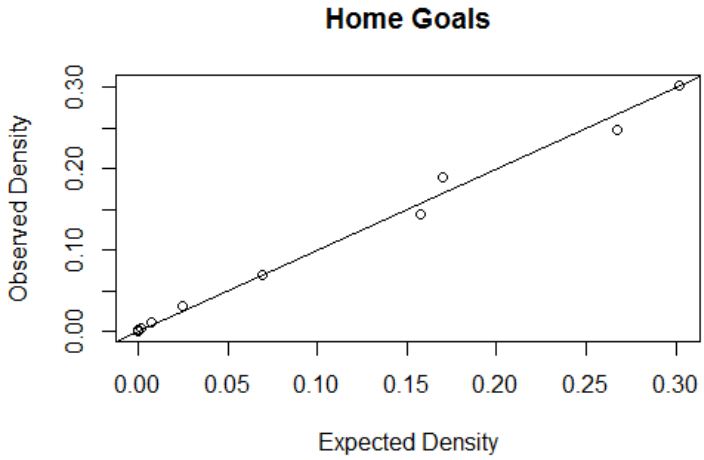
Author: Jaime Martin
Student No: 2914569

Subject: Big Data Analytics

Title: Assignment 3 – Part 2

Do Goals Scored Follow a Poisson Distribution?

On the right we have a plot of the observed density of number of goals scored versus the expected density. This has been done for home goals, away goals, and total goals. Home goals deviated furthest from the line so if we can prove mathematically that this fits a Poisson distribution then we will assume that away goals and total goals do too.



To check this we will use the Chi-Square Goodness of Fit test by comparing the actual frequencies (goals scored) to the theoretical frequencies that would be expected by a Poisson distribution.

H_0 : Goals Scored follows a Poisson distribution
H_1 : Goals Scored does not follow a Poisson distributions

In the table alongside we can see the number of goals scored by the home team and how often that occurred in the actual games played in the Observed column.

In the Expected column we see the number of games a Poisson distribution (with mean = 1.7702) over 190096 games, which is our sample size in this case.

Games with 9 or more goals have been grouped because one of the assumptions of this test is that each level has at least 5 expected outcomes.

Goals	Observed	Expected
0	35999	32723
1	57563	57306
2	47082	50722
3	27291	29929
4	13111	13245
5	5699	4689
6	2158	1383
7	784	349
8	249	77
9	160	15

The summary statistics for our test are as follows:

```
X-squared = 0.018726, df = 9, p-value = 1
```

With such a high p-value* we must accept the null hypothesis, in fact the null hypothesis is almost certain.

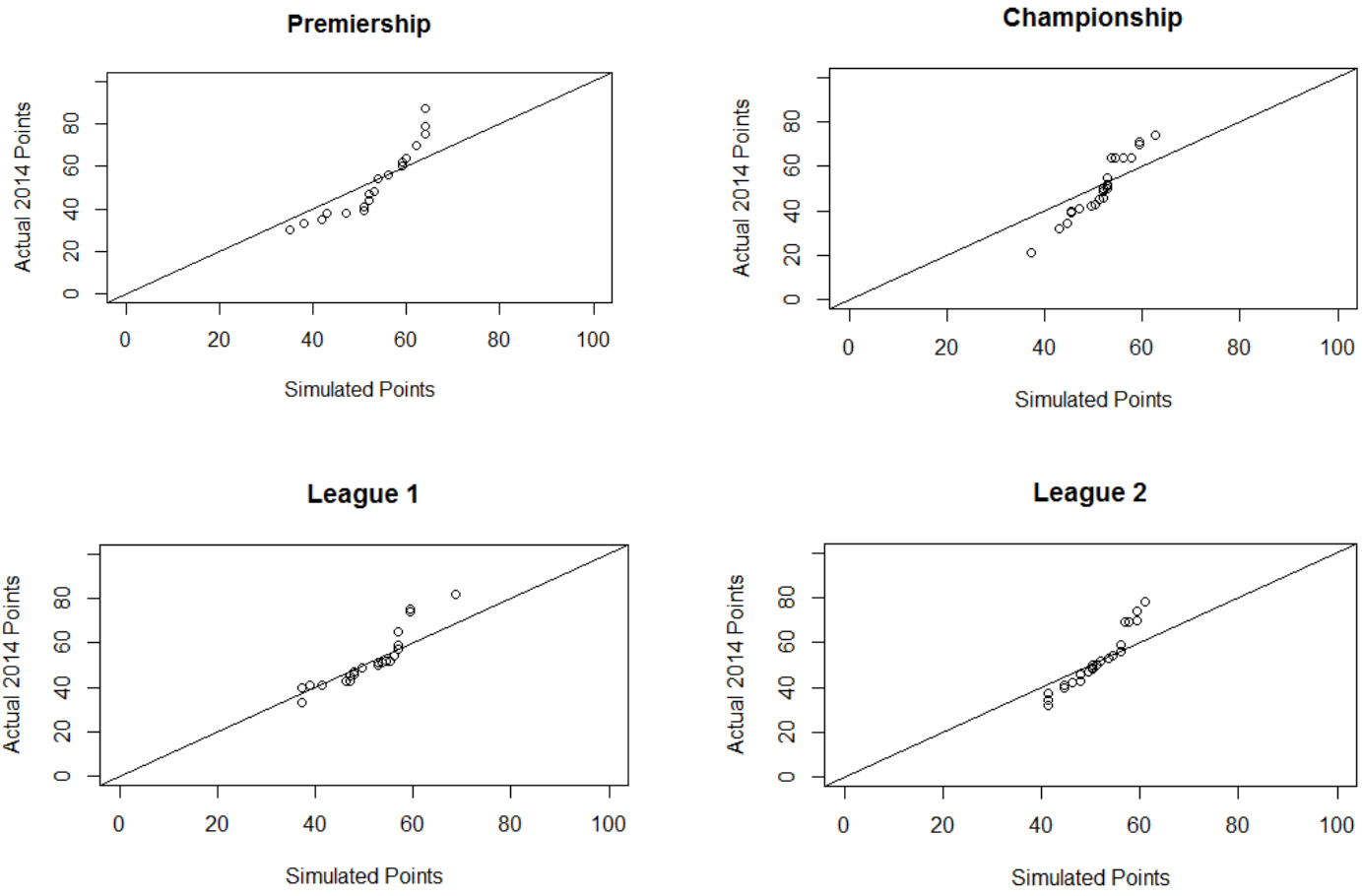
** The p-value is not the most accurate that could be obtained but all operations in R must be run in memory and my computer did not have enough memory to run it.*

Is Football a Game of Skill?

This would appear to be a question with an obvious answer but it is an assumption that is being made and so it should be tested. To do this we will take the four tiers in 2014 and test them separately. We are separating them because perhaps there is skill at higher levels that disappears lower down.

We will use each league's mean home goals scored and mean away goals scored and simulate a league where all teams have identical skill. To simulate a match we sample one number from a Poisson distribution with mean home goals scored and one number from a Poisson distribution with the away goals mean. Each team in the league plays a match home and away against every other team and their points are summed.

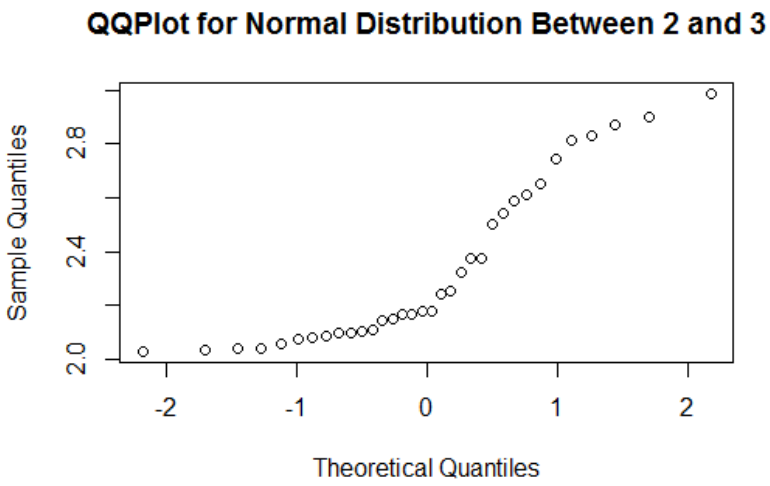
Here are the results for the four divisions:

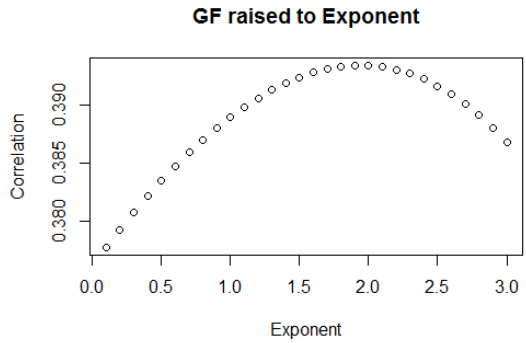


What we see here are indications of positive or right-skew indicating that the extreme values are at the top (less so in the Championship). To better understand this look at the diagram on the next page.

The diagram on the right has a similar pattern to the diagrams on the previous page. This diagram was created by taking a standard normal distribution and selecting the values between 2 and 3. Obviously the further from the mean the less frequent the values.

My theory is that the talent of all teams is normally distributed but by creating leagues and sorting the quality of teams via promotion and relegation you are creating divisions where the talent has positive skew.





Bootstrapping the Chi-Square Test: To see if the observed values fit the expected values we can again Chi-Square Test again – with null hypotheses that the results produced are by chance. If we assume that their is no skill then, via the principle of indifference, any of the expected outcomes from the simulated leagues can be matched with any of the observed outcomes. Thus their is no fair way to match them up from the values we currently have. Also using the mean does not work as R expects to be able to create tabular inputs with at least two levels.

Instead we will sample each vector of values with replacement to create 1,000 matched pairs. A nice problem solved but unfortunately it yielded the following insignificant p-values:

Premiership: 0.64
Championship: 0.8
League 1: 0.28
League 2: 0.56

Even though we are not able to reject the null hypothesis for any league we are far from accepting it. It seems to do that we need to find a model to predict the expected point for individual teams. The problem with the current model is that many teams will finish in the middle in the observed outcomes as most teams are predicted to in a random model. Thus the burden of showing skill falls disproportionally on the top and bottom teams.

Can we predict a team's points from their previous year's performance?

We can try this by joining data from a season to the points achieved in the following season. Here is an example of the data

	home	Season	tier	GP	W	D	L	GF	GA	GD	GFpg	GApg	GDpg	Ppg	GDRatio	TIER	PPG
1	Accrington	2006	4	46	13	11	22	70	81	-11	58	67	-9	41	0.87	4	42
2	Accrington	2007	4	46	16	3	27	49	83	-34	40	69	-28	42	0.58	4	41
3	Accrington	2008	4	46	13	11	22	42	59	-17	35	49	-14	41	0.71	4	50
4	Accrington	2009	4	46	18	7	21	62	74	-12	51	61	-10	50	0.84	4	60
5	Accrington	2010	4	46	18	19	9	73	55	18	60	45	15	60	1.33	4	47
6	Accrington	2011	4	46	14	15	17	54	66	-12	45	55	-10	47	0.82	4	45

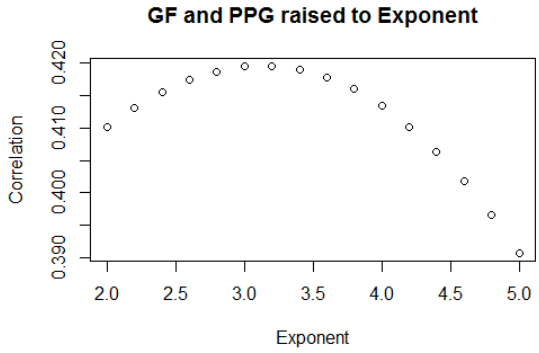
PPG all capitals is our response variable. In 2010 Accrington had a PPG of 47 which is found by looking at their actual points per game (Ppg) from the following year. Now we can look at other variables as explanatory variables to help us make predictions. Simply using the previous seasons points gives a correlation of 0.46 and is the target for others to beat.

Goals For (GFpg) and Agains (GApg):

GFpg has a correlation of 0.39 and GApg has a correlation of -0.37. We can improve the correlation of GF slightly by raising it to a positive power.

We can see that this get to about 0.395 at about 2 which is only a modest improvement.

However if we also raise PPG to an exponent we get another increase to 0.42 when the exponent is 3.



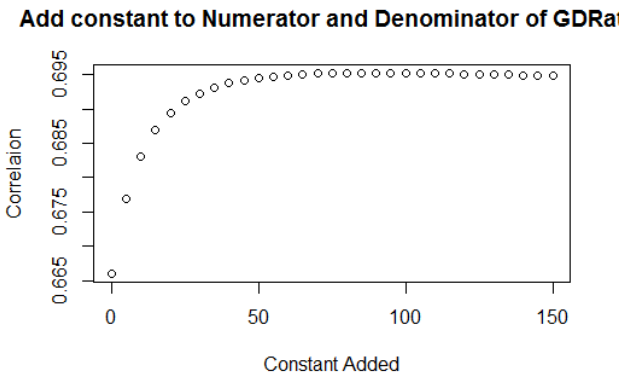
When we break this down by division we find that most of this value, as expected from earlier results is concentrated at the top. The correlation by division is as follows:

Premiership: 0.59
Championship: 0.24
League One: 0.22
League Two: 0.32

League Two is somewhat of anomaly being higher than the two leagues above it but it is also the one that retains the most teams each year (Premiership keeps 17/20, Championship 18/24, L1 17/24 and L2 19/24)

Combining GApg and GFpg:

GDpg: The simplest combination is already provided in the data frame Goal Difference Per 38. I also attempted to square both sides of the correlation with no luck.
GDRatio: Is GFpg/GApg. Here I found a slight increase in correlation by adding a constant term to the numerator and denominator as can be seen on the right in Premiership Correlations.

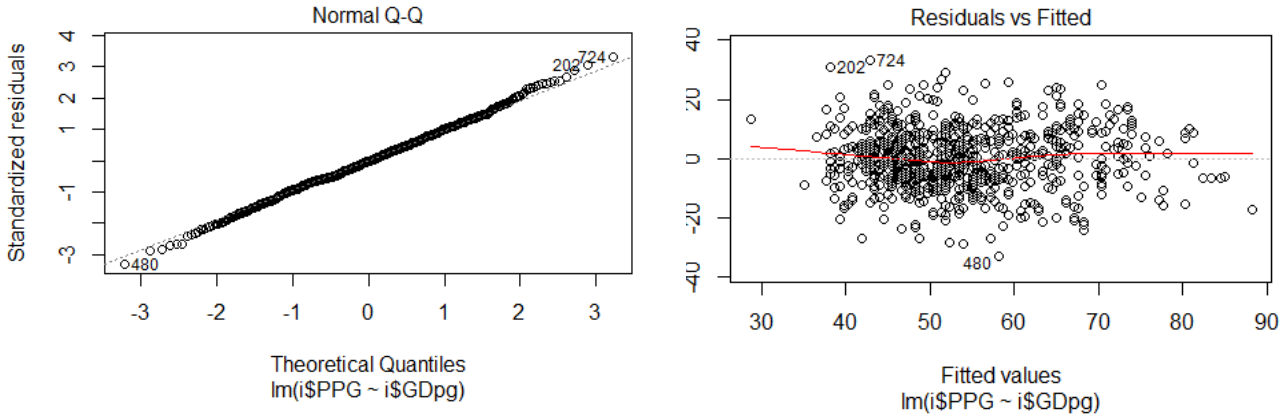


Pythagorean: Borrowed from baseball (https://en.wikipedia.org/wiki/Pythagorean_expectation) this involves (GFpg^2)/(GApg^2 + GFpg^2). This was tried with different exponent as1 which is essentially goals cored as a fraction of all goals.

All of these were tried with z-scores and logs and sometimes even logs were applied after translating to z-scores. None of this produced an improvement over the best predictor which is GDpg. Here are the results for each division:

Premiership: 0.7
Championship: 0.39
League 1: 0.33
League 2: 0.39

Here we can see that GD provides a good linear model as the residuals do not form a pattern and are normally distributed:

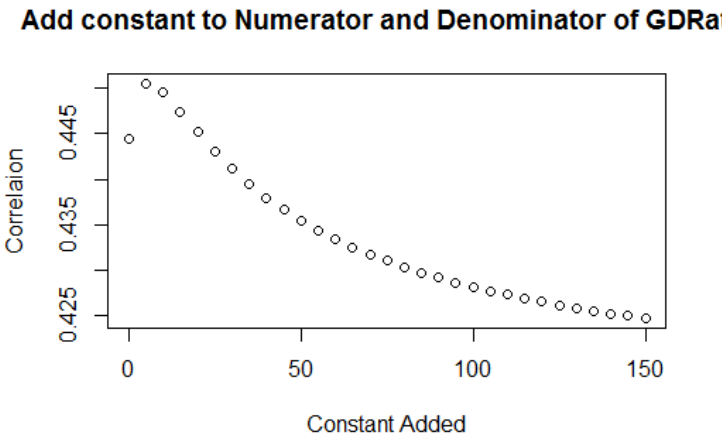


Can we predict good teams within a season?

Here we have divided our dataset into two subsets: games played from August to December and games played from January to June. We will try to use the first dataset to predict PPG from the second dataset.

This should be harder (to find good correlations) than the previous question as immediately we are throwing away half the data. First let's check our champion from the previous question. GDpg is as follows:

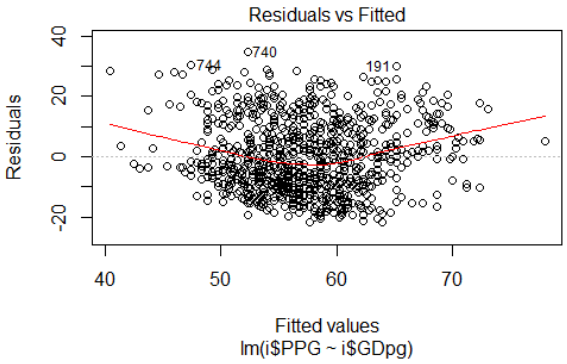
Premiership: 0.42
Championship: 0.09
League 1: 0.16
League 2: 0.19

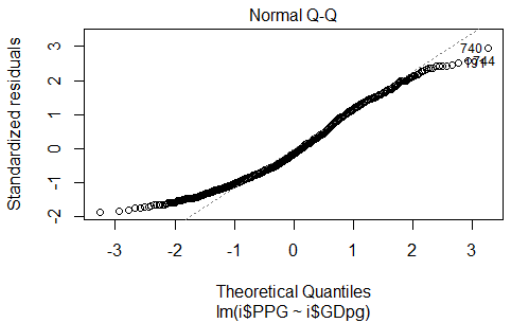


As we can see these numbers are much lower. A hint for why that is the case may be seen in the diagram on the right. It is adding a constant to GDRatio as we did in an earlier question. Here we can get the correlation up to 0.45, which is the highest correlation I could get with this set.

Here the affect of adding a constant drops off much sooner. The purpose of adding a constant is because division changes quite rapidly with low numbers. Perhaps non-linearly is a better term than rapidly. With fewer data points the constant is overwhelming the signal quite soon.

If we re-examine the Residual Plots for the Premiership as we did previously we can see a different story this time:





Interestingly when the model misses big it tends to be miss on the lower side. My theory here is that good teams, who are normally richer, when they perform mediocrely in the first half spend more than they would have when the transfer window opens in January. Thus in the second half they benefit from regression to their true talent level and a higher true talent level.