{ij} we get similar expressions for $\hat{\gamma_i}$ and $\hat{\delta_j}$.

If we now calculate the second derivative of the loglikelihood function with respect to α_i we can determine whether we are dealing with a maximum or a minimum. This gives:

$$\frac{\partial^2 \log L}{\partial \alpha_i^2} = -\frac{\sum_{j\neq i} x_{ij}}{\alpha_i^2}$$

Because $\sum j \neq i x_{ij}$ and α_i^2 are always positive, the second derivative is always negative and therefore we are dealing with a maximum and that is exactly what we want. If we now impose the two following constraints:

$$\sum_i \alpha_i = \sum_i \beta_i$$

$$\sum_i \gamma_i = \sum_i \delta_i$$

and use the earlier found expressions for $\hat{\alpha}, \hat{\beta}, \hat{\delta}$ and $\hat{\gamma}$ we get the following estimates:

$$\hat{\alpha_i} = \frac{\sum_{j\neq i} x_{ij}}{\sqrt{\sum_i \sum_{j\neq i} x_{ij}}}, \hat{\beta_j} = \frac{\sum_{i\neq j} x_{ij}}{\sqrt{\sum_i \sum_{j\neq i} x_{ij}}}, \hat{\gamma_i} = \frac{\sum_{j\neq i} y_{ij}}{\sqrt{\sum_i \sum_{j\neq i} y_{ij}}}, \hat{\delta_j} = \frac{\sum_{i\neq j} y_{ij}}{\sqrt{\sum_i \sum_{j\neq i} y_{ij}}}$$

In words, the attacking strength of team i when playing at home is the total number of home goals scored by team i in the season divided by the squareroot of the total number of home goals scored by all teams in the Dutch competition that season. On the other hand, the weakness of team i's defence is the total number of away goals scored by the opponents of team i while team i was playing at home, divided by the total number of away goals scored by all teams during the entire season.

4 Data

To estimate these estimates I will use all the results from the seasons 2008-2009 until 2012-2013. This data contain the scores of the 1530 matches played in these seasons. After I have estimated the parameters, I will fit my model on the previous season (season 2013-2014) to see whether it is a good fit of reality and whether I can find a strategy to defeat the bookmakers.
I ran into a big problem when I took a good look at the data of these seasons. It was the problem of promotion and relegation. To estimate the strength's of the teams, the teams need to have played the same amount of matches against each team, as you can see in the estimates that I derived in the previous section. However, for example, Go Ahead Eagles promoted to the Eredivisie in the 2012-2013 season and played in the Eredivisie for the first time in several years. Because of this, it was impossible to estimate the parameters of Go Ahead Eagles as there was no data for me to work with.
A way to solve this problem was to also include the teams of the Jupiler League, the second division in the Netherlands, in my model. In this way I could also depend the strengths of the teams in the Jupiler League and in that way solve the problem. However, the problem with that solution is that when teams promote, or relegate, their strength's change as well. This is due to the fact that when a team for example relegates it's budget decreases with a big amount. Therefore, the club needs to sell its best players and this will have a big influence on the team's strength in the next season. Next to that, also the number of supporters when playing at home will decrease, as supporters always want to see their team in the best competition and playing against attractive opponents, this will have an influence on the home advantage of the club. On the other hand, when a team promotes, the budget goes up with a considerable amount, the club will attract better players and the number of supporters supporting the club will increase. Again the strength of the team makes a big change. Therefore, I chose not to solve the problem by including the teams playing in the Jupiler League.
To solve the problem I created four "promotion teams" and named them $P1$, $P2$, $P3$ and $P4$. I started in the season 2008-2009 and gave the four teams that would regelate each one of the names. Then every season I named the promoted clubs again. The team that was champion in the Jupiler League got the name of the 'best' relegator, in the sense of the relegator with the highest place in the ranking. And then the second best in the Jupiler League was named after the second best relegator in the Eredivisie, etc... In this way the problem of promotion and relegation is being solved, and also each club plays the same amount of matches against any other club. This makes it possible to also estimate the parameters of these four "promotion teams".

5 Application to the Dutch Eredivisie in the 2013-2014 season

The data of the five seasons 2008-2009 until 2012-2013 of all the results of the matches in the Dutch competition of the 18 clubs, including the four "promotion teams", is enough to calculate the estimates of the 72 parameters that have to be found. I chose to use the data of five seasons because I wanted my dataset to be large enough, to reduce the probability of big errors. On the other hand, I chose not to add any more seasons as the strength of the teams nowadays is mostly comparable with previous results, and in average players don't stay at a club for more than four/five years.
The estimates that were derived, are only to be calculated for one season only as the formulas for the estimates include the total number of home/away goals scored in a season by all teams, and by team i. Therefore, I calculated the parameters for all of the five seasons per team. This resulted in the following tables for the moe parameters:

Table 1: Home attack parameters

-	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013
Ajax	1,791093316	2,785242495	1,557848117	2,025900665	1,721325932
AZ Alkmaar	1,660037708	1,436140662	1,516852114	1,571106638	1,2479613
Den Haag	0,917389259	0,913907694	1,475856111	0,909588054	1,075828707
Feyenoord	1,485296896	1,392621248	1,475856111	1,61245155	1,807392228
Groningen	1,397926491	1,08798535	1,516852114	1,199002434	0,860662966
Heerenveen	1,791093316	1,08798535	1,434860108	1,529761727	1,2479613
Heracles	0,961074462	1,566698904	1,59884412	1,240347346	1,635259635
NAC Breda	0,917389259	1,131504764	1,106892083	1,116312611	1,075828707
Nijmegen	1,092130071	0,87038828	1,229880093	1,0749677	0,817629818
P1	1,135815274	0,783349452	0,901912068	0,785553319	0,989762411
P2	0,961074462	1,000946522	1,147888086	0,992277877	1,032795559
P3	0,742648448	1,30558242	0,819920062	1,199002434	0,946729262
P4	1,179500476	0,783349452	1,147888086	0,620173673	0,989762411
PSV Eindhoven	1,791093316	1,74077656	2,008804151	2,356659957	2,495922601
Roda JC	0,961074462	1,610218318	1,393864105	1,405726992	1,2479613
Twente	1,747408113	1,610218318	1,557848117	1,695141373	1,420093894
Utrecht	1,179500476	0,913907694	1,557848117	1,61245155	1,161895004
Vitesse	1,179500476	0,957427108	0,942908071	1,240347346	1,463127042

Table 2: Home defence parameters

-	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013
Ajax	0,483843	0,209657	0,454569	0,68973	0,485643
AZ Alkmaar	0,376322	0,838628	1,010153	0,492665	1,116979
Den Haag	0,967686	1,467599	1,262691	1,280928	1,3598
Feyenoord	1,128967	0,733799	0,808122	0,738997	0,582772
Groningen	0,752645	0,786214	1,010153	1,133129	1,214107
Heerenveen	1,182727	0,995871	1,161675	1,52726	1,456929
Heracles	0,913926	1,310356	0,85863	1,034596	1,554057
NAC Breda	1,182727	0,891042	1,06066	1,083862	1,311236
Nijmegen	1,021446	1,415185	0,959645	1,034596	1,456929
P1	1,021446	1,36277	1,111168	1,872126	1,408365
P2	1,720331	1,520013	2,424366	0,936063	0,922722
P3	1,505289	1,205528	1,767767	1,231662	2,0397
P4	1,451529	1,572427	1,515229	1,428727	1,69975
PSV Eindhoven	0,806405	0,681385	0,757614	0,985329	0,825593
Roda JC	1,344008	1,467599	0,808122	1,231662	0,874157
Twente	0,698884	0,524142	0,656599	1,034596	0,825593
Utrecht	1,021446	0,628971	1,06066	1,625793	0,582772
Vitesse	1,021446	1,467599	1,111168	0,936063	0,874157

There are several ways to determine the parameters for the 2013-2014 season. The most simple way is to take the average of the parameters of the five seasons, and apply those parameters on the 2013-2014 season. Let's call the model in which the parameters are determinde in this way model 1. The parameters for model 1 can be found in the following table:

Club	$\alpha_i, homeattack$	$\gamma_i, homedefence$	$\delta_j, awayattack$	$\beta_j, awaydefence$
Ajax	1,976	0,465	1,990	0,954
AZ Alkmaar	1,486	0,767	1,326	0,958
Den Haag	1,059	1,268	1,057	1,506
Feyenoord	1,555	0,799	1,137	1,084
Groningen	1,212	0,979	1,023	1,288
Heerenveen	1,418	1,265	1,338	1,467
Heracles	1,400	1,134	1,002	1,518
NAC Breda	1,070	1,106	0,910	1,392
Nijmegen	1,017	1,178	1,002	1,274
P1	0,919	1,355	0,784	1,559
P2	1,027	1,505	0,670	1,492
P3	1,003	1,550	0,588	1,505
P4	0,944	1,534	0,641	1,821
PSV Eindhoven	2,079	0,811	1,694	0,899
Roda JC	1,324	1,145	1,105	1,657
Twente	1,606	0,748	1,448	0,787
Utrecht	1,285	0,984	0,941	1,104
Vitesse	1,157	1,082	1,018	1,274

However, as the strength of the teams is likely to be most comparable to the most recent results, a better way to determine the parameters may be to give more recent results more weight. I chose to determine the parameters in another way by using the following formula:

$$\hat{\theta_{total}} = \frac{\hat{\theta}_{0809} + 2\hat{\theta}_{0910} + 3\hat{\theta}_{1011} + 4\hat{\theta}_{1112} + 5\hat{\theta}_{1213}}{15}$$

Where, θ is equal to $\alpha, \beta, \gamma, \delta$. In this way, the matches played in the season 2012-2013 have way more influence on the strength of the teams in season 2013-2014 than the matches played in the season 2008-2009. Let's give this model the name model 2. The parameters for the 2013-2014 season for model 2 can be found in the following table:

Club	α_i, **home attack**	γ_i, **home defence**	δ_j, **away attack**	β_j, **away defenc**
Ajax	1,916	0,497	2,029392	0,902
AZ Alkmaar	1,440	0,843	1,277	1,065
Den Haag	1,079	1,308	1,062	1,496
Feyenoord	1,612	0,726	1,161	1,099
Groningen	1,148	1,064	0,936	1,333
Heerenveen	1,375	1,337	1,367	1,399
Heracles	1,469	1,201	1,048	1,575
NAC Breda	1,090	1,136	0,845	1,367
Nijmegen	0,994	1,210	1,050	1,331
P1	0,900	1,441	0,777	1,418
P2	1,036	1,359	0,681	1,470
P3	1,023	1,623	0,606	1,492
P4	0,908	1,557	0,628	1,808
PSV Eindhoven	2,214	0,834	1,757	0,972
Roda JC	1,348	1,067	1,135	1,788
Twente	1,568	0,799	1,509	0,805
Utrecht	1,329	0,992	1,011	1,132
Vitesse	1,213	1,027	1,141	1,210

To determine the estimate of the standard errors of the maximum likelihood estimators I make use of the Fisher information. If we take a look at the maximum likelihood estimator $\hat{\alpha}$, the Fisher information is denoted as:

$$I(\alpha_i) = E((\frac{\partial \log L}{\partial \alpha_i})^2 \mid \alpha_i).$$

You can see the likelihood function as a random curve. The observed log likelihood function which is found by using the data is slightly different from the "true" likelihood $E(logL(\theta))$. Now, as the sample size increases the observed likelihood converges to the true likelihood. Next to that, the derivative of the log likelihood function with respect to θ, also called the score function, converges to the derivative of the true likelihood function with respect to θ. So, the fisher information indicates the steepness of the observed log likelihood curve around the maximum likelihood estimator. The Fisher information can also be written as:

$$I(\alpha_i) = -E(\frac{\partial^2 \log L}{\partial \alpha_i^2} \mid \alpha_i)$$

, because

$$-E(\frac{\partial^2 \log L}{\partial \alpha_i^2} \mid \alpha_i) = -E(\frac{\partial}{\partial \alpha_i} \frac{\frac{\partial}{\partial \alpha_i} L(\alpha_i)}{L(\alpha_i)}) = -E(\frac{\frac{\partial^2}{\partial \alpha_i^2} L(\alpha_i)}{L(\alpha_i)} - \frac{(\frac{\partial}{\alpha_i} L(\alpha_i))^2}{(L(\alpha_i))^2}),$$

due to the quotient rule, and thus:

$$-E(\frac{\partial^2 \log L}{\partial \alpha_i^2} \mid \alpha_i) = \frac{\partial^2}{\partial \alpha_i^2}(1) + E((\frac{\partial \log L}{\partial \alpha_i})^2 \mid \alpha_i) = I(\alpha_i).$$

Now, we can fill in the formula. As earlier determined, the second derivative of the loglikelihood function with respect to α_i is equal to

$$\frac{\partial^2 \log L}{\partial \alpha_i^2} = -\frac{\sum_{j \neq i} x_{ij}}{\alpha_i^2}.$$

This gives:

$$I(\alpha_i) = E(\frac{\sum_{j \neq i} x_{ij}}{\alpha_i^2}).$$

Now that we have determined the Fisher informations, we can find the standard errors of the estimates by making use of the Cramer-Rao bound. The Cramer-Rao bound states that:

$$var(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

As the maximum likelihood estimator α_i is both unbiased(its expectation equals α_i) and efficient we can use the Cramer-Rao bound to determine the standard errors of the estimates:

$$SE(\hat{\alpha_i}) = \frac{1}{I(\alpha_i)} = \frac{\hat{\alpha_i}^2}{\sum_{j \neq i} x_{ij}}$$

For more information about the Fisher information and the use of the Cramer-Rao bound take a look at chapter 8.5 of the book *Mathematical statistics and* by John A. Rice. I have calculated all standard errors for both models and they are all smaller than 0.04. This is a small value and therefore we can use the estimates and apply them on the 2013-2014 season.

The maximum likelihood estimates from both models now can be used to estimate the means of X_{ij} and Y_{ij}, for a match between team i and team j. Since $X_{ij} \sim Poisson(\alpha_i \beta_j)$ and $Y_{ij} \sim Poisson(\gamma_i \delta_j)$ and since both random variable are assumed to be independent, the probabilities that $X_{ij} = x$ and $Y_{ij} = y$ may be calculated using the probability density function:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

The expected score distributions can now be found by calculating these probabilities for each match, thus for each pair i and j, and then summing these probabilities. These expected score distributions then can be compared with the observed score distributions. In the following table this is done for model 1:

	Home		**Away**	
Number of goals	obs.	exp.	obs.	exp.
0	38	63,66	79	103,66
1	103	92,08	107	101,72
2	78	73,19	80	58,96
3	55	42,73	25	26,43
4	16	20,56	8	10,14
5 or more	16	13,78	7	5,09

In the following table again the expected score distributions are determined, this time for model 2:

	Home		**Away**	
Number of goals	obs.	exp.	obs.	exp.
0	38	62,69	79	100,29
1	103	91,63	107	101,04
2	78	73,40	80	60,15
3	55	43,12	25	27,76
4	16	20,89	8	10,99
5 or more	16	14,27	7	5,09

In those two tables, it can be clearly seen that the estimates used for model 2 are a better fit of reality than those used for model 1. This can also be seen by calculating the value of the test statistic of a chi-squared test. This chi-squared statistic is given by the following formula:

$$\chi^2 = \sum_k \frac{(observed_k - expected_k)^2}{expected_k}$$

For model 1 this results in a test statistic of 16.84 for the home scores and 14.89 for the away scores. Both these chi-squared statistics are quite big and this mainly is a direct result of the fact that in both home and away scores the times that zero goals are scored are overestimated with a huge amount. The test statistics for the distributions in model 2 are somewhat lower than those of model 1, namely 16.04 for the home scores and 12.75 for the away scores. We are dealing with a chi-squared statistic with $6-1=5$ degrees of freedom,

and this gives at a significance level of 1 percent a critical value of 13.39. So, only the expected away score distribution when using the weighting-formula-parameters is a good fit of reality when using this test. However, for all four test statistics the fact that the zero goals matches were overestimated, had a great influence on their value. Therefore, by slightly adjusting the model this problem could easily be solved. Since the parameters used for the second table clearly give a better fit of reality we will continue with these parameters.

For a match between team i and team j the probabilities of a home win, a draw and a away win can be calculated. To calculate these probabilities we need to make use of the probability mass function for the Skellam distribution. The Skellam distribution is the probability distribution of the difference between two independent variables which are both Poisson distributed:

$$Skellam(\lambda_1, \lambda_2) = Poisson(\lambda_1) - Poisson(\lambda_2).$$

The probability mass function of a Skellam distributed variable is given by:

$$f(k; \lambda, \mu) = e^{-(\lambda+\mu)} \cdot (\frac{\lambda}{\mu})^{k/2} \cdot I_k(2\sqrt{\lambda\mu}),$$

where, k is the difference in scores; the number of home goals minus the number of away goals. Next to that, λ is the mean of X_{ij}, μ the mean of Y_{ij} and I_k is the modified Bessel function of the first kind.
For the derivation of this formula I refer to the article "The Frequency Distribution of the Difference Between Two Poisson Variates Belonging to Different Populations" by J.G. Skellam.

The probability of a home win is now the sum of the probability mass functions over $k = 1, 2, 3, ...$, the probability of a draw is the value of the probability mass function for $k = 0$ and the probability of an away win is the sum of the probability mass functions over $k = -1, -2, -3, ...$ In the following table these probabilities for some matches of the season 2013-2014 together with their observed score.

Match	Home win	Draw	Away win	Observed score
Roda JC-Ajax	0.22	0.20	0.58	1-2
Heracles-Den Haag	0.54	0.20	0.26	1-0
PSV-Nijmegen	0.79	0.13	0.08	5-0
Utrecht-Groningen	0.57	0.23	0.21	1-0

Overall, the model can predict the winner of a match very well, whereas it has difficulties with matches where one or both teams score no goals.

6 Conclusion

This simple model is already a good fit of the reality, but this model could be improved in many ways. The model I used is a static model and as you can probably imagine teams are not constant over the year. By using the weighting function I have taken care of the fact that previous results show more about the strengths of teams nowadays than results from earlier matches. However, there may be weighting function that work better and that can also be used during the season.
Next to the fact that the model is static, there are also other factors other than just the number of goals scored that affect the strengths of the teams. Think of the budget of the club and recent transfers that have been made. Also the sacking of the manager, injuries and suspensions of important players, and the strength of the team as a whole may influence the results. The number of matches played besides the competition, such as European matches and Cup matches, may influence the condition of the players which can also influence the strength of the team. Also factors as the referee, the number of supporters or even the weather may influence the result of the match. Of course, it is impossible to include all these factors in the model and as proved previously this simple model already works well.
The model also could be adjusted to the game status, as on average more goals are scored later in the match and may play more attacking when they are losing the match. Finally, as was seen in the tables, the model overestimates the number of matches in which one or both teams score no goals. An adjustment to the likelihood function could be made to solve this problem.
To conclude, the model is working well but a lot of things could be improved as well.