STAT 654 Spring 2020

Improved NBA scoring model

Submitted by

Jinsu Soh(827009246)
Donghyun Ko(529005755)
Yash Atulbhai Mehta(228002156)
Saurabh Kumar Suresh Jain(527002462)
Vraj Thakkar(329006917)

I . INTRODUCTION

Both ‘team specific average score model’(SA model) and ‘predicting same mean game score for every game / team’(PSM model) did not consider any uncertainty, so they have very low level of prediction. In addition, the last suggested model (OD model with Block-relaxation algorithm) didn’t consider two things : overdispersion and ‘home-and-away’ influence. These invoke the idea of building a model which can predict the score more accurately. There are mainly 3 reasons why the current models are not enough to predict the scores compared to our improved model.

First, Poisson model is assuming that the expectation value of $\hat{P}_{ij}$ (the predicted score when ith team plays against jth team) is the same as its variance. In other words, in order to apply Poisson model, the variance $(P_{ij} - \hat{P}_{ij})^2$ should be the same as $E(\hat{P}_{ij}) = 48e^{oi-dj}$. However, by showing the result of comparing $(P_{ij} - \hat{P}_{ij})^2$ to $E(\hat{P}_{ij})$, 35% of the whole games in the training set are associated with ‘overdispersion’. It means the assumption of Poisson model cannot make sense in every game. Therefore, we should make a new assumption that some games follow the negative binomial distribution in order to solve ‘overdispersion’. We found out the games which didn’t follow the assumption of Poisson distribution, made negative binomial model with new uncertainty parameter θ_{ij}(‘theta’) and predicted their scores with this new model. If this could be proved as valid assumption, the prediction level will be higher than the first two original models(SA and PSM) and it works at least as well as Poisson model. If Negative binomial model works at least as well as Poisson model, we should use Negative binomial model to predict the over-dispersed games (data with overdispersion trend) since the assumption of Poisson model is contradictory in this case(over-dispersed data).

Second, the OD model predicts P_{ij} based on the ‘offensive’ and ‘defensive’ parameters regardless of the home/away position. For example, San-Antonio Spurs was categorized as the strongest team from the current OD model. But if it plays better when it is a home team, and worse when it is an away team, the predictions of score should be separately considered depending on the play-position in order to improve the prediction level.

Lastly, the current OD model used the averaged game length which was around 48min. However, depending on the result of the 4th quarter, some of games had 1 or 2 more quarter(s), each of which has 5 min. In other words, the model without considering the extra quarters may neglect the potential 9-18 additional points, thus the number of outliers can increase.

II. Knowledges

A. Block-relaxation method

Block relaxation divides the parameters into disjoint blocks and cycles through the blocks, updating only those parameters within a single block at each stage of a cycle. Block relaxation is the most successful when these updates are exact. The current model is a simplified version of a sports model in which assumed the scores of ith

team against jth team follow a Poisson process with intensity e^{oi-dj} . This model has two parameters which are ‘offensive strength’ and ‘defensive strength’. But our improved model will have a total 4 parameters by dividing the offensive and defensive parameter into home/away further, and this will yield more prediction accuracy with less MAE.

B. Overdispersion

Poisson model is based on the assumption that the expectation value of the response should be equal to its variance. However, if the variance is larger than the expectation value, this assumption is contradictory. This trend is called “Overdispersion” and we cannot use Poisson regression model in this case. Therefore, we should use a different model, Negative binomial model, instead of Poisson distribution. In Negative binomial model, we assume that there exists a parameter θ associated with overdispersion : Var(Y) = $\mu + \theta\mu^2$, where μ is the mean of Y based on Poisson model. This new parameter, θ, doesn’t affect the mean, but affect the variance. In accordance with this assumption, Negative binomial model with θ has similar mean with comparing to that of Poisson model, but it should make more sense in modeling since it has more flexible assumption about the variance rather than E(Y) = Var(Y).

C. Negative binomial

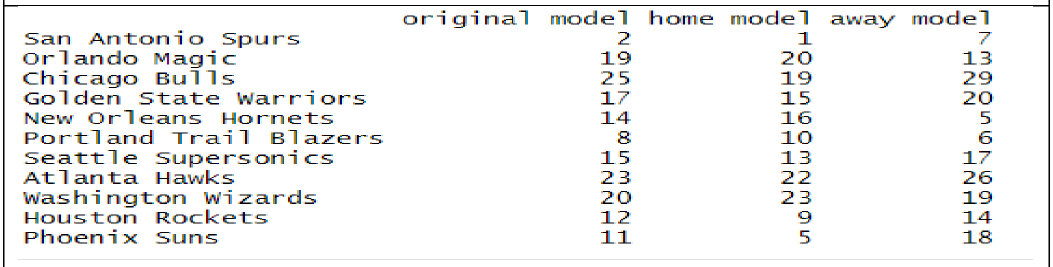
Negative binomial is one of the discrete probability distributions. A random variable ‘k’ follows the Negative binomial distribution based on ‘r’ and ‘p’; f(k;r,p) = $\binom{k+r-1}{k}(p)^r(1-p)^k$, where ‘k’ is the number of failures before r-th success, ‘r’ is the number of success and ‘p’ is the probability of success. Its expectation value is $\frac{r(1-p)}{p}$ and the variance is$\frac{r(1-p)}{p^2}$.

III. Model improvement

A. The original O-D model

As briefly mentioned in the previous sections, the original O-D model did not consider the home/away effect. Rather it just calculated the predicted score per game for each team by o-d parameters without considering home/away position, and used them to compute the MAE. As like our introduction said, each team might have different strengths when they are at home or away position, thus the model should compute the predicted score based on each team’s position(home/away). This is because some teams which are strong at home are not always strong when they are at away, either. The following table(Fig.1) indicates that 11 out of 29 teams showed that the rank of strengths at home or away is extremely different from each other.

Fig. 1. List of teams which home and away strength are extremely different(>4)



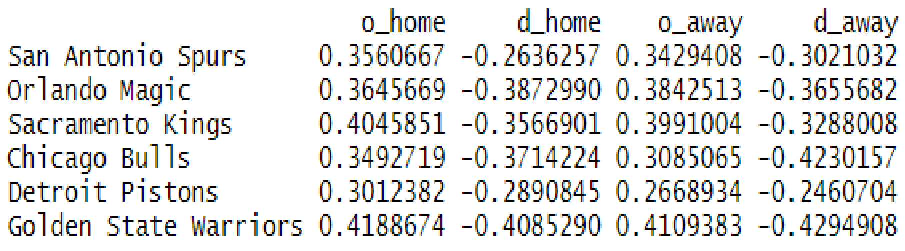
```
                        original model home model away model
San Antonio Spurs                    2          1          7
Orlando Magic                       19         20         13
Chicago Bulls                       25         19         29
Golden State Warriors               17         15         20
New Orleans Hornets                 14         16          5
Portland Trail Blazers               8         10          6
Seattle Supersonics                 15         13         17
Atlanta Hawks                       23         22         26
Washington Wizards                  20         23         19
Houston Rockets                     12          9         14
Phoenix Suns                        11          5         18
```

Other deficiency of this model is the fact that it used the averaged game length which was around 48min. But by considering that some games had 1 or 2 overtime quarter(s), it is not difficult to expect that assuming the game time as 48 min will be able to make a bit more outliers. Lastly, when we used Poisson model to predict the score, it was found that there existed more than 280 over-dispersed data. It means Poisson model is not always valid in every game since its assumption of E(Y) = Var(Y) is sometimes contradictory. Therefore, we need to make another model which can predict these over-dispersed data(games) without contradiction.

B. The improved model

First, we considered the home/away effects by optimizing parameters for home position and away position separately. This resulted in a different rank in each 'original, home, and away model'. For example, San-Antonio Spurs was the 2nd rank in the original model. However, home/away model told it was the most strong at home position, but the 7th rank at away position. This fact is enough to show that home/away effect is significant to be considered. In addition, o-d value of the same team, in most teams, are different depending on the home/away position.

Fig. 2. O-d values from home/away positions



```
                        o_home     d_home    o_away     d_away
San Antonio Spurs     0.3560667 -0.2636257 0.3429408 -0.3021032
Orlando Magic         0.3645669 -0.3872990 0.3842513 -0.3655682
Sacramento Kings      0.4045851 -0.3566901 0.3991004 -0.3288008
Chicago Bulls         0.3492719 -0.3714224 0.3085065 -0.4230157
Detroit Pistons       0.3012382 -0.2890845 0.2668934 -0.2460704
Golden State Warriors 0.4188674 -0.4085290 0.4109383 -0.4294908
```

Then, we considered the different game length for each game; $P_{ijk}=(X1+X2+X3)e^{(o_{ik}-d_{jk})}$, where $X_1 = 48$ if $nOT_{ijk} = 0$ or $X_1 = 0$ otherwise, $X_2 = 53$ if $nOT_{ijk} = 1$, or $X_2 = 0$ otherwise, and $X_3 = 58$ if $nOT_{ijk} = 2$, or $X_3 = 0$ otherwise. In this formula, 'k' refers to team i's home/away position. This sensibly makes the model less affected by time discrepancy than the original one which applied the average game time for all teams. As a result, we named the model made so far as 'improved O-D model or improved Poisson model'.

Lastly, we found out the games with over-dispersion. These games couldn't be predicted by Poisson model since it could not guarantee that the expected value of the predicted score was equal to its variance. Therefore, we used a Negative binomial model for each over-dispersed game(data). First, by using the training data, we found out over-dispersed data and set $(P_{ij} - \widehat{P}_{ij})^2 = \widehat{P}_{ij} + \widehat{P}_{ij}^{\,2}\theta_{ij}$, where θ_{ij} is a parameter associated with uncertainty in a game between team i and j, and $\widehat{P}_{ij}$ is the score predicted by Poisson model. Moreover, θ_{ij} is assumed to generate the over-dispersion in the game between team 'i' and 'j'. In succession, we calculated θ_{ij} through $(P_{ij} - \widehat{P}_{ij})^2 = \widehat{P}_{ij} + \widehat{P}_{ij}^{\,2}\theta_{ij}$ and used them to make a Negative binomial regression model $f(\widehat{P}_{ij}; r_{ij}, p_{ij})$ where its random variable $\widehat{P}_{ij}$ is the predicted score of team 'i' against 'j', $r_{ij} = 1 / \theta_{ij}$ and $p_{ij} = 1/(1+ \theta_{ij}\widehat{P}_{ij})$. Since θ_{ij} doesn't affect the mean, this model will have similar mean as Poisson model, but it makes more flexible assumption that the expected value of the predicted score between i and j is not needed to be equal to its variance. In addition, θ_{ij} is always different depending on 'i' and 'j', so we concluded that Negative binomial model with θ_{ij} has ‘heterogeneity’ with different variance depending on the game. Our new assumption resulted in the new likelihood function $\prod_{(i,j)\in S} \frac{e^{(48e^{oik-djk})}(48e^{oik-djk})^{Pij}}{Pij!}$ x $\prod_{(i,j)\in L}$ f($\widehat{P}_{ij}$; r_{ij}, p_{ij}) where S is the space consists of the games(data) without overdispersion, L is the space consists of the games(data) with overdispersion, S∪L = ‘The total data set’, and f($\widehat{P}_{ij}$, r_{ij}, p_{ij}) is a Negative binomial model to predict P_{ij}.. Because Negative binomial model with θ_{ij} has ‘heterogeneity’ and θ_{ij} is always different in each over-dispersed game, the choice of $\widehat{P}_{ij}$ $\forall\ (i,j) \in L$ which makes the likelihood function maximized is $\widehat{P}_{ij}$ which has the most likely score distributed by f($\widehat{P}_{ij}$, r_{ij}, p_{ij}). From this perspective, we made codes with using dnbinom function in R and predicted the score of the game between team i and j $\forall\ (i,j) \in L$. Other than these over-dispersion cases, the games which have under-dispersion trend or keep the assumption of E(P_{ij}) = Var(P_{ij}) will be predicted by the improved O-D model. In conclusion, our new model is the mixture of Poisson and Negative binomial, which predict the games without overdispersion via Poisson model and the others via Negative binomial model.

IV. Result

First, in terms of MAE, our new model(the mixture of Poisson and Negative binomial) has the lowest MAE. Therefore, we can conclude that our mixture model predict the scores better.

Fig. 3. MAE comparison

Model	PSM	SA	Original O-D	New model
MAE	10.0	9.6	9.15	8.97

In addition, the original O-D model isn’t valid, because 35% of the whole games are over-dispersed data and they make the assumption of E($\hat{P}_{ij}$) = Var($\hat{P}_{ij}$) contradictory. However, our new mixture model not only reduces MAE (= works better than the original model), but also is valid with more flexible assumption for the variance (it means there is no contradictory case). In conclusion, all the above results support the fact that the final model shows a better and more stable prediction.

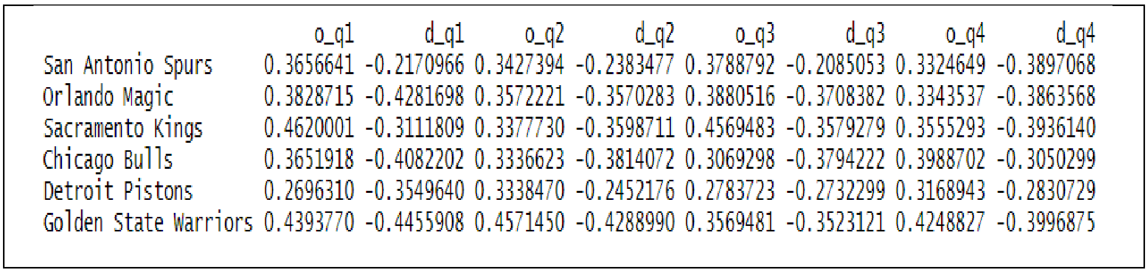
V. Further suggestion

According to the property of Poisson distribution, the parameter for the total score λ should be the same as $\lambda_{Q1}+\lambda_{Q2}+\lambda_{Q3}+\lambda_{Q4}$. Therefore, the predicted scores with parameters generated by the quarter scores were almost the same as those with only one parameter generated by the total scores. However, we got the meaningful things from the quarter parameters, so we want to suggest the idea for the further research.

Our new assumption is “Coaches usually use different player line-up and strategy in each quarter”. It means coaches have their preferred player line-up and strategy in each quarter. For example, the stats in NBA website show that coaches usually allow each player to play an average of 33minutes in a single game. It means star players(=important players) such as Lebron James, James Harden usually play very short time in some quarters. On the other hand, star players usually play the entire time in the fourth quarter since it is the most important moment in the game.

To understand easily, suppose Popovich, who are the coach of San Antonia Spurs, usually use player line-up A,B,C,D in each quarter. In addition, each line-up consists of 5 players and Tim Duncan, the star player in Spurs, usually play in the 1st, 2nd and 4th quarters. Then, Tim Duncan belongs to line-up A,B and D. When we calculate the o/d parameter for each quarter, then each o/d is also affected by each player line-up associated with each quarter; $\lambda_{Q1} = \lambda_A$, $\lambda_{Q2} = \lambda_B$, $\lambda_{Q3} = \lambda_C$, $\lambda_{Q4} = \lambda_D$. For example, o/d value for the first quarter is associated with line-up A.

Fig. 2. Quarter effects



```
                          o_q1       d_q1      o_q2       d_q2      o_q3       d_q3      o_q4       d_q4
San Antonio Spurs     0.3656641 -0.2170966 0.3427394 -0.2383477 0.3788792 -0.2085053 0.3324649 -0.3897068
Orlando Magic         0.3828715 -0.4281698 0.3572221 -0.3570283 0.3880516 -0.3708382 0.3343537 -0.3863568
Sacramento Kings      0.4620001 -0.3111809 0.3377730 -0.3598711 0.4569483 -0.3579279 0.3555293 -0.3936140
Chicago Bulls         0.3651918 -0.4082202 0.3336623 -0.3814072 0.3069298 -0.3794222 0.3988702 -0.3050299
Detroit Pistons       0.2696310 -0.3549640 0.3338470 -0.2452176 0.2783723 -0.2732299 0.3168943 -0.2830729
Golden State Warriors 0.4393770 -0.4455908 0.4571450 -0.4288990 0.3569481 -0.3523121 0.4248827 -0.3996875
```

Quarter parameters are not important to predict the scores in general games. However, it is very important to predict the scores in the games with unusual player line-up such as ‘garbage game’. For example, unlike the general games, star players don’t play in the fourth quarter if the opponent is overwhelmingly ahead and we call this kind of game ‘garbage game’. Then, Popovich also won’t use Tim Duncan in the fourth quarter if the opponent is overwhelmingly ahead. Then, in order to predict the score for this case, λ is not $\lambda_A+\lambda_B+\lambda_C+\lambda_D$, but $\lambda_A+\lambda_B+\lambda_C+\lambda_B$.

This kind of analysis can be applied to other factors. For example, we can calculate o/d parameters for the games with or without Tim Duncan. Like this, we can calculate o/d parameters for the games with or without specific players. Moreover, we also can calculate o/d parameters associated with 2-pt or 3-pt success rate. In order to do this further research, collecting more data related to what are needed in the analysis should precede. With these further research, we can predict the score more precisely.

REFERENCES

[1] Penn-State Statistics Department STAT 654 teaching material

[2] https://esajournals.onlinelibrary.wiley.com/doi/10.1890/10-1831.1