What makes a winning NBA team?

Luca Martucci, Francesco Vinciguerra

January 2024

Abstract

Basketball, as a sport, is a great mixture of athleticism, strategy, and precise quantitative analysis. The game's polyhedric nature has increasingly been scrutinized through statistical methodologies to decode and predict player and team performances. Acknowledging basketball as one of the most quantifiable sports, this paper employs a series of statistical techniques to analyze datasets encompassing teams statistics for the 2021-2 and 2022-3 NBA seasons, with the ideal target of detecting the most relevant parameters influencing the winning percentage of each team.

We will describe the dataset for 2021-2 and our visual and numerical exploration of it, then we will report the linear regressions we fitted, provide some possible explanation for their problematic outcomes, run tests to justify our solution to the problem, and validate our model by evaluating the predictions over data from 2022-3, then making a broader analysis of the evolution of particular statistics over the years, underlying the fundamental limitations of our approach.

1 Dataset

1.1 Description

We merged two datasets [1], [2] containing, for each of the 30 teams of the 2021-2 season, our target variable: the winning percentage `WIN.`, the following *per-game statistics*[4]:

`X3P.`	3 Point Field Goal Percentage
`X2P.`	2 Point Field Goal Percentage
`AST`	Assists per-game
`TOV`	Turnovers per-game
`STL`	Steals per-game
`ORB`	Offensive Rebounds per-game
`DRB`	Difensive Rebounds per-game
`X3PA`	3 Point Field Goal Attempted
`X2PA`	2 Point Field Goal Attempted
`FTA`	Free Throw Attempted
`PACE`	per game number of possessions

and the following *team statistics*:

· `oEFF` (*offensive Efficiency*): points scored averaged by the possessions of the team; intuitively, game actions that start when a player gets the ball and end in a field goal or free throw made, an opponent's defensive rebounder or a turnover;

· `dEFF` (*difensive Efficiency*): points allowed averaged by the possessions of the opponents;

· `eDIFF` := `oEFF - dEFF`

· `SOS`[3]: the *Strength of Schedule* is defined as

$$\text{SoS} := \frac{2\text{OW}\% + \text{OOW}\%}{3}$$

where the *Opponent's winning percentage* OW% is the sum of all the winning percentages of the opponents encountered, while OOW% is the *Overall Opponent Winning Percentage* of the opponents faced by the team, both averaged by the games played by the team in analysis.

1.2 Exploration

In the dataset, no `N/A` or duplicates were present and hence, before fitting some regressions, we explored some interesting relations between our target `WIN.` and relevant parameters, such as `SOS`, which indeed showed a good negative linear correlation as can be seen in the plot in Figure 1, confirmed by a correlation coefficient $\rho = -0.7424491$.

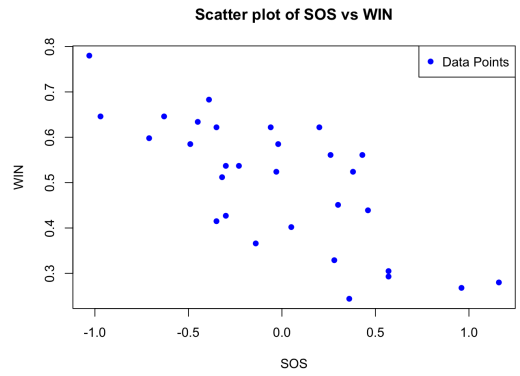

Figure 1.

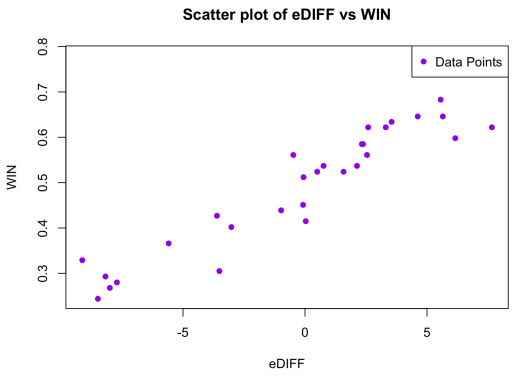

Figure 2.

Also notable were the relationships with `oEFF`, `dEFF` and `eDIFF`, the last of which can be visualized in Figure 2 and is almost trivial with a correlation coefficient $\rho = 0.9451835$: indeed `eDIFF` is just the difference between the previous two parameters and can thus be neglected by our linear model since such collinearity would cause major problems when doing regressions and would contribute nothing to our previous knowledge.

2 Linear regression

We decided to approach the analysis by grouping our statistics in order to avoid collinearity and fit a regression for each group, using both visual and quantitative analyses, such as the Shapiro-Wilk test, to check the assumptions.

The first group is formed by the *team statistics*: `SOS`, `oEFF`, `dEFF` and `PACE`; the assumptions are respected, as can be both seen by Figure 3 and p-value of 0.5571 resulting from a Shapiro - Wilk test.

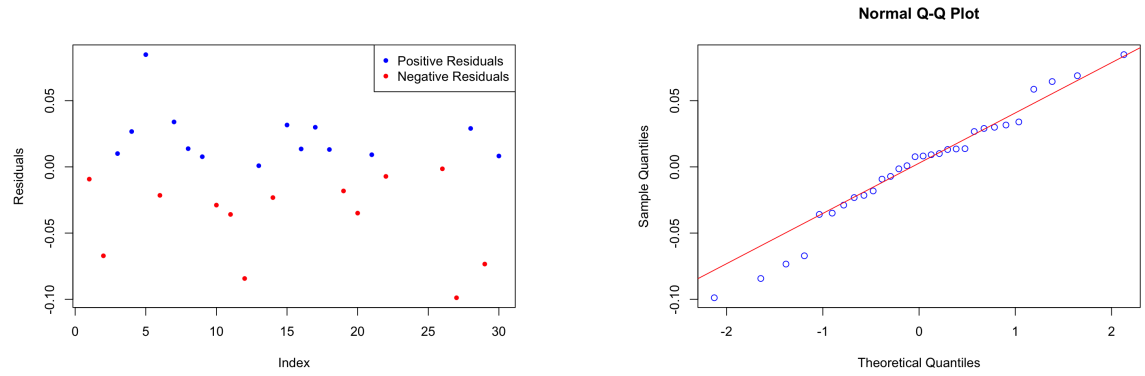

Figure 3.

The linear model for the second group, consisting of the *percentages*: `X3P.` and `X2P.`, gives unsatisfying results, with a $R^2 = 0.5238$ even though the assumptions seem to be respected, as can be seen in Figure 4.

We thought that this surprisingly low linear correlation could be explained by the fact that, in the course of an NBA season, some teams tend to play purposely in a *non-competitive* way, to get the benefits of being a low-ranked team, e.g. higher chances of obtaining lottery picks in the draft; this strategy is called *tanking*[5]. We decided to test our claim in Section 3.1.

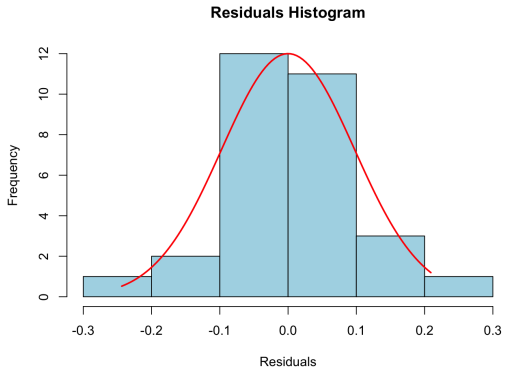

Figure 4.

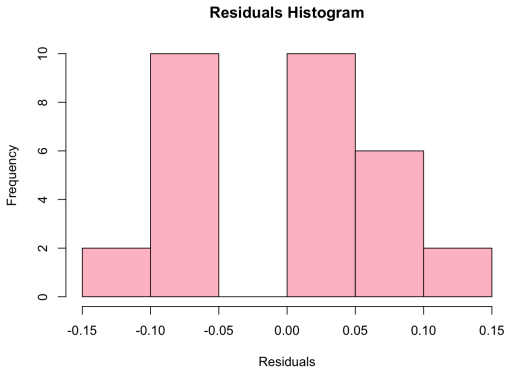

Figure 5.

The third group contains the *game statistics*: `AST`, `TOV`, `STL`, `ORB`, `DRB`, `X3PA`, `X2PA` and `FTA`. The linear model gives a $R^2 = 0.7551$ and while the homoscedasticity and mean assumptions are respected, the normality is not, as can be seen in Figure 5 and by a relatively low p-value of 0.05728 resulting from a Shapiro - Wilk test. We claim that also the problems that showed up in this model are due to the presence of teams that lose on purpose i.e. *tanks*.

3 Tests

We introduced a distinction between *contenders* and *tanks*, respectively teams who play competitively and teams that do not. In particular, we gave the status of *tanks* to the last eight ranked in the 2021-2 season, even though the concept of *tanking* is not rigorous and the ranking does not provide the only indicator to determining it. To *tank*, teams start less-skilled players: we claim that there is statistical evidence that *contenders* shot with a higher percentage from the field, as shown in Figure 6.

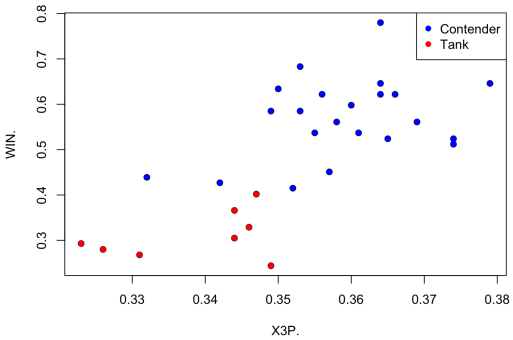

Figure 6.

Except for some outliers, all contenders have a `X3P` above a given threshold, around 0.35.

3.1 t-tests

We decided to test the *influence* of being a contender on `X3P.` i.e. the Null Hypothesis:

$$H_0 = (\mu_c = \mu_t)$$

where μ_c is the mean of `X3P.` for *contenders* and μ_t that for *tanks*, against the Alternative Hypothesis:

$$H_1 = (\mu_c \neq \mu_t)$$

By setting $\alpha = 0.05$ as usual, we rejected the Null Hypothesis with a p-value = 0.0004431 and thus concluded that there is enough statistical evidence that being a *contender* influences the `X3P.` parameter.

We successfully found a similar influence on the parameters `PACE`, `AST` and `DRB`, with respective p-values for the t-test: 0.02375, 0.01043 and 2.2×10^{-16}; indeed even though *contenders* start better players, they also need to implement strategies which increase these statistics: there is statistical evidence that playing, for instance, with a higher pace has proved a winning tactic.

There wasn't instead enough statistical evidence to reject the same Null Hypothesis for `X2P.`, since the p-value = 0.1118. Indeed, in the last fifteen years, teams noticed the significative impact of three points shots in team statistics: to match the expected value for three and two points shots, the relative percentages should respectively be 33% and 50%; in our data there is statistical evidence that the best teams shoot *threes* way better than *twos*, with respect to the losing teams[6]; this will be further justified in Section 5.

3.2 Step-up

Finally, to understand which parameters are relevant to determine our target `WIN.` for *contender* teams, we used the step-up method, starting with an empty model i.e. only intercept, testing each $H_0 = (\vartheta_i = 0)$ separately $\forall i$, and adding the parameter with the smallest p-value thus repeating the procedure with the larger model.

The resulting model, with a $R^2 = 0.7987$, adjusted to 0.7513, contains 4 covariates: `oEFF`, `dEFF`, `X3PA` and `X3P.`, as can be seen below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.134830	0.821342	0.164	0.871543
dEFF[status == "contender"]	-0.027879	0.004457	-6.255	8.72e-06
oEFF[status == "contender"]	0.028913	0.006186	4.674	0.000218
X3PA[status == "contender"]	-0.005893	0.003191	-1.846	0.082307
X3P.[status == "contender"]	1.263079	0.946797	1.334	0.199783

Residual standard error: 0.04454 on 17 degrees of freedom
Multiple R-squared: 0.7987, Adjusted R-squared: 0.7513
F-statistic: 16.86 on 4 and 17 DF, p-value: 9.418e-06

Even though `X3P.` comes with a p-value $\simeq 0.2$, the step up procedure suggests that it contributed to improve the AIC score. The ambiguity is due to the small size of the sample, the distribution of the data, since most of the values are very close to the mean (0.3589) and being *far* from it generates either a drastic drop or a steep increase in `WIN.`, and the restriction of our analysis to *contenders* only.

4 Validation

To test the validity of our model, we decided to predict `WIN.` for the 2022-3 season's data, found again in two different datasets[7], [8], containing the same variables as those for 2021-2.

Since main teams' strategies don't change dramatically over one year, we assumed that teams' statistics from both seasons are identically distributed and independent, thus the previous assumptions for the multiple linear regression, i.e. zero mean, homoscedasticity and normality, are still valid.

We thus computed the following indicators: RMSE = 0.04540824, MSE = 0.002061908 and $R^2 = 0.6163482$, which suggest that our model makes accurate predictions, while the low *adjusted* $R^2 = 0.2339756$ indicates the presence of some parameters with a low incidence.

5 Conclusions

The relevance of `oEFF`, `dEFF`, resulting from the step-up method, was expected since they measure how much each team is able to convert possessions into points made and unable to stop the other team from doing the same.

The presence of `X3P.` and `X3PA` was instead not foregone but, with further analysis, we realized that they indeed summarize an evident trend in the latest seasons[9], from 1979 to nowadays, i.e. the strategy of shooting *threes* rather than *twos*: this is evident in Figure 7, containing the plot of the `X3P.` and `X3PA` over the recent years.

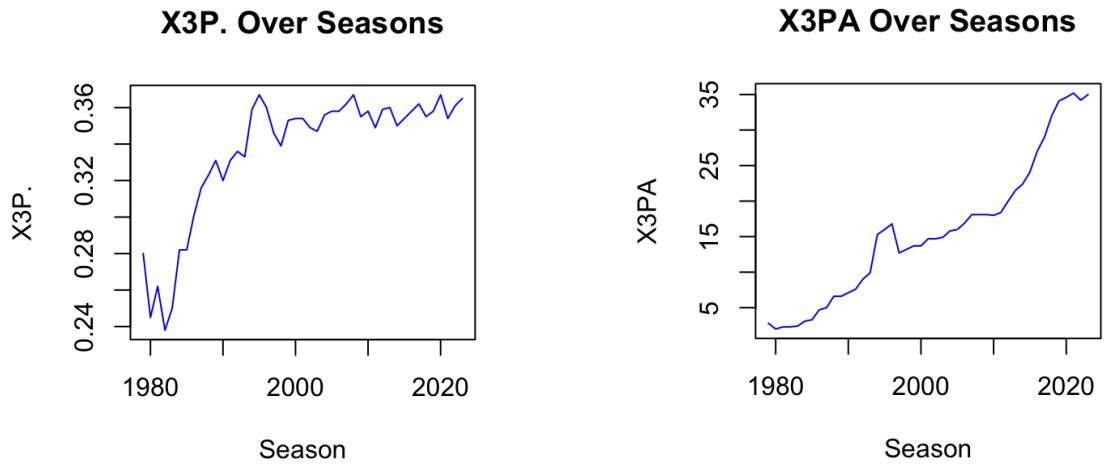

Figure 7.

Although `X3P.` has been growing slower in the last twenty years, the actual number of shots attempted, i.e. `X3PA`, has been continuously increasing, indeed in our model it has a much lower p-value.

As can be seen in Figure 8, the total of points scored by each team, `PPG`, does not follow a clear pattern, but `X3PPG`, the total of points *from three* over the total points scored by each team, has been increasing year by year; this suggests that the favorite strategy to score is attempting *threes*.

Further analysis may be carried out to determine whether the trend described will continue on this steep pattern or reach a *plateau*, as already attempted by some authors[10].

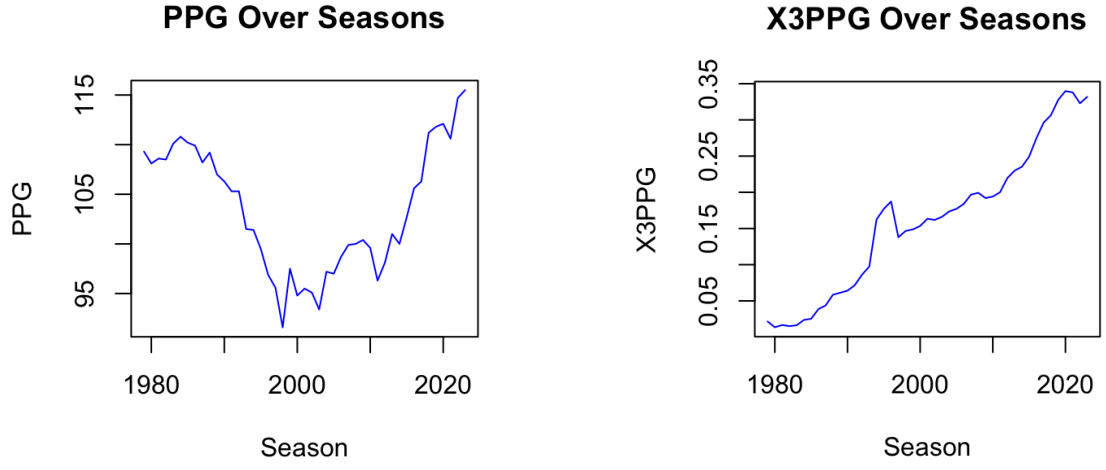

Figure 8.

5.1 Limitations of our study

The datasets we used all had the same *weaknesses*: they contained few data, due to their nature, and among them only a part was useful for our purpose, thus restricting the available information even more. Moreover, we decided not to use the so-called *advanced statistics*, plenty of which are relevant and widely used in the field of sports analytics. Such flaws are the main cause of the ambiguity behind the validation of our model, summarized by the *adjusted* R^2.

Furthermore, restricting the analysis to linear regressions overlooks potential non-linear relationships or interactions among variables, thereby potentially oversimplifying the intricate nature of team performance and game outcomes; such an approach might yield predictions or conclusions that do not accurately reflect the multifaceted reality of the sport.

References

[1] https://www.nbastuffer.com/2021-2022-nba-team-stats/

[2] https://www.basketball-reference.com/leagues/NBA_2022.html#all_per_game_team-opponent

[3] https://hackastat.eu/en/learn-a-stat-strength-of-schedule-sos/

[4] https://www.basketball-reference.com/

[5] https://en.as.com/nba/what-is-tanking-in-the-nba-and-why-do-teams-tank-n/

[6] https://nycdatascience.com

[7] https://www.nbastuffer.com/2022-2023-nba-team-stats/

[8] https://www.basketball-reference.com/leagues/NBA_2023.html

[9] https://www.basketball-reference.com/leagues/NBA_stats_per_game.html

[10] https://link.springer.com/article/10.1007/s12122-014-9193-5