Time-Independent Poisson Regression Case Study

Motivation

Instead of trying to show how much passionate I am about data science and football, I would prefer to share a piece of my own study, which I came up with while studying maximum likelihood estimation method for my econometrics exam. My goal was to create a model that would be able to assign probabilities to Premier League game outcomes based solely on clubs' past performance. To do so, I gathered historical data from the last decade and compared the model with odds offered by bookmakers to measure how accurate the model is.

Methodology

The approach I decided to take is as follows:

Suppose we consider a game between two teams a and b having certain attacking and defending parameters:
$a = \{a_A,\ a_D\},\ b = \{b_A,\ b_D\},$
$a, b \in S$ where S - teams taking part in a given competition

We want to model expected number of goals scored by each team. This will depend both on attacking strength of a team as well as defensive strength of the opposite side.
Let's denote expected number of goals scored by a against b and vice versa as:

$\lambda_{a,b} = e^{a_A - b_D}$
$\lambda_{b,a} = e^{b_A - a_D}$

Here, we will use Poisson distribution to model the probability of game ending with a $g_a : g_b$ result.

$$P(X = g_a : g_b) = \left(e^{-\lambda_{a,b}} \frac{\lambda_{a,b}^{g_a}}{g_a!}\right)\left(e^{-\lambda_{b,a}} \frac{\lambda_{b,a}^{g_b}}{g_b!}\right) \qquad \text{Eq. (1)}$$

Consider all N games within one Premier League season. We can calculate their joint likelihood of all these N games given parameters:

$$\mathcal{L} = \prod_{a,b \in S}^{N} \left(e^{-\lambda_{a,b}} \frac{\lambda_{a,b}^{g_a}}{g_a!}\right)\left(e^{-\lambda_{b,a}} \frac{\lambda_{b,a}^{g_b}}{g_b!}\right)$$

Our goal is to find such parameters for every team in the league that will maximize the above likelihood. In order to simplify the above problem, let's consider log-likelihood:

$$\ln(\mathcal{L}) = \sum_{a,b \in S}^{N} \left(-\lambda_{a,b} + g_a \ln(\lambda_{a,b}) - \ln(g_a!) - \lambda_{b,a} + g_b \ln(\lambda_{b,a}) - \ln(g_b!)\right)$$

Substituting for $\lambda_{a,b}$ and $\lambda_{b,a}$ gives:

$$\ln(\mathcal{L}) = \sum_{a,b \in S}^{N} \left(-e^{a_A - b_D} + g_a(a_A - b_D) - \ln(g_a!) - e^{b_A - a_D} + g_b(b_A - a_D) - \ln(g_b!)\right)$$

Differentiating with respect to offensive strength of team x we get:

$$\frac{\partial \ln(\mathcal{L})}{\partial x_A} = -\sum_{i \neq x}^{S} \left(e^{x_A - i_D}\right) + G_x \qquad \text{where } G_x \text{ - total goals scored by team } x.$$

By the same principle differentiating with respect to its defensive strength gives:

$$\frac{\partial \ln(\mathcal{L})}{\partial x_D} = \sum_{i \neq x}^{S} \left(e^{i_A - x_D}\right) - L_x \qquad \text{where } L_x \text{ - total goals lost by team } x.$$

Optimization requires both equations to be zero hence:

$$\sum_{i \neq x}^{S} \left(e^{x_A - i_D}\right) = G_x \qquad \text{and} \qquad \sum_{i \neq x}^{S} \left(e^{i_A - x_D}\right) = L_x$$

Thus

$$\ln\left(\sum_{i \neq x}^{S} \left(e^{x_A - i_D}\right)\right) = \ln(G_x) \qquad \text{and} \qquad \ln\left(\sum_{i \neq x}^{S} \left(e^{i_A - x_D}\right)\right) = \ln(L_x)$$

Although this is not entirely true, I will assume for *simplicity* of the model that all opponents playing against team x had their defensive and offensive strengths equal to

$i_D = \bar{i_D} = 1 \qquad \text{and} \qquad i_A = \bar{i_A} = 1$

Assuming the team played p games during the season, this gives us the following:

$$\ln(p e^{x_A - 1}) = \ln(G_x) \Rightarrow x_A = 1 + \ln\left(\frac{G_x}{p}\right)$$

$$\ln(p e^{1 - x_D}) = \ln(L_x) \Rightarrow x_D = 1 - \ln\left(\frac{L_x}{p}\right)$$

Notice that in both cases the terms within logs correspond to average goals scored/lost per game. Again, this result appears due to the simplification we made above. Now, this tells us that having data on the last p matches of both teams allows us to calculate the Poisson means.

Let say teams x and y are playing against each other. The expected number of goals scored by team x can be then expressed as:

$$\lambda_{x,y} = \exp(x_A - y_D) = \exp\left(1 + \ln\left(\frac{G_x}{p}\right) - 1 + \ln\left(\frac{L_y}{p}\right)\right) = \exp\left(\ln\left(\frac{G_x}{p}\right) + \ln\left(\frac{L_y}{p}\right)\right) = \left(\frac{G_x L_y}{p^2}\right)$$

$$\lambda_{y,x} = \left(\frac{G_y L_x}{p^2}\right)$$

Implementation

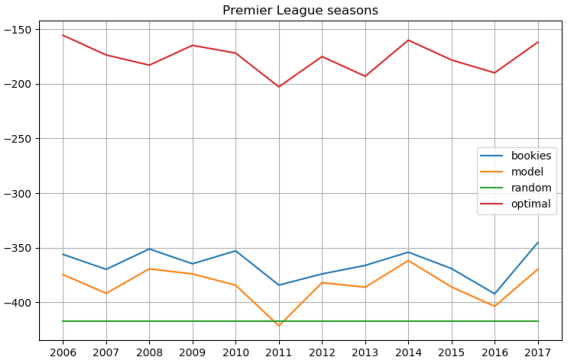
I wrote some Python code that tests the above model and compares it with probabilities implied by bookmakers, optimal model (which I will explain later on) and assumption that all outcomes are equally likely. In order to do so, I downloaded a batch of datasets including all games played within the last decade. To calculate the offensive and defensive strengths at the beginning of each season I collected goals scored and lost by each team in the preceding season. Doing this, I had the data for the last 38 games of each team (apart for the newly promoted teams, for which I assumed 40:60 performance, which is an estimate for an average performance of a newly promoted team). Having calculated offensive and defensive strength I could create a matrix of outcomes probabilities for each game (recall eq. 1):

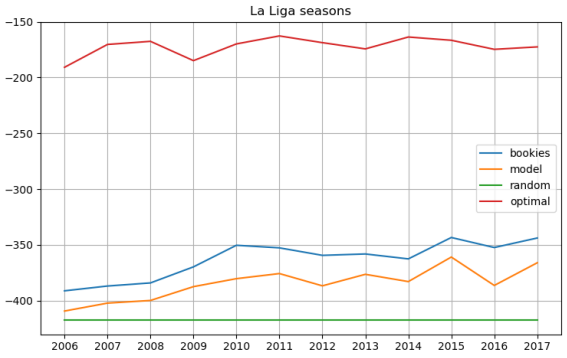
$$\begin{pmatrix} x/y & 0 & 1 & 2 & 3 & 4 & \dots \\ 0 & P(0:0) & P(0:1) & \dots & \dots & \dots & \dots \\ 1 & P(1:0) & \dots & \dots & \dots & \dots & \dots \\ 2 & \dots & \dots & \dots & \dots & \dots & \dots \\ 3 & \dots & \dots & \dots & \dots & \dots & \dots \\ 4 & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

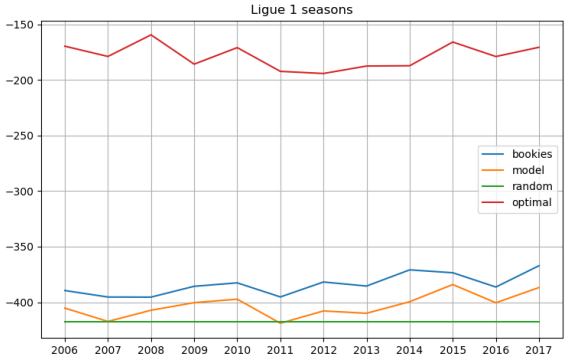
It is clear to see that the sum of diagonal entries gives the probability of a draw, lower/upper triangular matrices less the diagonal give probabilities of home/away side winning . Based on this logic, the Python code calculates probabilities of home/draw/away outcomes for each game. After the game is played, the $\frac{G_x}{p}$ ratio is updated to $\frac{\frac{p-1}{p} G_{x+g_x}}{p}$ where g_x is the number of goals scored in the given game. Similar process goes for goals lost. Notice, that this process assigns probabilities solely based on historical performance of each team. See the Github link and ReadMe file attached for the entire code.

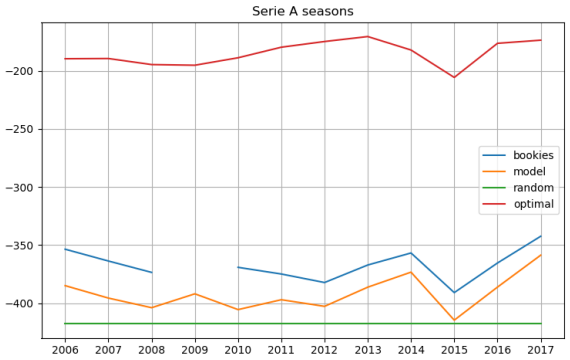
Performance

A good way to measure the performance of the model would be to compare its accuracy with bookmakers probabilities coming from their odds. Also, to put these two predictions into perspective, I will compare these models with the assumptions that all outcomes are equally likely i.e. P(home) = P(draw) = P(away) = 1/3. This could be assumed to be a lower performance bound. On the flip side, a perfect model would predict Poisson means equal to the actual outcome of the game i.e. if Liverpool wins against Leicester 4:3, a perfect model would predict $\lambda_{\text{Liv, Lei}} = 4$ and $\lambda_{\text{Lei,Liv}} = 3$. This can be assumed to be an upper performance bound. All of these 4 models give different probabilities for every outcome in a Premier League season. To measure their performance over the entire season I sum the log-likelihood for all outcomes during the Premier League season.
Having established these metrics I tested its performance over the last twelve years and compared log-likelihoods given by each model. I not only did this for Premier League, but other major leagues too.


Premier League seasons
−150
−200
−250
−300
−350
−400
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
bookies
model
random
optimal


La Liga seasons
−150
−200
−250
−300
−350
−400
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
bookies
model
random
optimal





As we can see the model is statistically valid, as it demonstrates some predictive potential in almost all instances. However, it still falls short of the bookmakers predictions which are still more accurate as their aggregate log-likelihood is always bigger. However, the correlation of movement between my model and bookmakers predictions indicates that foundations of both models can be alike.

Potential areas of improvement

Most importantly, the assumption regarding homogeneous offensive and defensive strengths is invalid. Iterating through the process a couple of times to find a true value of x_A satisfying the equation $\ln\left(\sum_{i \neq x}^{S} \left(e^{x_A - i_D}\right)\right) = \ln(G_x)$ may give more accurate value of x_A as well as x_D.

Also, there are plenty of important factors missing from the model like home advantage for instance. This, alongside the offensive strength would contribute positively towards the expected

number of goals scored and lost.
I.e means would be $\lambda_{a,b} = e^{a_A - b_D + a_H}$ and $\lambda_{b,a} = e^{b_A - a_D - a_H}$ assuming that a are the hosts.
Weather conditions could also be an important factor affecting the expected values via some certain scaling factor. Similar case may be made for time of the year and team motivation to win (small if the result doesn affect anything at the end).

}