An Introductory Statistical Analysis on Ten Years of NBA Playoff Performance Metrics

CHOI, Ho Chung; CHOW, Hau Cheung Jasper; LIN, Chuan-en; YANG, Guang (alphabetically)
Hong Kong University of Science and Technology
{itsc1, hcjchow, clinaf, itsc3}@connect.ust.hk

Introduction

The National Basketball Association (NBA) is arguably one of the most well-known professional basketball leagues in the world. At these levels of professional play, players and coaches have to be exhaustive and rigorous in squeezing out any edge they can get, as it could spell the difference between victory and defeat. In fact, there are numerous reports [1] that show the increasing tendency for professional coaches to turn to data analytics and statistical methods to pinpoint weaknesses and capitalize on their team's strengths. The aim of this report is therefore to provide an introductory analysis into how some strategic factors may affect how far teams progress in the playoffs bracket (where the team with the most wins prevails as the season champion). More specifically, we wish to investigate two hypotheses as follows:

1. Several years ago, the scores in playoff matches often only reached around 90 points. However, we have seen an increasing trend where scoring has become more and more competitive, where each match often exceeds 100 points and sometimes even reaches 130 points. We suspect that winning teams have now strategized more towards improving offense. Therefore, our first hypothesis is: **Higher offensive rating leads to more wins** (H_1) **.**

2. Another trend we see in playoffs is that winning teams are often dominated by a few select "star players". For example, for many years, the Cavaliers have barely been able to make it into the playoffs. However, since Lebron James joined the team, the Cavaliers have constantly dominated the league, grabbing a championship and multiple first runner-ups within just a few years. Hence, our second hypothesis is: **Teams with more star players win more** (H_2)**.**

To evaluate our hypotheses, we performed several statistical analyses on data collected from the NBA Advanced Stats website (*). The manually curated and cleaned dataset that we used can be downloaded in the link in the bibliography (**). The data we collected was calculated by statisticians on the NBA's official stats team and is publicly available. This is an observational study since we are examining a sample of the games played, namely, those taken over the past decade, in order to infer whether or not our two hypotheses H_1 and H_2 are true over the population of all playoff games. The data was collected for many purposes, but among them is providing a metric for a player/team's performance and judging how

much a team/player is worth. These statistics play a key role in explaining the price difference between the Atlanta Hawks, who were sold for $850 million USD, and the Houston Rockets, who were sold for $2.2 billion USD [3].

Methods

Our data (CSV) consists of ten years of NBA playoff data, ranging from 2010 to 2019. For each year, we have 16 rows representing each team that made it in the playoffs for the given year. The teams are also ranked in descending order, from the team with the most wins to the team with the least wins. For each row entry, we have various basketball metrics relating to the team's overall playoffs performance, such as offensive rating, assist percentage, and so on. We defined our Y (response) variable to be the number of wins and X_i (explanatory) variable to be the ith performance metric, where *i* ranges from 1 to 14. To test our hypotheses, we are only concerned with the "Offensive Rating (OffRtg, denoted by X_1)" and "Player Impact Estimate (PIE, denoted by X_{14})" performance metrics. The Offensive Rating measures a team's points scored per 100 possessions. The PIE measures a player's overall statistical contribution against the total statistics in games they play in. The formula for PIE is:

> PIE = (PTS + FGM + FTM - FGA - FTA + DREB + (.5 * OREB) + AST + STL + (.5 * BLK) - PF - TO) / (GmPTS + GmFGM + GmFTM - GmFGA - GmFTA + GmDREB + (.5 * GmOREB) + GmAST + GmSTL + (.5 * GmBLK) - GmPF - GmTO)

In plain English, the PIE takes into account factors like how many points the player scored (PTS), the percent of free throws they made for their team (FTM), the percentage of a team's blocks that a player has made while on the court (BLK), among a variety of other factors, to determine how important the player is to that team's success. Further information on what the rest of these numbers mean can be found in the bibliography [2].

The statistical methods employed in this report are confidence intervals, hypothesis testing, computation of correlation coefficients, and linear regression.

A confidence interval is an interval estimator. A $(1-\alpha)*100\%$ confidence interval for a population parameter x in the form $x \in [a,b]$ with $a \leq b$ implies that in repeated sampling, $(1-\alpha)*100\%$ of the samples would produce a confidence interval that contains the true value of x.

Hypothesis testing is a key technique used to infer a truth about the entire population, i.e., the set of all playoff matches. To this end, we construct two competing hypotheses (that say different things about the population), namely the null hypothesis and the alternative hypothesis. From this, we compute some statistic T based on our observations of the sample of data we obtain and calculate the probability that we observe a more extreme value of T given that the null hypothesis is true. This value is the p-value. If the p-value is less than the significance level, α, then it suggests the observed data is unlikely when the null hypothesis is true, indicating we should reject the null hypothesis.

Unfortunately, due to the very nature of how we randomly select our sample, we are prone to making errors in our inferences. We may choose some sample from which we determine we should reject the null hypothesis when it is in fact true. This is termed the type I error, and is commonly denoted as α. Alternatively, we may choose some sample from which we determine we fail to reject the null hypothesis when we should. This is termed the type II error, commonly denoted as β. For a fixed sample size, type I error is inversely proportional to type II error, so we cannot mitigate both kinds of error simultaneously without increasing the sample size.

The correlation coefficient is the measure of association between two quantitative variables. It does not imply causality. Its values range from -1 to 1; the larger the absolute value of the coefficient, the stronger the linear relationship is between X and Y. A correlation coefficient of 0 means that there is no linear relationship between X and Y. However, it is important to preface our analysis that the correlation coefficient is sensitive to outliers, and the linear relationship may fall apart outside the given range. In the data we collected, the PIE values all fell between 32.8 and 58.8; outside of this range, the linear relationship may break down. Once we have established a linear relationship between an explanatory variable X and a response variable Y, linear regression is the method of summarising that relationship with a linear function. Most commonly, it is expressed as $Y = \alpha + \beta X + \varepsilon$, where ε is a normally distributed error term with mean 0 and variance σ^2. A good choice for a linear relationship is therefore the one that minimizes the square of the sum of the errors; known as the least squares estimator. This is defined to be: $\hat{Y} = \hat{\alpha} + \hat{\beta} X$, where $\hat{Y}$ is an estimate of Y, and $\hat{\alpha}$, $\hat{\beta}$ are estimators for the coefficients of the linear relationship based on the sample, and ε = error = $Y - \hat{Y}$.

Result

We first investigated whether or not there exists a linear relationship between each X_i (X_1 = OffRtg, X_{14} = PIE) and Y for each of our hypotheses. We computed the estimated Pearson's correlation coefficients and conducted hypothesis testing in which our null hypothesis was that the population correlation coefficient $\rho = 0$, and the alternative hypothesis was that $\rho \neq 0$. We set our estimated Pearson's correlation coefficient threshold to 0.5 with significance level $\alpha = 0.01$, thereby limiting the amount of type I error to be at most 0.01, and meaning we deemed each X_i and Y to be correlated only if the computed coefficient exceeds 0.5 with p-value less than 0.01.

With our estimated Pearson's correlation coefficient threshold and significance level, we cannot say PIE (X_{14}) and Wins (Y) are not correlated, since p-value = $2.2 * 10^{-16} < \alpha = 0.01$ and the estimated correlation coefficient was 0.672 > 0.5. Therefore, we continue our statistical analysis for our H_2. However, even though the p-value = $1.725 * 10^{-11} < \alpha = 0.01$, we found the sample correlation coefficient between OffRtg (X) and Wins (Y) to be relatively weak, below the set threshold of 0.5. (In both cases, we failed to reject the null hypothesis that states that the variables are uncorrelated.)

Therefore, we may conclude that there exists a linear relationship between both (PIE and Wins) and (OffRtg and Wins), but we will only examine (PIE and Wins), i.e. H_2, since we calculated its estimated

correlation coefficient to be above our desired threshold of 0.5. In order to have conducted this t-test, we need to have assumed that the two variables are (bivariately) normally distributed, and that each of the variables are normal and have homogenous variance at all levels of the other variable. In order to verify these assumptions, we can run the Shapiro-Wilk test and Bartlett's test.

Bartlett's test is a hypothesis test where the null hypothesis is that the samples come from populations that exhibit homogeneity of variance, and the alternative hypothesis is that there exists one or more pairs of populations with non-identical variances. The p-value returned from the R command *bartlett.test()* is exactly the p-value for this hypothesis test. In our example, the p-value of 0.051 is greater than our desired significance level of $\alpha = 0.01$, so we can conclude that we do not have enough evidence to reject the hypothesis that the variances of the population are equal at this significance level.

The Shapiro-Wilk test is a hypothesis test where the null hypothesis is that the sample came from a normally distributed population, and the alternative hypothesis is that there exists one or more populations that are not normally distributed. The p-value returned from the R command *shapiro.test()* is exactly the p-value for this hypothesis test. We ran this test on both the PIE and the Wins, and the p-values were 0.00806 and $1.741 * 10^{-10}$ respectively. As both were less than our desired significance level of $\alpha = 0.01$, we can conclude that we must reject the null hypothesis that the samples came from a normally distributed population. As the assumption of normality is not met, we can instead use the non-parametric Spearman and Kendall's tau rank. The p-values for these tests are both less than $2.2 * 10^{-16}$, much lower than $\alpha = 0.01$, suggesting that we should reject the hypothesis that the true correlation coefficient is 0. Thus, we may conclude there exists a linear relationship between PIE and the number of wins.

Continuing our analysis for H_2, we performed simple linear regression between PIE (our explanatory variable) and Wins (our response variable). We also plotted our data points as a scatterplot with a fitted linear regression line and residual errors (see Fig. 1).

Nonetheless, we learned that linear regression is built on top of several assumptions. Therefore, to evaluate the validity of our model, we used the built-in function in R to visualize our diagnostic plots for linear regression so as to check the validity of each of the assumptions. The assumptions are: (1) linearity, (2) independence, (3) normality of error, and (4) homogeneity of variance.

1. Linearity: From the "Residuals vs Fitted" graph in Fig. 2, we see a loosely horizontal line around the zero point, although it is a bit weak. Therefore, we can say that even if there is a linear relationship among the data, it is not a particularly strong one.

2. Independence: we assume this to be true.

3. Normality of error term: From the Q-Q plot in Fig. 2, we see that the residuals largely follow the straight diagonal line. Therefore, we can say that the error in our data is largely normally distributed.

4. Homogeneity of variance: From the "Scale-Location" graph in Fig. 2, we see a loosely horizontal line. Similar to the "Residuals vs Fitted" graph, it is a bit weak and the data points are

not too equally spread across. Therefore, we can say that even if our data has loose homogeneity of variance, to some degree there is still the presence of non-constant variance.

While not being one of the aforementioned three assumptions of a linear regression model, it may be useful to identify outliers that may skew our regression results. From the "Residuals vs Leverage" graph in Fig. 2, we can generally identify the outliers as the data points on the upper or lower right corners of the graph.

From our linear regression model, we have the following linear formula: **[Wins] = 0.82859 * [PIE] -29.50336.** Our coefficient of determination (R^2) is around 0.45, meaning that this equation 'explains' 45% of the variability between PIE and Wins. We can also compute the 99% confidence interval for the coefficients of our linear model to be: $\hat{\alpha} \in [-38.664505, -20.342218]$, $\hat{\beta} \in [0.6392288, 1.017949]$. In our R code, we have also calculated the 99% prediction interval for X = PIE at 50.0 to be $[0.444625, 23.40755]$. (Readers may further adjust the parameters provided in our R code to compute the prediction intervals at a different significance level and/or at different predicted value(s)). See Fig. 3 (the red dashed lines denote the 99% prediction interval, and the gray shaded region denotes the 99% confidence intervals of the coefficients of the linear model.)

Conclusion

Ultimately, we concluded that there did not exist a significant enough correlation between OffRtg and Wins, since the estimated Pearson's correlation coefficient was less than 0.5. Conversely, we concluded that there did exist a linear relationship between PIE and Wins, since both non-parametric and parametric correlation coefficient tests revealed a moderate linear relationship between the variables. However, from our diagnostic plots in Fig. 2, we can see that the linearity and homogeneity of variance assumptions are fairly weak and that the data also has several outliers. Therefore, future analysis should focus on the investigation of non-linear models, such as applying log transform or adding a quadratic term. Furthermore, our prediction model may also benefit from removing more extreme data points.

It is also worth noting that because the statistics we collected are aggregates of a team's performance over the entire season, it does not account for what team they played against and the teams' specific player lineup in each match. For example, if a high performing team **(high performing team 1)** played a match against another high performing team **(high performing team 2)** compared to a weaker team **(low performing team),** a win might be substantially less likely. Our statistics also covers nothing about the nature of the win; a 1-point lead and a 20-point lead are treated the same. For more meticulous study in the future, we could consider examining a variable like the Win/Loss ratio for a given team T, perhaps also cross-calibrated to the number of wins/losses played against each team T faced. This bears resemblance to a lot of game theory and Nash equilibrium situations, where, for example, team A has an advantage over team B, and team B has an advantage over team C, but team C has an advantage over team A. This may be due to a variety of factors such as the degree to which one team's players passes to teammates and whether the strategies adopted by the two teams are tailored to be unique to the opponent they are facing.

Bibliography

1. D. Kopf (Oct 2017). Data Analytics have made the NBA unrecognizable. Quartz. https://qz.com/1104922/data-analytics-have-revolutionized-the-nba/
2. NBA Stat Glossary. NBA Stats. https://stats.nba.com/help/glossary/
3. T. Cato (Sep 2017). How much did each NBA owner pay to buy their teams? SBNation. https://www.sbnation.com/nba/2017/9/5/16255168/nba-teams-sold-highest-record-price-all-30

(*) Original data obtained from https://stats.nba.com/teams/advanced/?sort=OFF_RATING&dir=-1&Season=2018-19&SeasonType=Playoffs.
(**) The cleaned dataset we used can be downloaded from https://filebin.net/katyfwqathrdyve5 (also comes with our .zip submission)

Appendix

Fig. 1 - Linear regression model with residual errors.

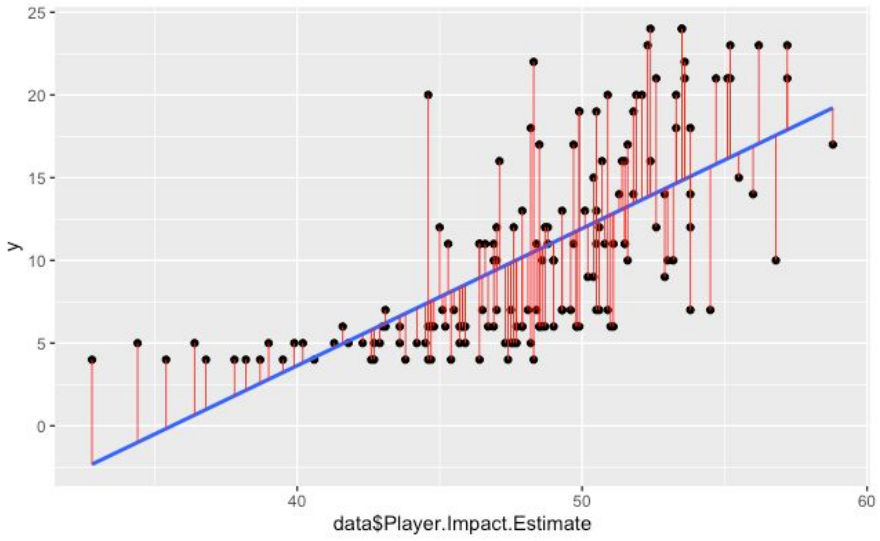

Fig. 2 - Diagnostic plots

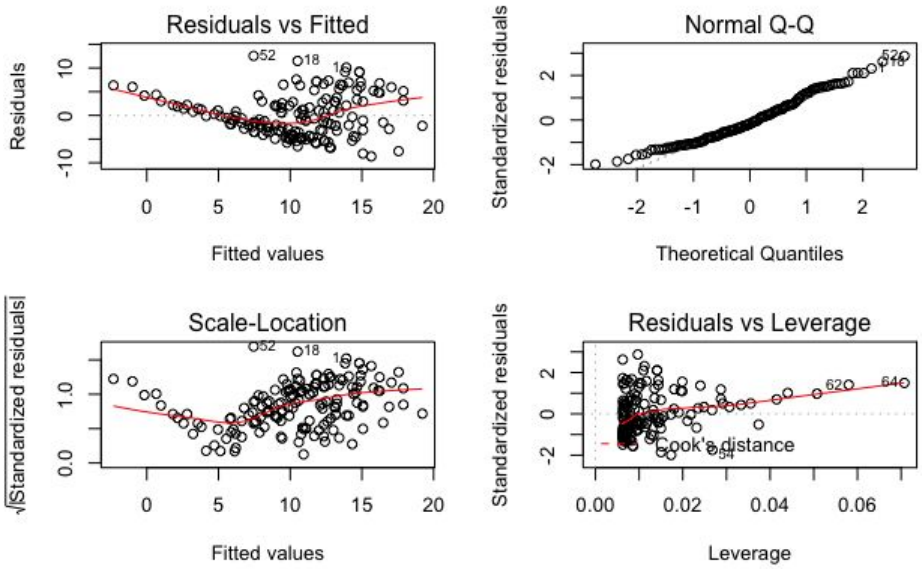

Fig. 3 - Linear regression model with 99% confidence and 99% prediction interval