MATH5004M: Bayesian Sports Modelling

Ryan Chan
200850644

May 22, 2018

Table of contents

Introduction

Project aims:

- Build a Bayesian hierarchical model to predict football results in the Premier League
- Implement the model using Hamiltonian Monte Carlo with the Stan programming language and R
- Look at different techniques to assess model performance and compare with Baio & Blangiardo's model (2010)

The Negative Binomial Model

- Here, we use the negative binomial distribution to model the number of goals scored by the home and away team.
- The use of the negative binomial distribution in football models have been largely ignored.
- Generally, an independent Poisson distribution is used to model the number of goals scored by each team.
- We use the parametrisation that Stan uses in terms of the mean μ and size n, which has the probability mass function:

The Negative Binomial Model

- Here, we use the negative binomial distribution to model the number of goals scored by the home and away team.
- The use of the negative binomial distribution in football models have been largely ignored.
- Generally, an independent Poisson distribution is used to model the number of goals scored by each team.
- We use the parametrisation that Stan uses in terms of the mean μ and size n, which has the probability mass function:

$$p(x) = \frac{(x+n-1)!}{(n-1)!x!}\left(\frac{n}{n+\mu}\right)^n\left(\frac{\mu}{\mu+n}\right)^x.$$

The Negative Binomial Model

- Let y_{g1} and y_{g2} denote the number of goals scored by the home and away team in the g-th game of the season, respectively.
- We believe these follow a negative binomial distribution, with mean μ_{gj} and size n_j, where $j = 1$ for the home goals and $j = 2$ for the away goals:

The Negative Binomial Model

- Let y_{g1} and y_{g2} denote the number of goals scored by the home and away team in the g-th game of the season, respectively.
- We believe these follow a negative binomial distribution, with mean μ_{gj} and size n_j, where $j = 1$ for the home goals and $j = 2$ for the away goals:

$$y_{gj} \mid \mu_{gj}, n_j \sim \mathrm{NB}(\mu_{gj}, n_j),$$

where μ_{gj} represents the mean number of goals expected to be scored by the home team ($j = 1$) and the away team ($j = 2$) in the g-th game of the season.

The Negative Binomial Model

- For the mean number of goals, we assume a log-linear effect, where

$$\log \mu_{g1} = home\_att_{h(g)} + away\_def_{a(g)},$$
$$\log \mu_{g2} = away\_att_{a(g)} + home\_def_{h(g)}.$$

The Negative Binomial Model

- For the mean number of goals, we assume a log-linear effect, where

$$\log \mu_{g1} = home\_att_{h(g)} + away\_def_{a(g)},$$
$$\log \mu_{g2} = away\_att_{a(g)} + home\_def_{h(g)}.$$

- For the home and away parameters for the attacking and defensive strengths for each team, $t = 1, ..., T$, where T is the number of teams in the league,

$$home\_att_t \sim \mathrm{N}(\mu_{h\_att}, \sigma^2_{att}),$$
$$away\_att_t \sim \mathrm{N}(\mu_{a\_att}, \sigma^2_{att}),$$
$$home\_def_t \sim \mathrm{N}(\mu_{h\_def}, \sigma^2_{def}),$$
$$away\_def_t \sim \mathrm{N}(\mu_{a\_def}, \sigma^2_{def}).$$

The Negative Binomial Model

- To impose identifiability constraints, we use a sum-to-zero constraint, so

$$\sum_{t=1}^{T} home\_att_t = 0, \sum_{t=1}^{T} away\_att_t = 0,$$
$$\sum_{t=1}^{T} home\_def_t = 0, \sum_{t=1}^{T} away\_def_t = 0.$$

The Negative Binomial Model

- Then the prior distributions for the hyperparameters are as follows:

$$\begin{aligned}
\mu_{h_att} &\sim \mathrm{N}(0.2, 1), \\
\mu_{a_att} &\sim \mathrm{N}(0, 1), \\
\mu_{h_def} &\sim \mathrm{N}(-0.2, 1), \\
\mu_{a_def} &\sim \mathrm{N}(0, 1). \\
\sigma^2_{att} &\sim \mathrm{Gamma}(10, 10), \\
\sigma^2_{def} &\sim \mathrm{Gamma}(10, 10).
\end{aligned}$$

The Negative Binomial Model

- Then the prior distributions for the hyperparameters are as follows:

$$\begin{aligned}
\mu_{h_att} &\sim \mathrm{N}(0.2, 1), \\
\mu_{a_att} &\sim \mathrm{N}(0, 1), \\
\mu_{h_def} &\sim \mathrm{N}(-0.2, 1), \\
\mu_{a_def} &\sim \mathrm{N}(0, 1). \\
\sigma^2_{att} &\sim \mathrm{Gamma}(10, 10), \\
\sigma^2_{def} &\sim \mathrm{Gamma}(10, 10).
\end{aligned}$$

- And the prior distribution for the size n_j for $j = 1, 2$ is given by

$$\begin{aligned}
n_1 &\sim \mathrm{Gamma}(2.5, 0.05), \\
n_2 &\sim \mathrm{Gamma}(2.5, 0.05).
\end{aligned}$$

The Negative Binomial Model

A graphical representation of this model is:

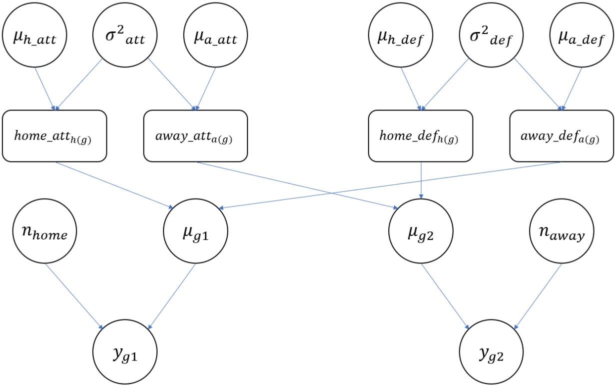

Figure: The DAG representation of the Negative-Binomial Model

The Negative Binomial Model - Results

- We use the data from the 2017/18 Premier League season, to obtain estimates for the attack and defence parameters for each team.
- The data is taken from the the football-data.co.uk website

The Negative Binomial Model - Results

- We use the data from the 2017/18 Premier League season, to obtain estimates for the attack and defence parameters for each team.
- The data is taken from the the football-data.co.uk website
- Higher attack parameter $\implies$ better attacking ability.
- Higher defence parameter $\implies$ worse defending ability.

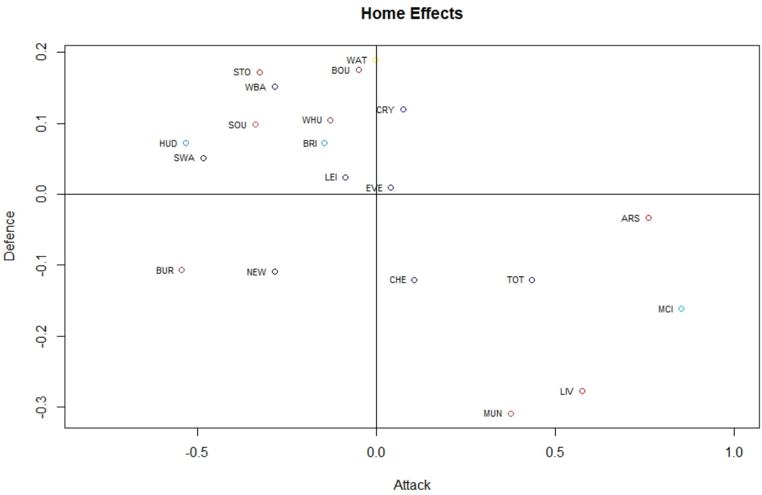

Figure: Plot of the posterior means for the home attack parameter against the home defence parameter for each team

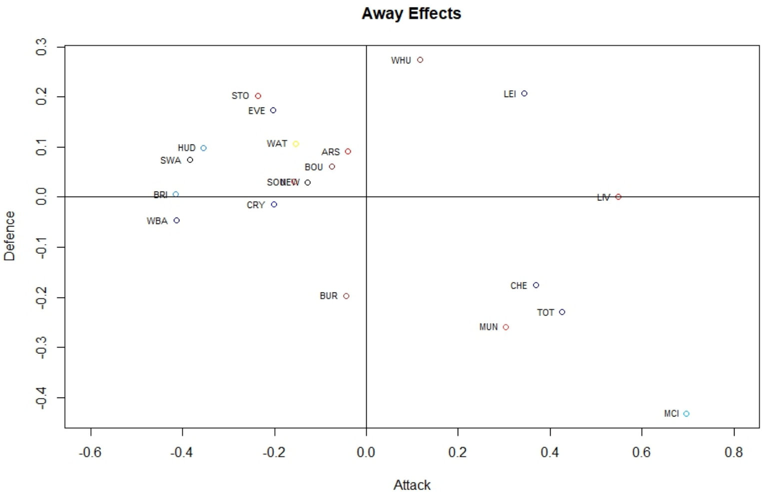

Figure: Plot of the posterior means for the away attack parameter against the away defence parameter for each team

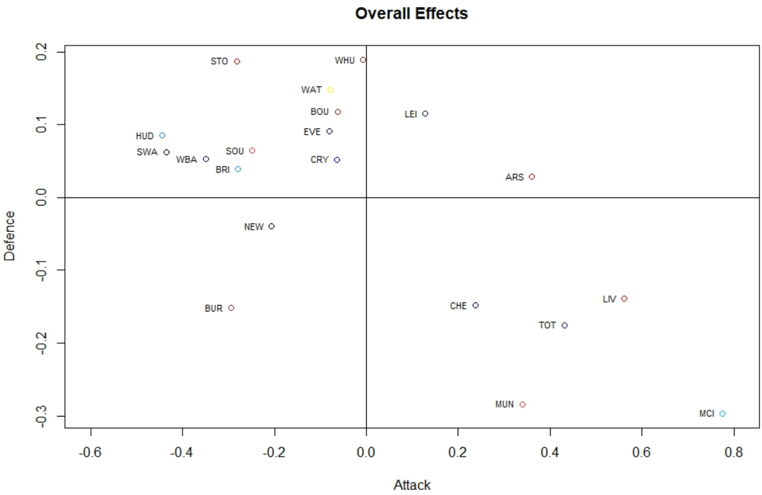

Figure: Plot of the posterior means for the attack parameter against the defence parameter

The Negative Binomial Model - using it for prediction

Using this model for prediction of football matches, we can obtain posterior probabilities for:

- match outcomes (home win / draw / away win),
- final scores.

The Negative Binomial Model - using it for prediction

Using this model for prediction of football matches, we can obtain posterior probabilities for:

- match outcomes (home win / draw / away win),
- final scores.

After using the Stan programming language and R to implement the model, we obtain a sample from our target density.
Once we have a sample from our posterior distribution, we can draw from a predictive distribution of unobserved data or future data.

The Negative Binomial Model - using it for prediction

In our case, to predict a football match between team A (playing at home) vs. team B (playing away):

The Negative Binomial Model - using it for prediction

In our case, to predict a football match between team A (playing at home) vs. team B (playing away):

1. Extract the samples for the attack and defence parameters for each team and for the size n_j, for $j = 1, 2$.

The Negative Binomial Model - using it for prediction

In our case, to predict a football match between team A (playing at home) vs. team B (playing away):

1. Extract the samples for the attack and defence parameters for each team and for the size n_j, for $j = 1, 2$.
2. Use the formula for μ_j, $j = 1, 2$, to get

$$\log \mu_1 = home\_att_A + away\_def_B,$$
$$\log \mu_2 = away\_att_B + home\_def_A.$$

The Negative Binomial Model - using it for prediction

In our case, to predict a football match between team A (playing at home) vs. team B (playing away):

1. Extract the samples for the attack and defence parameters for each team and for the size n_j, for $j = 1, 2$.
2. Use the formula for μ_j, $j = 1, 2$, to get

$$\log \mu_1 = home\_att_A + away\_def_B,$$
$$\log \mu_2 = away\_att_B + home\_def_A.$$

3. Obtain draws from our likelihood, $\pi(y_j^* \mid \mu_j, n_j)$, for $j = 1, 2$, (from a negative binomial distribution).

The Negative Binomial Model - using it for prediction

In our case, to predict a football match between team A (playing at home) vs. team B (playing away):

1. Extract the samples for the attack and defence parameters for each team and for the size n_j, for $j = 1, 2$.
2. Use the formula for μ_j, $j = 1, 2$, to get

$$\log \mu_1 = home\_att_A + away\_def_B,$$
$$\log \mu_2 = away\_att_B + home\_def_A.$$

3. Obtain draws from our likelihood, $\pi(y_j^* \mid \mu_j, n_j)$, for $j = 1, 2$, (from a negative binomial distribution).
4. Now we have a sample from the predictive distribution for number of goals scored by each team, and we can use these for prediction.

The Negative Binomial Model - using it for prediction

- To predict the outcome of a match, we estimate the probabilities as:

$$\Pr(\text{Home Win}) = \frac{\text{Number of times } y_1^* > y_2^*}{\text{Number of samples}},$$
$$\Pr(\text{Draw}) = \frac{\text{Number of times } y_1^* = y_2^*}{\text{Number of samples}},$$
$$\Pr(\text{Away Win}) = \frac{\text{Number of times } y_1^* < y_2^*}{\text{Number of samples}}.$$

The Negative Binomial Model - using it for prediction

- To predict the outcome of a match, we estimate the probabilities as:

$$\Pr(\text{Home Win}) = \frac{\text{Number of times } y_1^* > y_2^*}{\text{Number of samples}},$$
$$\Pr(\text{Draw}) = \frac{\text{Number of times } y_1^* = y_2^*}{\text{Number of samples}},$$
$$\Pr(\text{Away Win}) = \frac{\text{Number of times } y_1^* < y_2^*}{\text{Number of samples}}.$$

- To predict the score of a match, we obtain the MAP estimate for the number of goals scored (find the mode).

The Negative Binomial Model - using it for prediction

- To predict the outcome of a match, we estimate the probabilities as:

$$\Pr(\text{Home Win}) = \frac{\text{Number of times } y_1^* > y_2^*}{\text{Number of samples}},$$
$$\Pr(\text{Draw}) = \frac{\text{Number of times } y_1^* = y_2^*}{\text{Number of samples}},$$
$$\Pr(\text{Away Win}) = \frac{\text{Number of times } y_1^* < y_2^*}{\text{Number of samples}}.$$

- To predict the score of a match, we obtain the MAP estimate for the number of goals scored (find the mode).
- Alternatively, we can estimate the probability of the match ending with team A scoring a goals and team B scoring b goals as:

$$\Pr(\text{Score ending at a-b}) = \Pr(y_1^* = a) \times \Pr(y_2^* = b).$$

The Negative Binomial Model - Tottenham Hotspurs vs. Leicester City

- Extract the samples for the attack and defence parameters for TOT and LEI and n_j for $j = 1, 2$.

The Negative Binomial Model - Tottenham Hotspurs vs. Leicester City

- Extract the samples for the attack and defence parameters for TOT and LEI and n_j for $j = 1, 2$.
- Use the formula:

$$\log \mu_1 = home\_att_{TOT} + away\_def_{LEI},$$
$$\log \mu_2 = away\_att_{LEI} + home\_def_{TOT}.$$

The Negative Binomial Model - Tottenham Hotspurs vs. Leicester City

- Extract the samples for the attack and defence parameters for TOT and LEI and n_j for $j = 1, 2$.
- Use the formula:

$$\log \mu_1 = home\_att_{TOT} + away\_def_{LEI},$$
$$\log \mu_2 = away\_att_{LEI} + home\_def_{TOT}.$$

- Simulate from a $NB(\mu_j, n_j)$ to obtain a posterior predictive sample for goals scored by each team.

The Negative Binomial Model - Tottenham Hotspurs vs. Leicester City

- The model predictions for the outcome for this match was:

$$\begin{aligned} \Pr(\text{Tottenham Win}) &= 0.509, \\ \Pr(\text{Draw}) &= 0.235, \\ \Pr(\text{Leicester Win}) &= 0.256. \end{aligned}$$

The Negative Binomial Model - Tottenham Hotspurs vs. Leicester City

- The model predictions for the outcome for this match was:

$$\begin{aligned} \Pr(\text{Tottenham Win}) &= 0.509, \\ \Pr(\text{Draw}) &= 0.235, \\ \Pr(\text{Leicester Win}) &= 0.256. \end{aligned}$$

- The MAP estimate for the number of goals scored predicted the score of the match would be 1-1.

The Negative Binomial Model - Tottenham Hotspurs vs. Leicester City

The probability estimates for the final score were:

		Tottenham						
		0	1	2	3	4	5	6+
Leicester	0	0.048	0.081	0.076	0.049	0.024	0.011	0.006
	1	0.055	0.093	0.088	0.056	0.028	0.012	0.007
	2	0.036	0.060	0.056	0.036	0.018	0.008	0.005
	3	0.016	0.027	0.025	0.016	0.008	0.004	0.002
	4	0.006	0.009	0.009	0.006	0.003	0.001	0.001
	5	0.002	0.003	0.003	0.002	0.001	0.000	0.000
	6+	0.000	0.001	0.001	0.000	0.000	0.000	0.000

Table: Score probabilities for Tottenham vs. Leicester

Model Assessment - methods

The scoring rules that were used to assess the model's performance for the prediction of football scores were:

- Cross-Validation
- The Brier score
- The rank probability score

Additionally, we assessed the model's performance by attempting to predict a league table using the model and using it as a basis of a betting model.

Model Assessment - results and comparison

We calculate the cross-validation score, Brier score, rank probability score and the profit/loss from betting £10 on the most probable outcome for the two models for the 2017/18 Premier League season.

Model	Cross-Validation	Brier score	Average RPS	Profit/Loss
BB	57.8%	0.532	0.173	£449.9
NB	59.7%	0.540	0.177	£873.3

Table: Results and comparison of the Negative Binomial model to Baio & Blangiardo's model (2010)

Model Assessment - betting results

Profit/Loss (PL)	Frequency
-10.00 (lost bet)	156
$0 \leq PL < 10$	132
$10 \leq PL < 20$	63
$20 \leq PL < 30$	13
$30 \leq PL < 40$	2
$40 \leq PL < 50$	3
$PL \geq 50$	1

(a) Baio & Blangiardo's model

Profit/Loss	Frequency
-10.00 (lost bet)	149
$0 \leq PL < 10$	133
$10 \leq PL < 20$	49
$20 \leq PL < 30$	27
$30 \leq PL < 40$	7
$40 \leq PL < 50$	3
$PL \geq 50$	2

(b) Negative Binomial model

Table: Frequency of each profit/loss for each model in £s

Discussion - strengths

Discussion - strengths

- By simply just using previous goals data, we are able to achieve a good model for prediction of football matches.

Discussion - strengths

- By simply just using previous goals data, we are able to achieve a good model for prediction of football matches.
- By assessing the model's usefulness as a basis of a decision rule for betting, it was able to turn a profit - has real world applications.

Discussion - strengths

- By simply just using previous goals data, we are able to achieve a good model for prediction of football matches.
- By assessing the model's usefulness as a basis of a decision rule for betting, it was able to turn a profit - has real world applications.
- By splitting up the attack and defence parameters for home and away and not using a constant home-advantage parameter as Baio & Blangiardo (2010), Dixon & Coles (1997), Lee (1997), Maher (1982), we are able to encode more information on each team's performances.

Discussion - weaknesses

Discussion - weaknesses

- Although the model was able to turn a profit, there was still 149 games that the model incorrectly predicted the outcome of the game ($\approx 40.3\%$), so there is still a lot of room for improvement.

Discussion - weaknesses

- Although the model was able to turn a profit, there was still 149 games that the model incorrectly predicted the outcome of the game ($\approx 40.3\%$), so there is still a lot of room for improvement.
- The model only uses goals to obtain estimates for parameters for each team.
  - Goals may not the best indicator for how well a team is performing - teams can be lucky or unlucky.
  - Possibly by incorporating more data, we can obtain more accurate estimates for the attack and defence parameters for each team.

Discussion - weaknesses

- Although the model was able to turn a profit, there was still 149 games that the model incorrectly predicted the outcome of the game ($\approx 40.3\%$), so there is still a lot of room for improvement.
- The model only uses goals to obtain estimates for parameters for each team.
  - Goals may not the best indicator for how well a team is performing - teams can be lucky or unlucky.
  - Possibly by incorporating more data, we can obtain more accurate estimates for the attack and defence parameters for each team.
- Model ignores other possible factors that can affect team performance, for example:
  - injury/resting of star players
  - fatigue of players / number of days rest between games
  - distance travelled for away team
  - effect of managerial changes

Summary

- We built a Bayesian hierarchical model for prediction of football results, which used a negative binomial distribution to model the goals scored by each team.
- By using several techniques for model assessment, there was not much difference between the negative binomial model and Baio & Blangiardo's model.
  - But the model was far superior when using it as a basis for a betting decision rule and gave a much higher profit return.

Thank you for listening