Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics

Project of CS282 - Machine Learning

May 2016

ABSTRACT

Not long ago, if the Golden State Warriors had wanted to figure out how to best defend Pelicans star forward Anthony Davis, they might have sent a scout to a game or watched video clips. For their recent first-round playoff matchup, however, they had another way. As of this year, almost every NBA team has access to sophisticated tracking data that can tell them the position of the ball and every player on the court for every second of every game. The data made available by advanced technology and analytic tools is starting to revolutionize professional basketball, influencing everything from game strategy and player conditioning to how fans interact with the sport. Players, coaching staff, fans as well as sports analysts are trying to take advantage of the abundance of data to satisfy their different analytical needs.

This thesis analyzes the correlation between individual player's statistics and their team's performance, and develops a prediction model that can be used to forecast regular season results of NBA teams based on common player statistics. Data from the past twenty seasons were collected via the Internet and analyzed using R. The outputs of least-squares regression analysis yielded robust models that had strong positive R^2 results and significant F-statistics from the Wald test that evaluated the model's goodness of fit. The predictions on current regular season's results based on the model proved to be satisfactory. In addition, possible complications and issues of the model were also thoroughly considered and discussed in the thesis.

TABLE OF CONTENTS

§ Introduction & Background

As an avid basketball fan, I watch NBA games online, on TV, in the bars and in the arenas almost religiously. Since I started college at Berkeley and went to a few Golden State Warriors' home games in Oracle arena, I quickly became a Warriors' fan. I absolutely enjoy watching Stephen Curry, Klay Thompson and company putting on a show every time they step on the floor. The Warriors have amazed their fans in every way possible, be it Curry's wizard behind-the-back dribble followed by a pull-up three point game winner, Thompson's historic 37-point quarter against Sacramento Kings, or Bogut's ferocious alley-oop dunk. Having been a solid team for several consecutive seasons, the Warriors finally unleashed their full potential and clinched the best record in the NBA for 2014–2015 season, thrilling fans here in the Bay as well as nationally and breaking a long list of franchise records along the way.

Unlike baseball teams that play 162 games during a single season, basketball teams in NBA only play 82 (or even fewer) regular season games. There, nonetheless, still are ample data generated from those games and the data are readily available from various sources. There are numerous well-written papers and books analyzing baseball teams' performance[i,ii,iii] based on their players' batting averages and other common statistics, yet significantly fewer researches have touched upon other sports like basketball. Therefore, in this paper, *I will try to utilize common basketball statistics that are accessible to all fans of the sport and explore the connection between the statistics of individual players and the performance of their respective teams. The ultimate goal of the paper is to develop a model that can satisfactorily predict NBA teams' regular season performance based on common statistics of their players.*

Though the analyses of game-related statistics have been popular among coaches for a long time[iv], the types of data collected and metrics created have been constantly evolving to better serve changing analytical needs. On the most basic and tangible level, most of the data concerning players' performance are presented in the form of boxscores, the tabulated results of the players after each game they play. The numbers in the boxscores gauge the value of each player that participated in the game as well as his/her teammates and opponents. At first, boxscores only included limited statistics such as players' scoring, rebounding, assisting numbers and field goal percentages. People, however, later realized that these statistics do not necessarily fully capture the value of a player and his/her impact on the game for a number of reasons. First, existing statistics are very offensively biased. They reward players who are more aggressive on the offensive end, thus racking up more points and assists, but discount players who contribute to the team mainly through their defense such as blocking shots and making steals. In the meantime, these statistics tend to favor those players who have more control over the ball on the court. Point guards, for instance, are responsible for driving the ball up the court and directing a team's offense, so they usually end up looking better on the stats sheet. Nevertheless, the more a player handles the basketball, the more likely he is to make a bad decision and turn the ball over to the opponents. The ability for a player to take care of the basketball is not reflected in anyway by existing statistics, but can be a crucial part of a player's game. Therefore, the league started to record turnovers of players from 1970–1971 season.

Nowadays, with the help of computers and powerful motion sensors, the data we can collect from the game and the analytical tools we have available have pushed sports science to another level. Teams like Golden State Warriors are utilizing big data[v] and other technologies to help players and coaches understand their games better and change their practice drills as well as decision-making procedure accordingly. This paper, however, is not intended to take advantage of those more advanced statistics, such as players' positioning on the court, miles covered in a game, efficiencies against opponents of different heights, etc., that teams collect proprietarily because those data are not uniform across teams and not readily available to general fans. Instead, the paper will focus on more common statistics like points scored, assists, rebounds, blocks, steals, etc. and analyze the usefulness of these variables in characterizing and understanding players' contribution to their teams' success.

Recently, some sports websites and analysts saw the value of combining simple statistics to coin more indicative metrics that are better adjusted for different teams and different styles they play with. For instance, the offensive statistics of a player who plays for a more defensively focused team may not be as glamorous as the statistics of a player who plays for a more offensively oriented team. However, that does not necessarily mean the former is less productive or efficient than the latter. To mitigate the discrepancies in teams, sports analysts tried to normalize the differences in a number of ways. One way, for example, is to normalize players' statistics by possessions. Fast-paced teams, by nature, have more possessions per game than slow-paced teams. Looking at players' productions per possession, therefore, provides more accurate insight into players' actual effectiveness.

One of the most powerful and widely recognized metrics, which is derived from basic player statistics to gauge a player's strength, is **Player Efficiency Rating (PER)** developed by *John Hollinger* from ESPN[vi]. Among many merits of this metric is that it is both a pace-adjusted and per-minute measure. As mentioned before, adjusting for the pace of the teams makes sure that players on slow-paced teams are not penalized for having slightly lower numbers on their stats sheet than their counterparts in offensive power houses. In addition, because PER is a per-minute measure, it is easy to compare players who have huge disparity in minutes played in games (for instance, the starters versus reserves). Since players' PERs are readily available, I will leverage this metric when building my model. In the meantime, I will also explore the possibility of developing a new metric similar to PER using support vector machine to connect players' statistics with their teams' performance. As stated earlier, the goal of the paper is to develop a model that can be used to forecast the regular season results of NBA teams and predict which teams would be able to make the playoffs given their current rosters and individual player's statistics. However, there are several complications that might undermine the model I am trying to build. For instance, players might get injured or be traded to other teams during the season; teams might change their coaching staff during or in between seasons and thus becoming a vastly different team with new styles of play; teams' records can be so close that the differences in players' statistics would fail to be accurate indicators of teams' relative strengths. I will discuss these factors in my model and address potential issues they might cause in *Discussion of Results* section.

§ Materials and Method

As mentioned in the prior section, NBA basketball is highly statistically developed and there is an abundance of sortable basketball statistics, most of which are free and easy to obtain. The data used in this paper are taken from the website *basketball-reference.com*. This website provides downloadable data for both players and teams from 1946-1947 season till present.

I chose to focus on the most recent 20 regular seasons and base my model on the data from these 20 seasons because as the game evolves over time, the characteristics that distinguish teams' winning and losing also change considerably. Therefore, using the statistics that were too far back in time might not be as relevant as using more recent data when it comes to developing a model to predict teams' performance in the near future. For instance, in 1980s, a team's success largely depended on the strengths of the team's big men[vii] (in other words, centers and/or power forwards) because, at that time, the pace of basketball games was rather slow and thus dominant post players were usually the ones that made the difference between a successful and a struggling team. This, however, is not the case in today's NBA games. It is not only because this is a golden era for smaller players (point guards and/or shooting guards) with amazing shooting abilities and ball handling skills like Stephen Curry, James harden and Kyrie Irving, but also because the league has made a number of rule changes that benefit offensive players, especially the shooters in hope of making the game more thrilling and exciting for fans to watch. In consequence, I will only use the data from last 20 seasons in this paper, with an emphasis on the data from more recent seasons.

The data collection process was more difficult than I expected. I first tried to gather data from *ESPN.com*[viii] since it has all the statistics that I need displayed on the website. However, the data was not easily downloadable. I wrote some Python scripts in an attempt to extract the data directly from the website's source code. Though I successfully accessed the data and downloaded it, the format of the data was very hard to interpret and made it difficult to perform further analysis on the data. Thus I decided to turn to other sources and found that *basketball-reference.com*[ix] offers CSV files of sortable data for download. I then downloaded the stats both for players and teams for the last 20 seasons, fed them into R, read and cleaned up the data using R scripts with the help of regular expressions.

I organized the data into three main categories: ***nba(YEAR)*** contains the statistics of all players played during YEAR-(YEAR+1) season; ***standings(YEAR)*** contains teams' record (including total, conference, division as well as monthly records) for the corresponding season; ***team(YEAR)*** contains data that aggregate all individual players' stats by teams they play for. With the data stored in proper form, the regression model could be built.

Since I am exploring the connection between teams' performance and players' statistics, I will use statistics for both teams and individual players. For teams, I will use **teams' win ratio (Wins / Total Games Played * 100)** as a proxy of teams' performance over a season. Win ratio, from my perspective, is more informative than teams' ranking at the end of the season because win ratio is more quantifiable and can show by how much margin one team is better than another for a given season. As for players, I will use **Hollinger's PER** along with other basic statistics to value their strengths. Since my model will largely rely on PER, I would like to go into some details about this metric in the following subsection.

❖ **What is PER?**

PER is a metric that aims to measure a player's effectiveness with a single number. It takes into account almost all statistics kept by the NBA, and weighs the player's production by minutes played per game, and number of team possessions per game.

❖ **How is PER Calculated[x]?**

The formula used to calculate unadjusted PER is as follows[1]:

```
uPER = (1 / MP) * [ 3P + (2/3) * AST + (2 - factor * (team_AST /
team_FG)) * FG + (FT *0.5 * (1 + (1 - (team_AST / team_FG)) + (2/3) *
```

[1] Detailed definitions of the terms used in uPER formula are included in the Glossary section in Appendix

```
(team_AST / team_FG))) - VOP * TOV - VOP * DRB% * (FGA - FG) - VOP *
0.44 * (0.44 + (0.56 * DRB%)) * (FTA - FT) + VOP * (1 - DRB%) * (TRB -
ORB) + VOP * DRB% * ORB + VOP * STL + VOP * DRB% * BLK - PF * ((lg_FT /
lg_PF) - 0.44 * (lg_FTA / lg_PF) * VOP)]
```

Some things to note in the formula are:

a) 1979-80 — debut of 3-point shot in NBA

b) 1977-78 — player turnovers first recorded in NBA

c) 1973-74 — player offensive rebounds, steals, and blocked shots first recorded in NBA

However, since we are only concerned with the data starting from 1995-1996 season, we do not have to worry about the aforementioned complications.

A few terms used in the formula above that are not quite self-explanatory are calculated as below:

```
a) factor = (2 / 3) - (0.5 * (lg_AST / lg_FG)) / (2 * (lg_FG /
   lg_FT))

b) VOP    = lg_PTS / (lg_FGA - lg_ORB + lg_TOV + 0.44 * lg_FTA)

c) DRB%   = (lg_TRB - lg_ORB) / lg_TRB
```

The formula stated above will give us the unadjusted PER for players, meaning the pace of the team has not yet been taken into account. To account for the different paces teams play with, we define pace adjustment as follows:

```
pace adjustment = lg_Pace / team_Pace
```

and pace factor (which is an estimate of possessions per 48 minutes by a team) for each team is:

```
48 * ((Tm Poss + Opp Poss) / (2 * (Tm MP / 5)))
```

Lastly, the formula for adjusted PER (which I will use in the model) is:

```
aPER = (pace adjustment) * uPER
```

- **Merits & Perils of the PER Statistics:**
  - *Benefits:* PER is a huge step up from looking at standard boxscore statistics. It is much more detailed and accurate than anything one can do with raw statistical totals or per-game numbers.
  - *Negatives:* One major weakness in the original PER concept is lack of consideration for defense. The formula itself largely measures offensive performance. Though there are blocked shots and steals, the formula doesn't account in any way for players who play great individual or team defense.

After carefully defining and discussing the player statistics that I will use, it is time to proceed to the model. I mainly used (multi)-linear regression in analyzing the relationship between players' and teams' stats. I will gauge the effectiveness and accuracy of the model based on the R-squared statistics it produces and the significance of F-statistics when correlating those two variables.

§ Data Analysis and Findings

For each season that I investigated, two major variables are calculated. First is teams' *win ratio*, which, as defined before, equal to the percentage of games that a particular team won during the season (multiplied by 100 for clarity). The other variable is *team PER*, which is calculated by: 1) *Sorting players on the roster of a specific team by minutes they played for that team over the entire season*, 2) *Then multiplying individual PER by the minutes they played, and*

3) *Summing the product (of individual PER * minutes played) for the first twelve players (by minutes played) on a given team.*

I calculated team PER the way as I described above because: 1) PER is a per-minute statistics, thus multiplying a player's PER by his minutes played can serve as an approximation of the total contribution he made towards his team over the entire season; 2) Some teams have made changes in their roster over the season through trade, due to injuries, or at coaching staff's discretion. However, the first 12 players on each team's roster are relatively stable, which indicates that those players are likely to be in the usual rotation of their teams' lineup.

One huge benefit of calculating team PER the way I did is that the formula automatically puts more weight on players with higher PER because high-PER individuals are, by the way the metric is designed, more capable and are more likely to play a lot more than the players with lower PER. Even when an all-star player (usually with very high PER value) gets injured, the formula is not affected because he will play much fewer minutes and thus have much less weight when calculating team PER. Derrick Rose of Chicago Bulls, for instance, had multiple surgeries over last season and thus played very limited minutes. Therefore, even though Rose is a very high caliber player (he is actually the Most Valuable Player of 2011-2012 season), his contribution to Bull's team PER is rather small because of the limited time he played.

The statistics summary shown below are outputs from least-square regression on **log(team PER)**[2] and **team's win ratio**. I only included the results and graphs for the most recent 3 seasons as illustrations of the model.

[2] I took the natural logarithm of team PER because it increases the R^2 statistics by ~1%. The relatively insignificant improvement on R^2 is due to the fact that the natural log function within the range of my PER value is almost linear.

Scatterplot and summary statistics for 2013-2014 Season

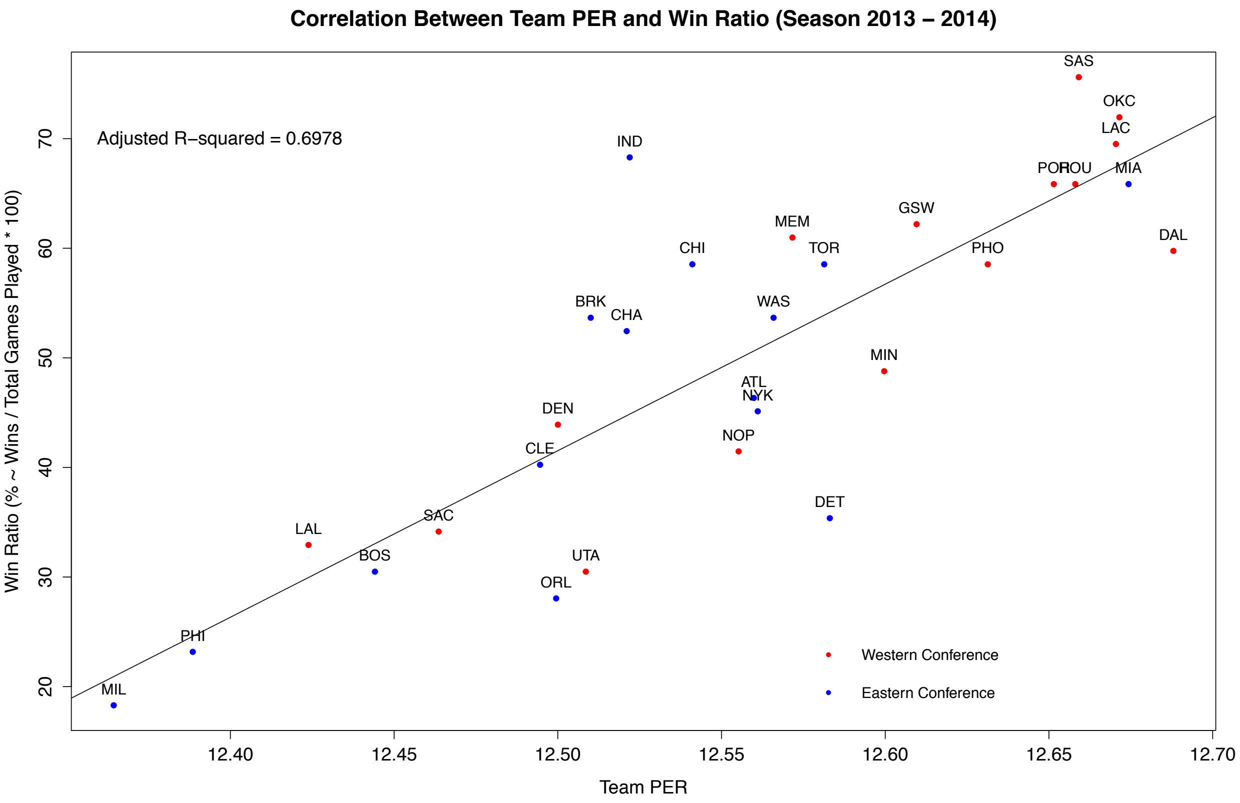


Figure 1: Scatterplot of Team Win Ratio versus Team PER for 2013-2014 Season

We can clearly see a linear trend from the scatterplot, which indicates that the teams with higher PER value are indeed more likely to perform better over the season. Since I color coded teams by conference, we can visually observe that there is a huge disparity between eastern and western conference teams as western conference teams are clustered in the upper right portion of the graph whereas eastern conference teams are clustered mostly in the lower left portion. This phenomenon is exactly what is to be expected since the imbalance of the two conferences has lingered for the past few years. In a way, the apparent difference between eastern and western conference teams as shown in the plot confirms the validity of using team PER as an independent variable and increases the credibility of the model. We then look at the summary statistics of the linear model for further information:

```
Call:
lm(formula = as.numeric(team_PER) ~ winR)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.119066 -0.033274 -0.004973  0.033922  0.095515 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.232e+01  2.959e-02 416.428  < 2e-16 ***
winR        4.661e-03  5.653e-04   8.244 5.68e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04794 on 28 degrees of freedom
Multiple R-squared:  0.7082,    Adjusted R-squared:  0.6978 
F-statistic: 67.97 on 1 and 28 DF,  p-value: 5.677e-09
```

We can see from the summary statistics that the linear model yields a strong R^2 result with an adjusted R^2 of 0.6978. The significant F-statistics from the Wald test further confirms that the linear model decently fits the data and could potentially be used to make future forecasts.

As demonstrated in the residual vs. fitted value plot and normal Q-Q plot below, the linear model results in a very good fit except on a few data points. One of the anomalies is IND (Indiana Paces), as we can easily see from **Figure 1** where it has rather low team PER but a much higher win ratio than expected. Data points like this will make the tail in Q-Q plot deviate from their expected position, and they should be analyzed case by case. However, in general, the linear model results in a satisfactory fit from scatterplot, statistical summary and residual plots.

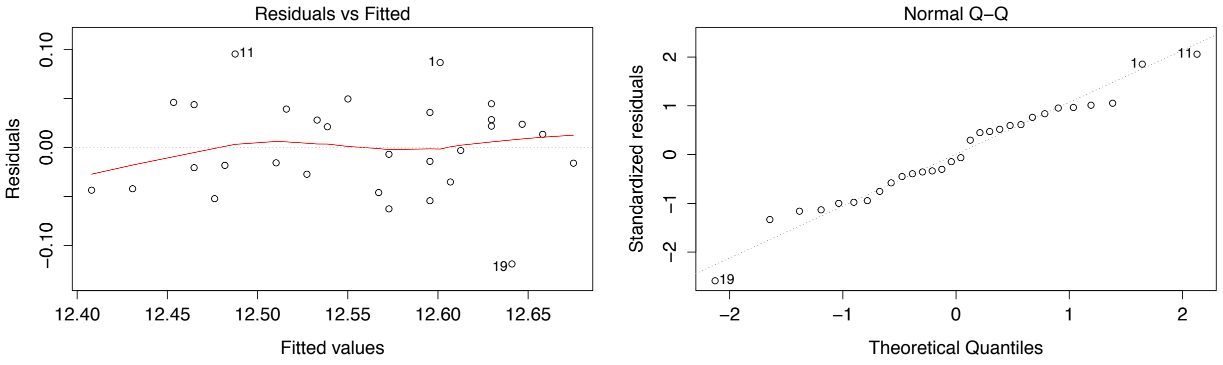


Figure 2: Residuals Plot and Normal Q-Q Plot for Linear Model for 2012-2013 Season

<u>Scatterplot and summary statistics for 2012-2013 Season</u>

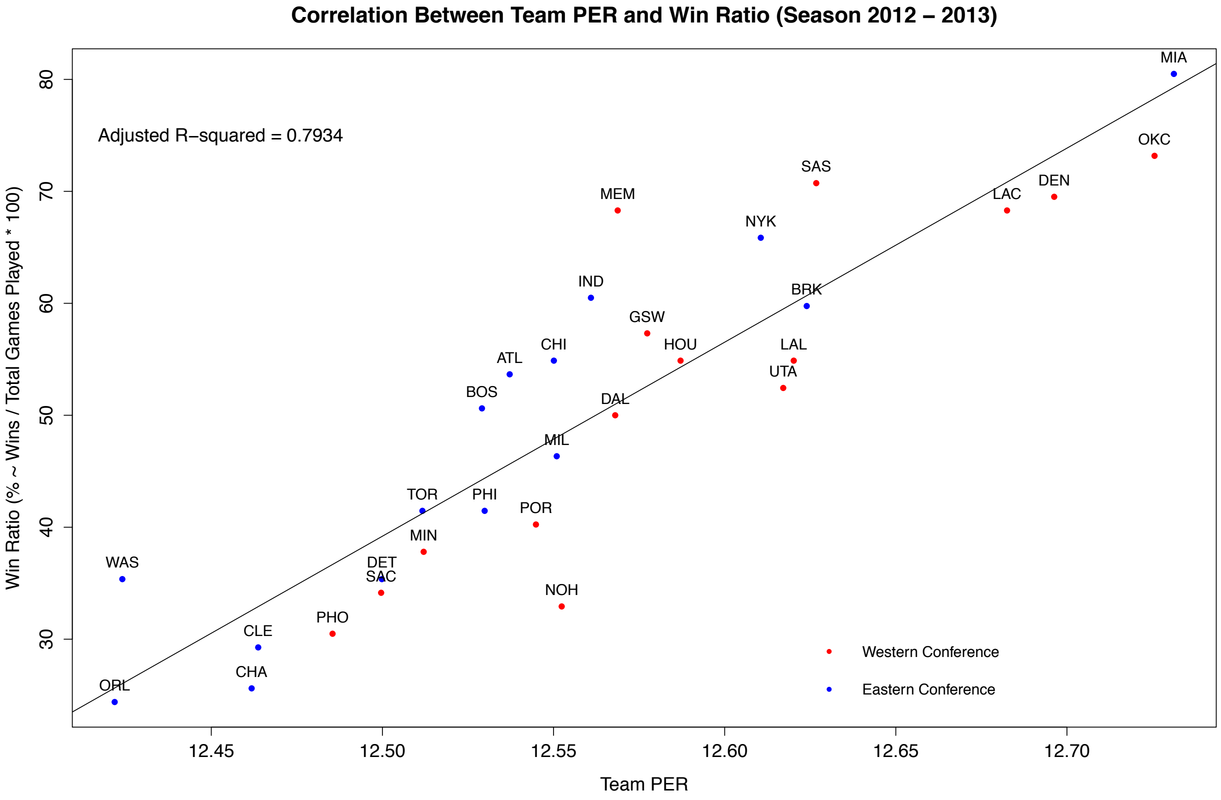


Figure 3: Scatterplot of Team Win Ratio versus Team PER for 2012-2013 Season

Like the scatterplot for 2013-2014 season, the linear trend in this scatterplot is also very apparent. In fact, the linearity is more pronounced in this plot than the previous one partially because the anomalies presented in this graph are less extreme than those from last season's graph. MEM (Memphis Grizzlies) and NOH (New Orleans Hornets), deviating considerably from the regression line, seem to be the ones that do not completely agree with the model this season. However, the discrepancy between eastern and western conference teams as I described earlier is still highly observable, conforming to our previously agreed upon knowledge. Next, we look at the statistical summary of the model:

```
Call:
lm(formula = as.numeric(team_PER) ~ winR)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.07810 -0.02433  0.00559  0.02484  0.06892 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.233e+01  2.278e-02   541.2  < 2e-16 ***
winR        4.620e-03  4.358e-04    10.6 2.62e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03642 on 28 degrees of freedom
Multiple R-squared:  0.8005,    Adjusted R-squared:  0.7934 
F-statistic: 112.3 on 1 and 28 DF,  p-value: 2.621e-11
```

The statistical summary of the linear model yields an even stronger R^2 value than 2013-2014 season, with an adjusted R^2 of 0.7934. Meanwhile, the F-statistics from the Wald test further confirms that the result is statistically significant.

As for the residual plots, we can see that most points in the Q-Q plot are supportive of the model except for a few at either end. This is to be expected because there are some data points that deviate more from the expected value given by the linear model than others. Besides MEM and NOH, WAS (Washington Wizards), in **Figure 3**, has a very low team PER but an unexpectedly high win ratio. The points in residual plot appear to be rather random as expected.

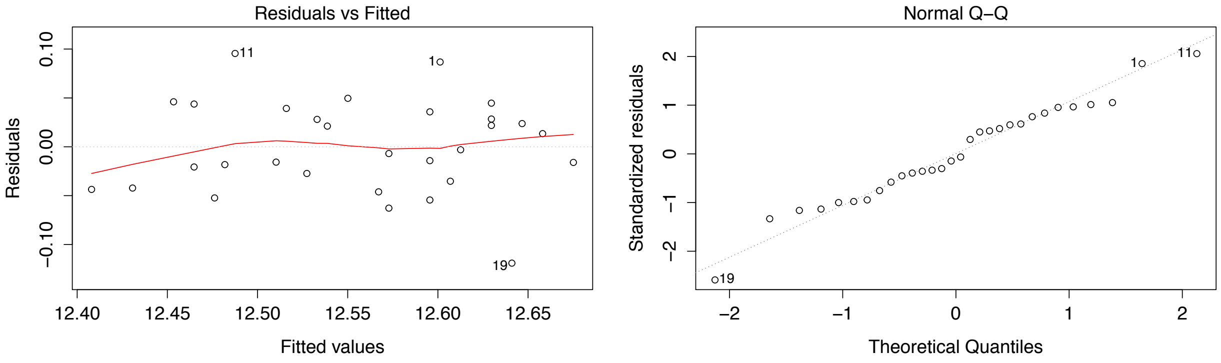


Figure 4: Residuals Plot and Normal Q-Q Plot for Linear Model for 2012-2013 Season

Scatterplot and summary statistics for 2011-2012 Season

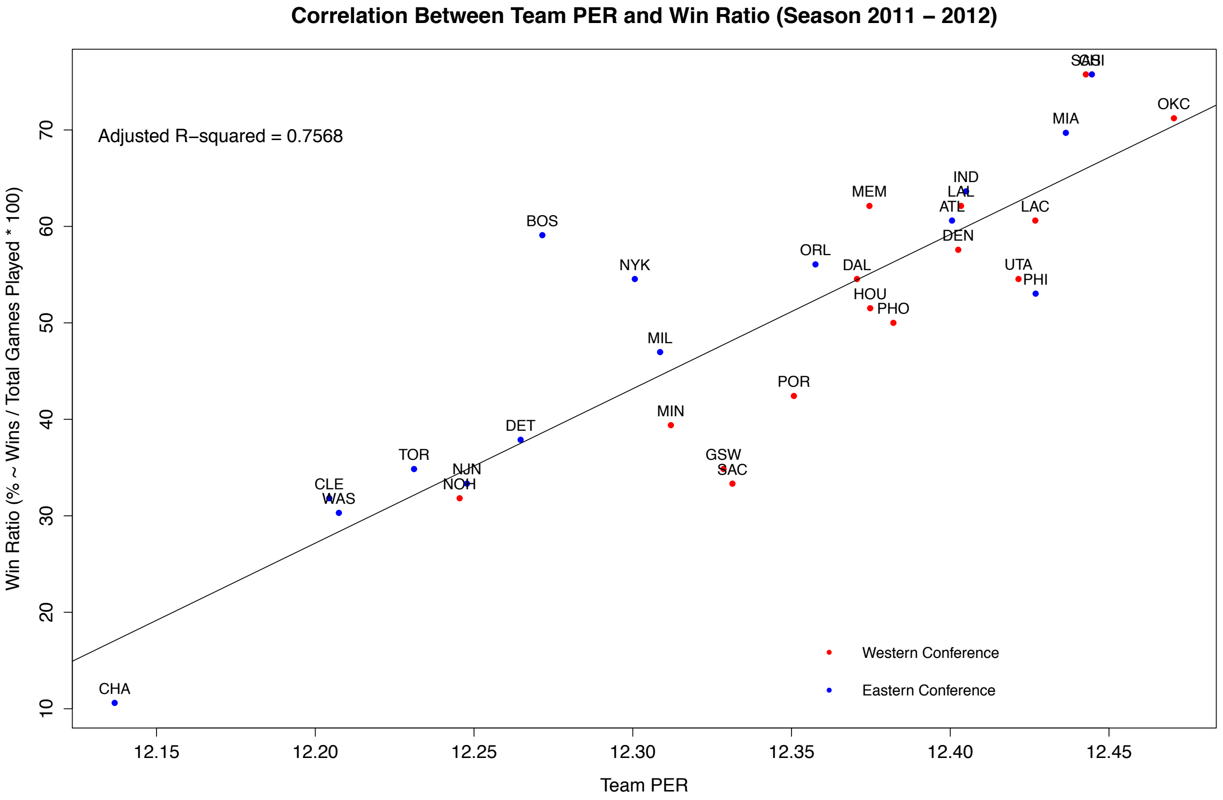


Figure 5: Scatterplot of Team Win Ratio versus Team PER for 2011-2012 Season

Similar to the previous two scatterplots, this plot also demonstrates clear linearity between team PER and team's win ratio. There are some notable anomalies such as BOS (Boston Celtics) which has a higher win ratio than its team PER would suggest. The R-squared value including BOS is 0.7568, indicating a decent fit of the model. However, if we exclude BOS, the R-squared value will increase to 0.8197 (or an improvement of 8.3%). The results of the model would be further improved if we take into account other anomalies such as GSW (Golden State Warriors) and SAC (Sacramento Kings) which have lower win ratio than their team PER would indicate. In general, the linear regression line seems to be a good fit graphically. We further confirm the goodness of fit by looking at its statistical summary.

Results with BOS included:

```
Call:
lm(formula = as.numeric(team_PER) ~ winR)

Residuals:
      Min        1Q    Median        3Q       Max
-0.114685 -0.020992 -0.001734  0.025915  0.069684

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.210e+01  2.619e-02 462.204  < 2e-16 ***
winR        4.783e-03  5.007e-04   9.552 2.62e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04207 on 28 degrees of freedom
Multiple R-squared:  0.7652,    Adjusted R-squared:  0.7568
F-statistic: 91.24 on 1 and 28 DF,  p-value: 2.621e-10
```

Results with BOS excluded:

```
Call:
lm(formula = as.numeric(team_PER[-23]) ~ winR[-23])

Residuals:
      Min        1Q    Median        3Q       Max
-0.068519 -0.023142 -0.007556  0.020585  0.066959

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.210e+01  2.269e-02  533.25  < 2e-16 ***
winR[-23]   4.937e-03  4.359e-04   11.33 9.21e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03641 on 27 degrees of freedom
Multiple R-squared:  0.8261,    Adjusted R-squared:  0.8197
F-statistic: 128.3 on 1 and 27 DF,  p-value: 9.21e-12
```

As one can see from the statistical summary of the model, it produces significant coefficients as well as F-statistics. The model confirms that, to a large extent, a team’s win ratio is linearly correlated to team PER value. The results without BOS indicate that the model sometimes fails to work with teams with high PER that underperform or teams with low PER that overachieve. I will discuss possible reasons for the existence of such anomalies in next section.

§ Discussion of Results

As we can see from the outputs of the model, there undoubtedly is a strong linear correlation between a team's win ratio and the team PER. It makes logical sense because teams with more talented players are supposed to win proportionally more games. The relatively high R^2 values and significant F-statistics confirm the goodness of fit of the model. There, however, still are some issues and complications about the model that I would like to discuss.

First, I want to address the issue that I mentioned in the prior section about anomalies in the model. As we can see from **Figure 1, Figure 3,** and **Figure 5**, there apparently are a number of anomalies in each graph where the model fails to apply. Some possible reasons for the existence of the anomalies include:

- For underperforming high PER teams:
    - Even the teams have high caliber individual players (with high individual PER values) like GSW's Stephen Curry and Klay Thompson, they have not found or employed a good system that can make the players capitalize on their individual brilliance and play well as a team.
    - Players were not consistent in their game and/or not as good when they were playing on the road versus at home, or vice versa.
    - Teams went on long stretches of losing because of major injuries from key players, and did not bounce back quickly once those players returned to the lineup.
- For overachieving low PER teams:
    - As opposed to underperforming high PER teams, teams with lower PER but higher win ratio might have figured out a system that can enhance each player's

game. After all, basketball is a team sport and it requires a concerted effort from all the players to win games. With appropriate system in place, it is not uncommon for teams without super stars to beat teams that rely too heavily on one or two really good players. For instance, ATL (Atlanta Hawks) does not necessarily have all-star players during 2011-2012 season, but they have a rather effective pick-and-roll offensive system that works well for them and thus they are sitting comfortably above the linear regression line in the scatterplot.

- Team PER sometimes fails to accurately capture a team’s strength when their best players don’t play enough minutes. BOS (Boston Celtics), for example, does not have a high team PER because they seldom let their best players (namely their big three – Kevin Garnett, Paul Pierce, and Ray Allen) log a lot of playing time unless they are in a close game till the end. This, in a way, will distort the team PER and undervalue those teams. Players with high PER have less weight when calculating team PER simply because they don’t have to play that many minutes for their team to win games. As we can see, BOS indeed is an anomaly this season with relatively low PER but a high win ratio.

Second, I would like to point out some other situations in which the model can potentially fail to work well:

I. When teams make major changes in their roster during the season. This can happen under a number of circumstances. For instance, some teams might make major trades during the season where they might send some of their best players to other teams in exchange of players, draft picks, and/or cash. Trades would have minimal impact on the model when one team trades its players for players on the other team. This way, even though the roster changes, new players coming to the team will still make contribution to the team's PER as they start to play for the team. If, however, the trade involves draft picks and cash, the model would be negatively impacted because it does not take into account non-player factors. In reality, most trades do have non-player factors because no two players are exactly comparable and further complications about their contracts and teams' management also have significant influence in these situations.

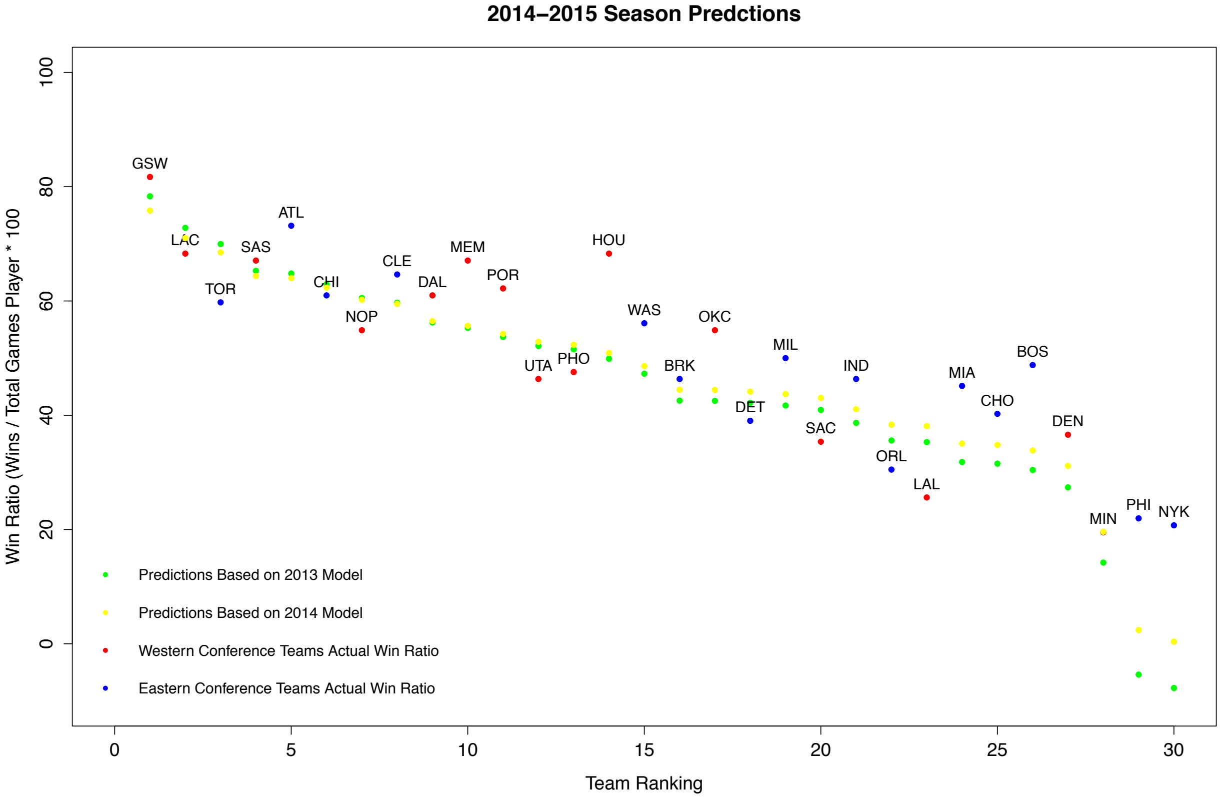
In addition to trades, injuries can also play a huge role in determining teams' fate and cause teams to make major changes in their rosters. If a team loses its most capable player due to injury for a certain number of games, it is much more likely for the team to lose those games than if there were no injuries on their best player. However, if the player recovered quickly enough, the total minutes he played over the season might not be that distinguishable from his teammates as well as opponents. That way, the injury the player has would have more impact on the team's win ratio than on the team PER, and this discrepancy could potentially cause the model to fail and create an anomaly in the output.

II. Another drawback of the model is that it might not work as expected when two teams have very close win ratios. Since the model is linear, when predicting teams' win ratios, it will yield a deterministic order of two teams with unequal team PERs. However, as we can see from the results, when two teams have close team PERs, their win ratios do not necessarily correspond to the relative value of their PERs. This is understandable because when two teams are relatively close in terms of the strength of their players, the number of games they can win depends more on other factors than the players themselves. For instance, in the last few seasons, western conference is much more competitive than the eastern conference. Because teams play against teams in the same conference much more than teams in the other conference, it is much harder for teams in the western conference to win the same number of games as teams in the eastern conference. Therefore, teams with comparable PER value might differ a lot in their win ratio due to other factors such as the conference to which they belong. This reasoning is actually visually observable in the scatterplots where western conference teams normally lie under the regression line (i.e. underperforming high PER teams) whereas eastern conference teams generally lie above the regression line (i.e. overachieving low PER teams). The model, therefore, may not be able to predict the exact standings of the teams, but should do a decent job in predicting which teams are going to make the playoffs (namely the first 8 teams in both conferences by win ratio).

III. Although the model proves to do a decent job in predicting team's performance based on players' stats, it is not possible to get rid of the luck factor involved in the game entirely. In an 82 game season, anything can happen. Sometimes the winning and losing between two teams do not necessarily depend on the relative strengths of their players. The league's worst team can surprise the best team and steal a game or two when they face each other on the court if the best team is feeling under weather, or more likely, if the best team is resting its players in preparation for the upcoming playoffs towards the end of the season. It is hard to factor the luck factor into the model because it largely depends on the match-up and other factors that do not show up tangibly in players' stats upon which our model is based. A possible mitigation is to add an additional term in the linear model to capture the luck (or match-up) factor.

§ Predictions and Analysis

After building and analyzing the model, I used the linear models from the past two seasons to predict 2014-2015 regular season results of the 30 teams currently in the league based on the players' stats (their individual PER) at the beginning of the season.



The red and blue dots in the scatterplot represent the actual win ratios of each team at the end of 2014-2015 season. Yellow dots are predicted win ratios based on the model from 2012-2013 season while the green dots are predicted win ratios based on the model from 2013-2014 season. We can see that there still is considerable variance in regards to the predicted value and actual value of teams' win ratios. Nonetheless, if we were to predict which teams would make the playoffs (8 teams from each conference with highest win ratios), we would only be off by two teams. To be exact, BOS (Boston Celtics) from eastern conference should make the playoffs instead of DET (Detroit Pistons), and HOU (Houston Rockets) from western conference should have a place in the top 8 instead of UTA (Utah Jazz).

The actual playoff teams from both conferences and my predictions according to the model are listed below:

Actual Playoff Teams From the Western Conference	**Predicted Playoff Teams From the Western Conference**	**Actual Playoff Teams From the Eastern Conference**	**Predicted Playoff Teams From the Eastern Conference**
Golden State Warriors	Golden State Warriors	Atlanta Hawks	Toronto Raptors
Houston Rockets	Los Angeles Clippers	Cleveland Cavaliers	Atlanta Hawks
Los Angeles Clippers	San Antonio Spurs	Chicago Bulls	Chicago Bulls
Portland Trial Blazers	Dallas Mavericks	Toronto Raptors	Cleveland Cavaliers
Memphis Grizzlies	Memphis Grizzlies	Washington Wizards	Washington Wizards
San Antonio Spurs	Portland Trial Blazers	Milwaukee Bucks	Brooklyn Nets
Dallas Mavericks	New Orleans Pelicans	Boston Celtics	Milwaukee Bucks
New Orleans Pelicans	Utah Jazz	Brooklyn Nets	Detroit Pistons

The reason why predicted win ratio of the Rockets is off by a huge margin is primarily because the influence that James Harden, a MVP candidate and the league's No.2 Scorer, has on the team is diluted by his less talented teammates. Harden single-handedly helped the Rockets won a number of close games by making numerous clutch shots and game winners. The team PER, as currently defined, is not particularly good at rewarding players/teams of winning close games. Consequently, even though Harden has been brilliant throughout the season, the lack of help from his teammates (due to multiple injuries suffered by other key players on the team such as all star center Dwight Howard and promising young power forward Terrence Jones) drags down the team PER, and thus the model does not accurately predict their win ratio. Also, this season's team win ratios, especially that of the 2nd – 6th seeds of the western conference, differed only by a very thin margin. Precisely speaking, the Rockets have exactly the same win ratio as the Clippers, and have won only 1 more game than Grizzlies and Spurs, but they hold the tiebreakers against the other teams, thus sitting as the No. 2 seed. In that sense, the model is not severely undermined even though it leaves out the second highest seeded team in the competitive conference from the playoff picture.

As for the Celtics, their win ratio is underestimated by the model precisely because of one of the pitfalls of the model that I discussed in the previous section. The Celtics are a team without star players, meaning they don't really have any individual with extraordinarily high PER. Nevertheless, they play very well as a team. Since the model is in a way more biased towards players' individual strengths, teams like the Celtics are negatively impacted and misjudged. A possible direction of further research into this topic is to find a solution to address this issue, devising a method (possibly adding an extra term or creating a team-efficiency multiplier, etc.) that could capture how much better/worse players are collectively.

§ Conclusion and Extensions

It might seem that the model above is common sense – that it is logical to predict a team's performance from the individual player's capability. However, very few basketball fans and pundits are actually leveraging the players' statistics to describe a team's abilities. Even when they do, they tend to over complicate things by using obscure metrics such as shooting efficiencies, rebounding opportunities, defensive impact, etc., which are almost certainly biased towards one certain type of players/teams or another. They also tend to exploit the abundance of data by utilizing insignificant historical statistics to predict a team's future performance. Too often we can hear analysts from ESPN or TNT refer to statistics like "some team has won 80% of the games they played in March for the past 10 years" or "some player shoots 20% more accurately the month following the all-star break". These statistics mean little more than historical coincident. It does not make logical or statistical sense to use what happens to a team/player in the past decade to predict what is going to happen because the game has surely changed on a lot of dimensions. The significance of the high correlation coefficients and statistical summary will hopefully convince a portion of the basketball masses that there are simple basketball statistics out there that are very useful and powerful in evaluating a player and his team.

Upon further research, I came across another metric cleverly developed by a basketball expert that can attribute the total number of games that a particular team has won to every player on that team. It is an extension of the system that *Bill James* developed to attribute baseball's wins to individual players. The metric is called Win Share[xi], and is comprised of Offensive Win Share and Defensive Win Share. Offensive Win Shares are credited to players based on *Dean Oliver*'s points produced and offensive possessions and are calculated using the following algorithm:

1. Calculate points produced for each player. In 2008-09, James had an estimated 2345.9 points produced.
2. Calculate offensive possessions for each player.
3. Calculate marginal offense for each player. Marginal offense is equal to (points produced) - 0.92 * (league points per possession) * (offensive possessions).
4. Calculate marginal points per win. Marginal points per win reduces to 0.32 * (league points per game) * ((team pace) / (league pace)).
5. Credit Offensive Win Shares to the players.

Similarly, we credit defensive win shares to players using the algorithm described as follows:

6. Calculate the Defensive Rating for each player.
7. Calculate marginal defense for each player. Marginal defense is equal to (player minutes played / team minutes played) * (team defensive possessions) * (1.08 * (league points per possession) - ((Defensive Rating) / 100)).
8. Calculate marginal points per win. Marginal points per win reduces to 0.32 * (league points per game) * ((team pace) / (league pace)).
9. Credit Defensive Win Shares to the players. Defensive Win Shares are credited using the following formula: (marginal defense) / (marginal points per win).

Then we put it all together by adding offensive and defensive win shares of a player to get the total number of wins that one specific player contributes to the team over a single season.

I tested the accuracy of the metric by correlating players' win shares with their teams' win ratios for the past few seasons, and found that this is an extremely effective metric that produces an R^2 value of over 0.9. However, because win share only attributes the number of games that a particular team has already won to individual players, it is ungrounded to use the win share of previous seasons to predict next season's result. However, win share definitely is a good metric to look at when we want to gauge players' actual contribution to their respective teams once the games have been played out.

Another direction where further research on this topic could pursue is to add additional inputs to the model. Currently, the model is restricted to one input (namely, team PER). One benefit of this is that the results of the model can be easily graphed and interpreted. However, there might, and surely will, be other variables that can be used to connect individual players' stats and their teams' performance. In that case, adding new variables to the model and use multi-regression methods might give us more insights into the linkage between players' capabilities and teams' success.

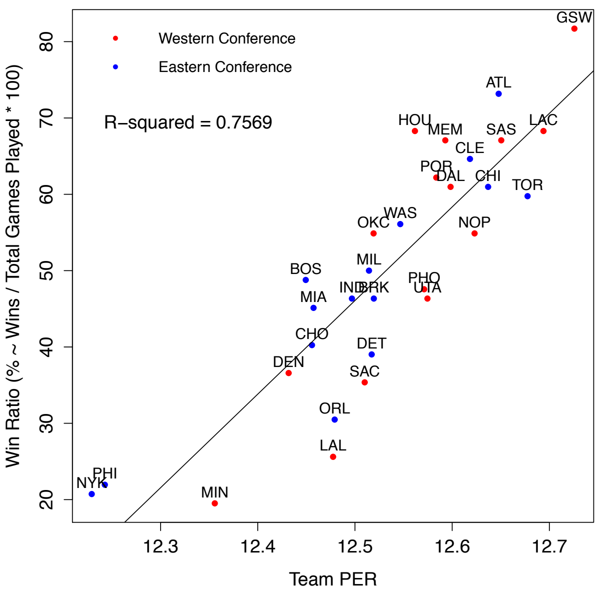
APPENDIX

§ Glossary

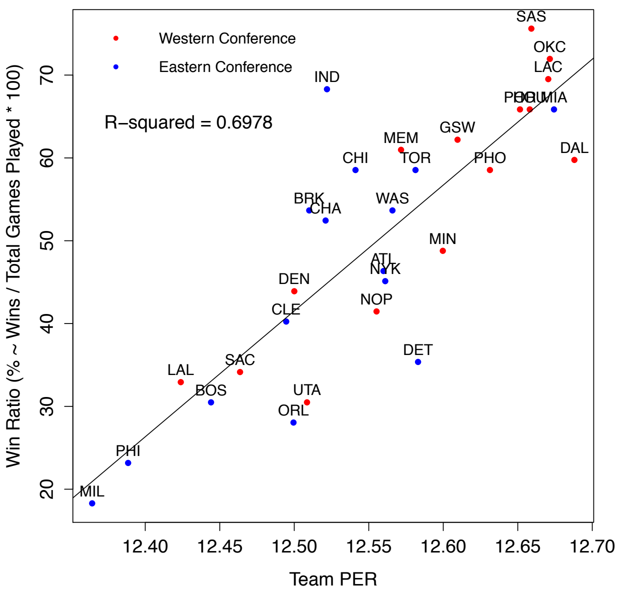
2P: 2-Point Field Goals
2P%: 2-Point Field Goal Percentage; the formula is 2P / 2PA.
2PA: 2-Point Field Goal Attempts
3P: 3-Point Field Goals (available since the 1979-80 season in the NBA)
3P%: 3-Point Field Goal Percentage (available since the 1979-80 season in the NBA); the formula is 3P / 3PA.
3PA: 3-Point Field Goal Attempts (available since the 1979-80 season in the NBA)
AST: Assists
BLK: Blocks (available since the 1973-74 season in the NBA)
DRB: Defensive Rebounds (available since the 1973-74 season in the NBA)
FG: Field Goals (includes both 2-point field goals and 3-point field goals)
FGA: Field Goal Attempts (includes both 2-point field goal attempts and 3-point field goal attempts)
FT: Free Throws
FTA: Free Throw Attempts
MP: Minutes Played (available since the 1951-52 season)
ORB: Offensive Rebounds (available since the 1973-74 season in the NBA)
PF: Personal Fouls
PTS: Points
STL: Steals (available since the 1973-74 season in the NBA)
TOV: Turnovers (available since the 1977-78 season in the NBA)
TRB: Total Rebounds (available since the 1950-51 season)

§ Scatterplots of Team Win Ratio versus Team PER for Past 8 Seasons

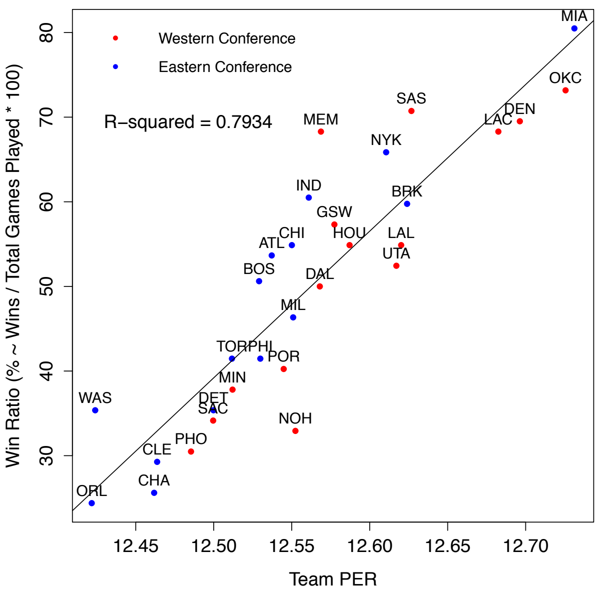
Season 2014 – 2015



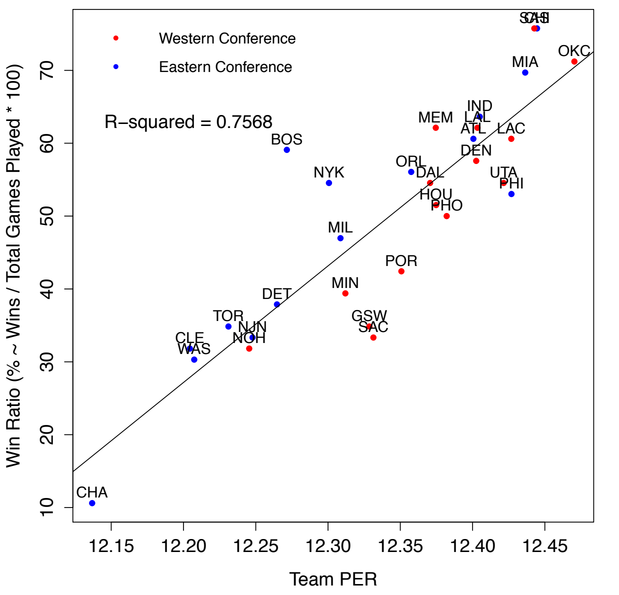
Season 2013 – 2014



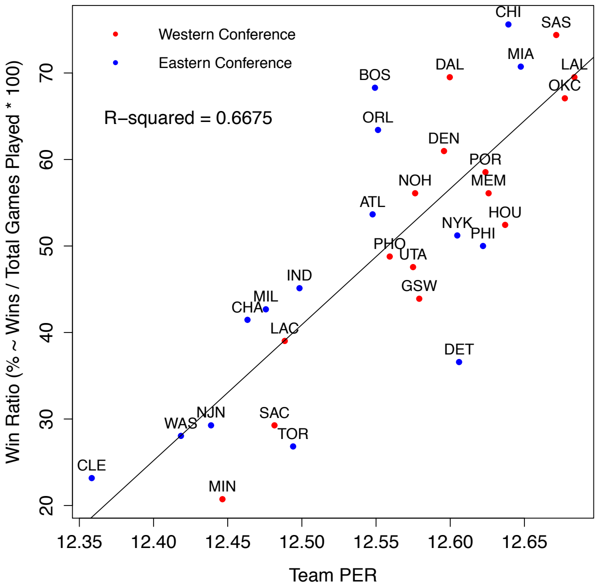
Season 2012 – 2013



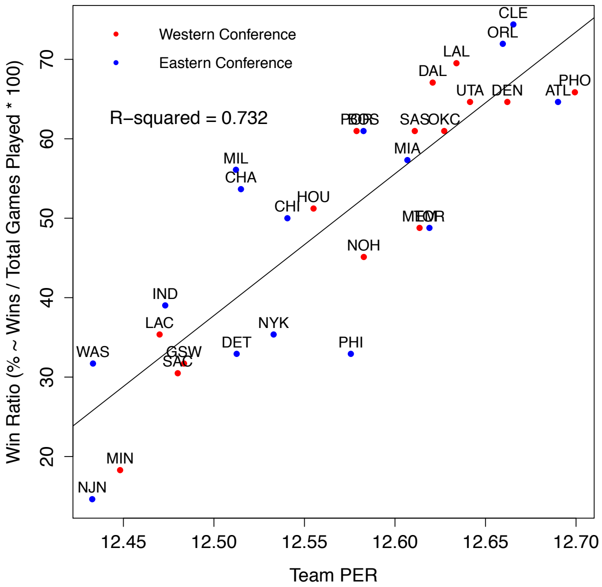
Season 2011 – 2012



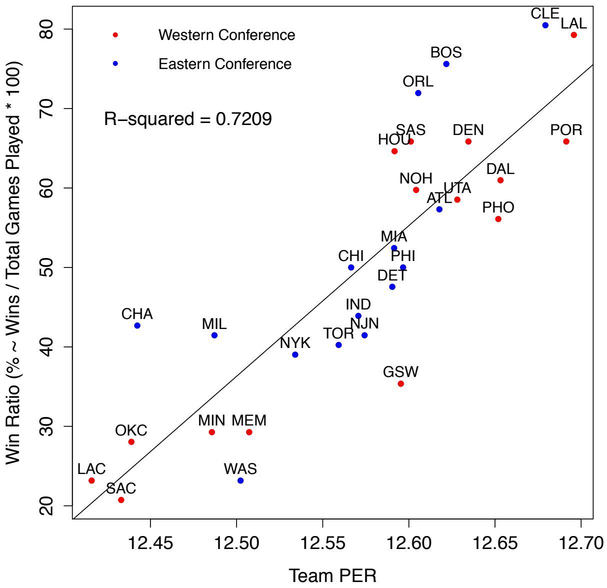
Season 2010 – 2011



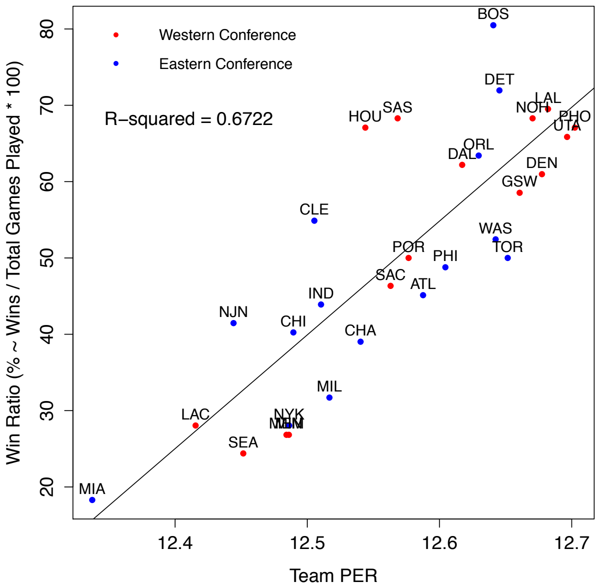
Season 2009 – 2010



Season 2008 – 2009



Season 2007 – 2008



REFERENCES

[i] JAMES, BILL (1985): *The Bill James Historical Abstract*. Villard Books.

[ii] ALBERT, JIM, AND BENNETT, JAY (2003): *Curve Ball*. Copernicus Books

[iii] SCHWARZ, ALAN (2004): *The Numbers Game: Baseball's Lifelong Fascination with Statistics*. Thomas Dunne Books.

[iv] SEVERINI, THOMAS: *Analytic Methods in Sports*. CRC Press

[v] SIMMONS, RUSTY: "*Golden State Warriors at the Forefront of NBA Data Analysis*."

[vi] HOLLINGER, JOHN: "*What is PER?"*

[vii] FIXLER, KEVIN: "*Disappearance of the Traditional NBA Big Man*."

[viii]ESPN.COM: *NBA Statistics*.

[ix] BASKETBALL-REFERENCE.COM: *NBA League Index*.

[x]FEIN, ZACH: *Calculating PER*.

[xi] KUBATKO, JUSTIN: *Calculating Win Shares*.