International Journal of Intelligent Information and Management Science
ISSN: 2307-0692, Volume 8, Issue 1, February, 2019

The Effect of Points per Game on the Number of Wins in NBA

Xinyi Shang
College of Liberal Arts and Sciences, University of Illinois at Urbana-Champaign, Champaign 61820, Illinois, USA

Abstract: In the National Basketball Association (NBA), it contains a seasonal league matches. In a construction of a perfect team, not only coaches and players, the work of collecting statistic of each match also plays an important role. In this study, I pick up some of data which I think impact much on the number of wins and do a statistic analysis. The data are provided by Sports Reference, a professional, easy-to-get and up-to-date website, which contains a branch of basketball. In the dataset, I choose six independent variables which may have strong influence on the dependent variable. The results of my regression model verify my suspicion.

Keywords: Basketball; NBA wins; Multiple linear regression; Normal distribution

1. Introduction

In the National Basketball Association (NBA), it contains a seasonal league matches. In a construction of a perfect team, not only coaches and players, the work of collecting statistic of each match also plays an important role. By analyzing the data in games, we can see the performance of each player more directly. What's more, the team is able to figure out which factor affects most and improve the specific one purposely. This method which has been used for several years, recording and analyzing, did have a great help.

In this study, I pick up some of data which I think impact much on the number of wins and do a statistic analysis. The data are provided by Sports Reference, a professional, easy-to-get and up-to-date website, which contains a branch of basketball. Using the data of last two seasons with a sum of sixty NBA teams, I run a regression model of possible factors against the number of wins to test which one is the most influential element in competitions. To find out the determinants of wins, I collect examples of 60 teams corresponding to 6 independent variables through the last 2015~2016 and 2016~2017 seasons. With the information provided, I'm able to run a multiple regression to see which independent variable has the most effect on the wins from the overall situation. The statistics can support a scientific and reliable analysis for my prediction.

From an effective multiple regression model, it satisfies all the potential assumptions. The error terms of data are normally distributed since the histogram of standard residuals is bell- shaped and the means are close to zero. After applying correlation table and reviewing the relationship of independent variables, I don't find anyone larger than 0.8, so the data has no serious multicollinearity. Moreover, with the scatter plot of residuals against predicted wins, the data has no serious outliers, no autocorrelation and almost constant variance of errors, that is homoscedasticity. Because at first the scatter plot is not so perfect, I also perform the other four transformations on the number of wins, but these are even not so good as the origin one. The original model can be considered as homoscedasticity. It's no need for me to add a trend variable with such a successful multiple regression model.

The results of my regression model verify my suspicion that the number of wins is highly affected by the points per game, in a positive relationship. Not only the highest one, the other independent variables such as opponent points per game, 3 points field goals per game, 2 points field goals per game and free throws per game are also correlated to the dependent variable, the number of wins. These independent variables are all important determinants in winning games.

2. Data Description

In this study, the data I collect are from the basketball branch of Sports Reference. It clearly lists all the statistics I need and provide them divide by different teams and every seasons. As my previous knowledge of basketball, I choose the number of wins as dependent variable. In order to test the most powerful factor, I shorten the independent variables to 6 from such a large database. For the initial full model, the 6 independent variables I select are points per game (pts/game), opponent points per game (oppt/game), 3 points field goals per game (3ptfg/game), 2 points field goals per game (2ptfg/game), free throws per game (frthr/game) and personal fouls per game (perfoul/game). These dependent and independent variables will help me explore which element is the most significant one.

In Figure 1 below, I create scatter plots of individual independent variables to make sure the data has no curvilinear relationships or serious outliers. After it, in Figure 2

and 3, I apply histogram and scatter plots of residuals against predicted number of wins. Through the graphs, we can directly see that there is no heteroscedasticity or non-normality.

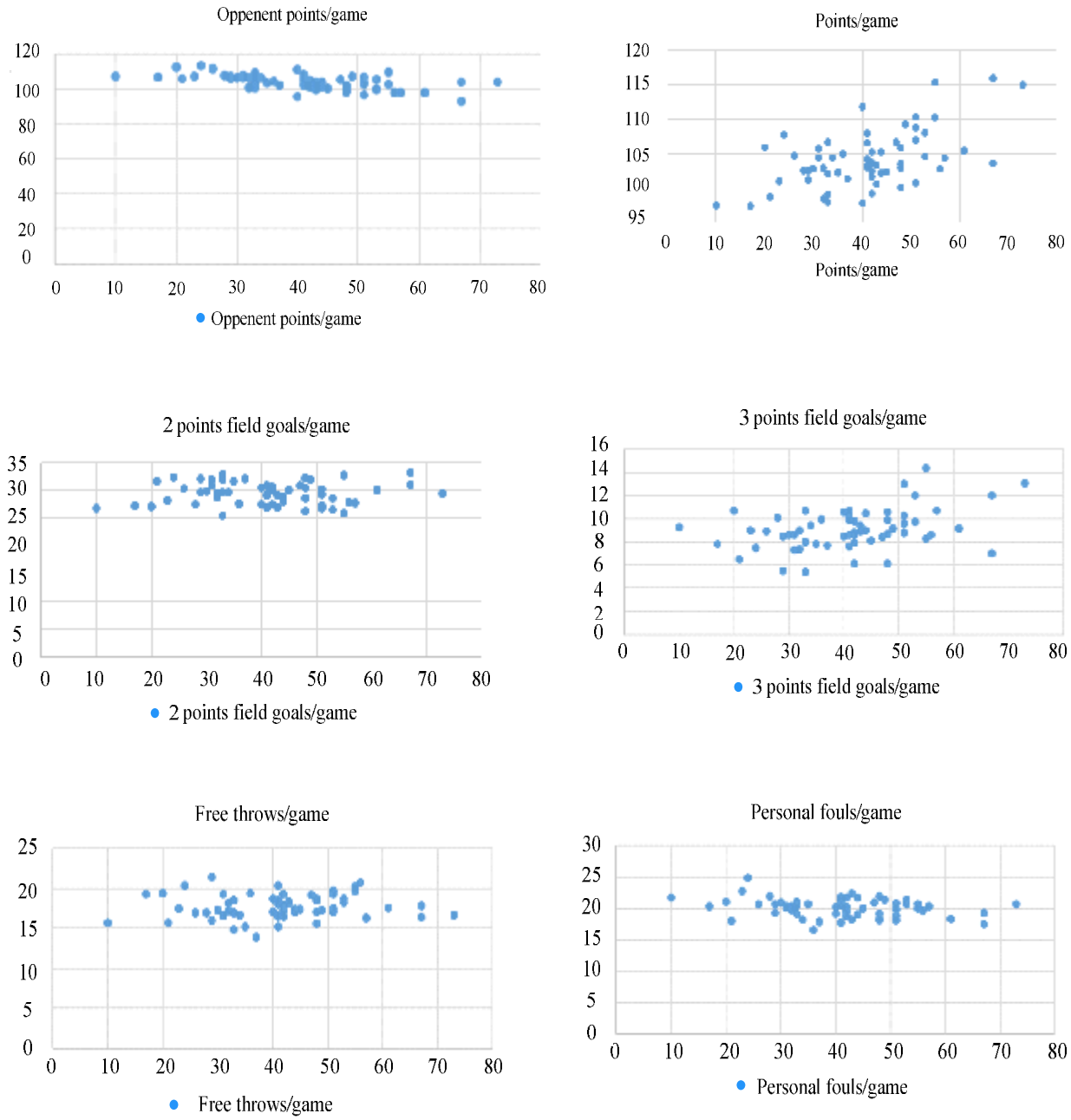

Figure 1. Scatter plots of the number of wins vs independent variables

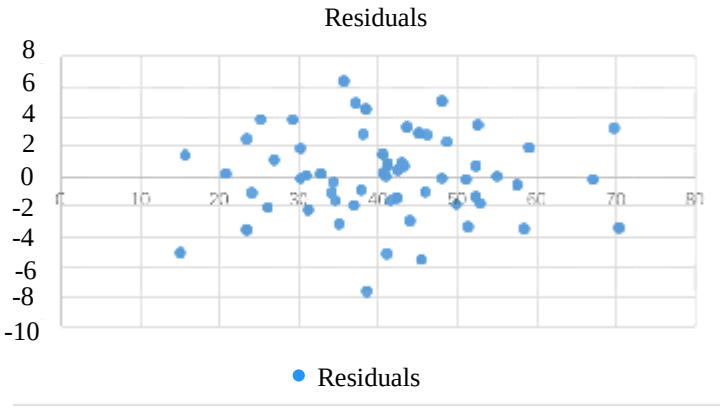

Figure 2. Histogram of standardized residual- Initial model

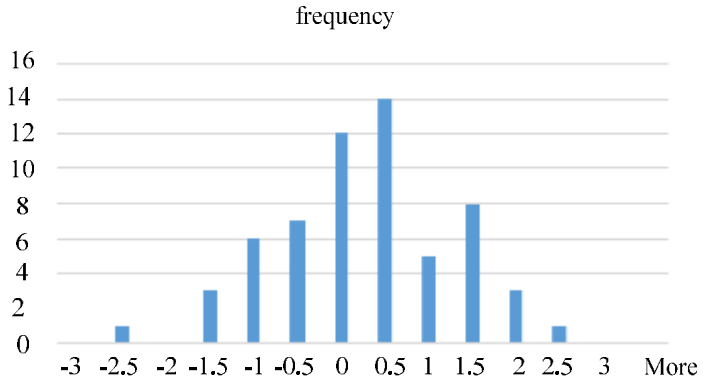

Figure 3. Scatter plots of residuals vs predicted wins- Initial model

In addition, Table 1 states the descriptive statistics of all variables. The data contains 60 teams over the past two seasons without any omission. Table 2, the correlation chart, is used to test if there is any relationship between independent variables. Since the multicollinearity is defined as the absolute value of correlation coefficient equals or larger than 0.8, there is no serious multicollinearity based on the chart. In the case that all charts and graphs conforms the data to five assumptions, it concludes that the database is normal for me to continue on more development.

Table 1. Descriptive statistics

Wins		**Points/game**	**Opponent points/game**		
Mean	41	Mean	104.12667	Mean	104.13
Standard Error	1.6138525	Standard Error	0.5418	Standard Error	0.5298241
Median	41.5	Median	103.45	Median	104.3
Mode	42	Mode	104.3	Mode	102.6
Stanard Deviation	12.500847	Stanard Deviation	4.196765	Stanard Deviation	4.1039997
Sample Variance	156.27119	Sample Variance	17.612836	Sample Variance	16.842814
Kurtosis	0.2220902	Kurtosis	0.956723	Kurtosis	0.1911912
Skewness	0.0927875	Skewness	0.8449856	Skewness	-0.192816
Range	63	Range	18.6	Range	20.4
Minimum	10	Minimum	97.3	Minimum	92.9
Maximum	73	Maximum	115.9	Maximum	113.3
Sum	2460	Sum	6247.6	Sum	6247.8
Count	60	Count	60	Count	60
3 points field	**2 points field goals/game**			**Free**	**Personal fouls/games**

goals/game				throws/game			
Mean	9.0783333	Mean	29.555	Mean	17.761667	Mean	20.075
Standard Error	0.226033	Standard Error	0.2565132	Standard Error	0.2000719	Standard Error	0.1899394
Median	8.9	Median	29.7	Median	17.55	Median	20.3
Mode	8.6	Mode	29.2	Mode	19.3	Mode	20.7
Stanard Deviation	1.750844	Stanard Deviation	1.9869425	Stanard Deviation	1.5497503	Stanard Deviation	1.4712643
Sample Variance	3.0654548	Sample Variance	3.9479407	Sample Variance	2.401726	Sample Variance	2.1646186
Kurtosis	1.0517244	Kurtosis	-0.938786	Kurtosis	-0.086563	Kurtosis	0.7857013
Skewness	0.5066738	Skewness	-0.063592	Skewness	0.0355192	Skewness	0.1951621
Range	9	Range	7.7	Range	7.5	Range	8.2
Minimum	5.4	Minimum	25.5	Minimum	13.9	Minimum	16.6
Maximum	14.4	Maximum	33.2	Maximum	21.4	Maximum	24.8
Sum	544.7	Sum	1773.3	Sum	1065.7	Sum	1204.5
Count	60	Count	60	Count	60	Count	60

Table 2. Correlations

	Wins	Points/game	Opponent points/game	3 points field goals/game	2 points field goals/game	Free throws/game	Personal fouls/game
Wins	1						
Points/game	0.55341541	1					
Opponent points/game	-0.5502981	0.357012746	1				
3 points field goals/game	0.39447548	0.632754017	0.18669756	1			
2 points field goals/game	0.03166221	0.077201522	0.026565635	-0.671171684	1		
Free throws/game	0.08118855	0.365805337	0.260969804	0.038604594	-0.061667262	1	
Personal fouls/game	-0.2634705	0.062421342	0.382727075	-0.019097744	-0.021089923	0.280262964	1

3. Regression Analysis

As I suggest, the selected independent variables are related to the number of wins. From my personal point of view, the six factors will directly influence the dependent variable. My initial regression model is as follows:

$$Numwins = b_0 + b_1 pts/game + b_2 oppt/game + b_3 3ptfg/game + b_4 2ptfg/game + b_5 frthr/game + b_6 perfoul/game + e$$

These elements are all worthy considerations due to the potential relationship they may have with the dependent variable. When we discuss about the number of wins, we have a large possibility to think about these factors.

As showed in Table 3-5, the initial model seems successful. It appears with an r-squared of 0.9502 and an F-stat of 168.4115. Due to the prerequisite of rejecting the null hypothesis, the error should be at the 5% level, that is, the p-value has to be smaller than 0.05. Although the whole model has been proven successfully, there are still some variables statistically insignificant. The 3 points field goals per game, 2 points field goals per game, free throws per game and personal fouls per game all have p-values larger than 0.05, especially the personal fouls per game is over 0.15. After considering that there is no serious multicollinearity, proven by correlation chart (Table 2), I decide to drop the variable personal fouls per game because of the highest p-value it produces.

Additionally, I make Figure 1, 2 and 3, individual variables scatter plots, a histogram of residuals and a scatter plot of residuals vs predicted number of wins, to ensure that there is no non-normality of errors, no heteroscedasticity, non-autocorrelation or no unnecessary outliers. The assumptions are all satisfied. Figure 4 and 5 can also prove for the reduced model later.

After removing one insignificant variable I choose, I create the reduced regression model, seeming to be valid. The r-squared and adjusted r-squared has slightly changed, but not so much, almost same as before, within a variation about 94.5%. The other remaining independent variables becomes to have lower p-values. They are at least below 15% and two of them are lower than 5%. What's more, it's necessary for me to run a partial F-test. The result shows that the test-statistic is 0.2028 with a p-value of 0.6543, which is smaller than 0.8239. It doesn't allow me to reject the null hypothesis so I decide to use the reduced model. The reduced regression is effective as shown in Table 6-8.

The final regression model is as follows:

$$Numwins = b_0 + b_1 pts/game + b_2 oppt/game + b_3 3ptfg/game + b_4 2ptfg/game + b_5 frthr/game + e$$

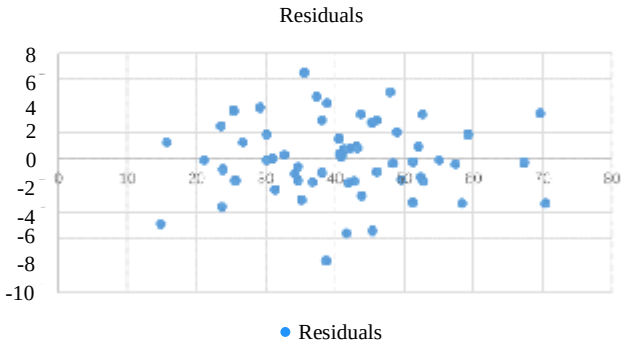

Figure 4. Histogram of standardized residual- Reduced model

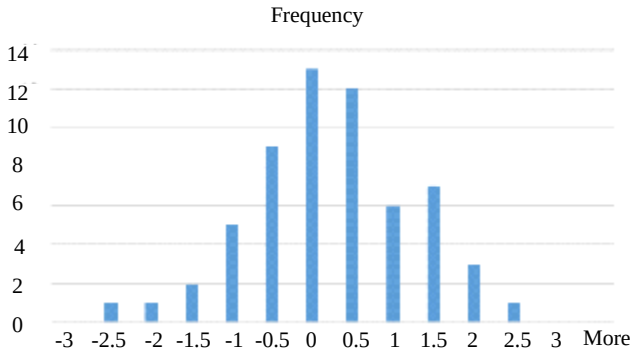

Figure 5. Scatter plots of residuals vs predicted wins-Reduced model

Table 3. Initial model regression results-Summary output

Regression Statistics	
Multiple R	0.9747631
R Square	0.9501631
Adjusted R Square	0.9445212
Standard Error	2.9444421
Observations	60

Table 4. Initial model regression results-Anova

	df	**SS**	**MS**	**F**	**Significance F**
Regression	6	8760.5038	1460.084	168.41152	1.085E-32
Residual	53	459.49619	8.6697395		
Total	59	9220			

Table 5. Initial model regression results

	Coefficients	**Standard Error**	**t stat**	**P-value**	**Lower 95%**	**Upper 95%**
Intercept	46.212552	12.937748	3.5719162	0.000764	20.262716	72.162388
Points/game	7.7052294	3.2204254	2.3926132	0.0203055	1.245874	14.164585
Opponent points/game	-2.628655	0.1086026	-24.20434	2.247E-30	-2.846484	-2.410826
3 points field goals/game	-15.33062	9.6327178	-1.591515	0.1174413	-34.6514	3.9901699
2 points field goals/game	-10.23447	6.3819721	-1.603653	0.114734	-23.03508	2.5661471

Free throws/game	-5.336839	3.281091	-1.626544	0.1097658	-11.91787	1.2441962
Prsonal fouls/game	0.1313436	0.2916721	0.4503124	0.6543219	-0.453677	0.71636338

Table 6. Initial model regression results-Summary output

Regression Statistics	
Multiple R	0.9746653
R Square	0.9499724
Adjusted R Square	0.9453402
Standard Error	2.9226265
Observations	60

Table 7. Initial model regression results-Anova

	df	**SS**	**MS**	**F**	**Significance F**
Regression	5	8758.7457	1751.7491	205.08094	7.93E-34
Residual	54	461.25425	8.5417454		
Total	59	9220			

Table 8. Initial model regression results

	Coefficients	**Standard Error**	**t stat**	**P-value**	**Lower 95%**	**Upper 95%**
Intercept	47.239197	12.640933	3.7370025	0.0004509	21.895653	72.58274
Points/game	7.7252054	3.1962617	2.4169502	0.0190591	1.3170865	14.133324
Opponent points/game	-2.610924	0.1004632	-25.98886	3.342E-32	-2.81234	-2.409507
3 points field goals/game	-15.4124	9.5596483	-1.612235	0.1127407	-34.57834	3.7535387
2 points field goals/game	-10.28902	6.3335462	-1.624527	0.1100868	-22.98701	2.4089798
Free throws/game	-5.334681	3.2567776	-1.638024	0.1072313	-11.86413	1.1947653

4. Empirical Results

In Table 6-8, the coefficients suggest that, holding other variables constant, if the points per game increase by one, the number of wins will be increase by 7.7252. Considering about the mean value of wins is 41.5, there would present a 18.615% increase in the number of wins associating with the points per game increase. It's a positive and effective expectation.

The other four factors are in same relationships with the dependent variables. As the opponent points per game decrease, the winning games becomes more and more. Also, reviewing for the 3 points and 2 points field goals per game, the more field goals the team get, the winning number will decline in sequence. Moreover, knowing the free throws per game has the fewest number of field goals, it affects less than other two variables but similar in negative relationship.

Through this study, I am surprised to find out that the 3 points field goals per game is the most influence determinant of the number of wins, then 2 points field goals per game, especially in a negative relationship. As a common sense, the more 3 points field goals a team gets in a match, the more points they will have. The statistics show exactly different conclusion as I expected before. In addition, with no surprise, more points per game lead to more success. The efforts on points per game will contribute to a good result.

5. Summary and Discussion

This study investigates the determinants of the number of wins in NBA. Specifically, I am interested in analyzing the effect of points per game on the winning number. Through the analysis, I test if the relationship between these has a theoretical empirically prediction. As a result, I'm able to conclude that points per game is one of the most significant factors of the number of wins, with an obvious positive relationship. Additionally, 3 and 2 points field goals per game and free throws per game also have much influence on the dependent variable, even more than the points per game, but in an inverse relationship.

One shortfall of this study is that I only focus on the matches from all over the two seasons, the sum of winning games from start to finish. However, a season of league match should be divided by two parts, preseasons and playoffs. As we all know, some teams will use strategy to reserve capability or avoid strong competitors. They sometimes exert full power for victory selectively so the data may not show the ability of each team entirely.

Future study should focus on the number of winning playoffs if the data is possible.

Another shortfall is that although the data is collected for NBA, the matches of American clubs, the league allows clubs to introduce foreign athletes. The foreign players also play important role in matches. With this missing information, we may consider the number of foreign player of each team as an independent variable in future study.

References

[1] Basketball Statistics and History. Retrieved from https://www.basketball- reference.com/

[2] Masaru T., Cross C. L., Rieger R. H., Maak T. G., Willick S. E. Predictive validity of national basketball association draft combine on future performance. Journal of Strength & Conditioning Research. 2018, 32, 396–408.

[3] Gonzalez A. M., Hoffman J. R., Rogowski J. P., Burgos W., Manalo E., Weise K., Stout J. R. Performance changes in nba basketball players vary in starters vs. nonstarters over a competitive season. Journal of Strength & Conditioning Research. 2013, 27, 611–615.

[4] Herring C. The count. Wall Street Journal - Eastern Edition. 2013, D5.