The defense of the three-point shot

Louis Nass

Department of Mathematics

Tulane University

New Orleans, LA 70118

December 17, 2017

Abstract

I designed a linear model of NBA wins where the predictors are two-point makes, three-point makes, two-point makes allowed, and three-point makes allowed. I will consider this model for three data sets during three consecutive seasons from different decades. They are 1985-88, 1995-98, and 2005-08. First, I will run hypothesis tests to see if various combinations of the predictors are arbitrary. Next, I will investigate the correlation matrices from each model in order to comment on the "style of play" that each sample represents. Finally, I will run a residual analysis of each of the models. Specifically, we will comment on the 'Residual vs Leverage' plots to see if the teams that impact the model do indeed correspond with the trends from the correlation matrix.

Key Words: Linear model, hypothesis test, leverage, predictors, correlation, NBA

1 Introduction

It is unquestionable that the three-point line has been a factor in NBA games since its introduction in the 1979-80 season. However, it has been suggested by numerous sports reporters that the game now is solely focused on the three-point shot, and as a result, the game has become offensively dominated. The quote "NBA players don't play defense" has been tossed around, in my opinion, unfairly. To do them justice I want to evaluate the importance of the two-point shot versus the three-point shot, as well as the difference between offense and defense by modeling by the only metric that counts, wins. Luckily, statistical analysis in sports is at an all time high, particularly in basketball.

This data set comes from the website Kaggle[1]. Within this data set, Kaggle records total wins, losses, baskets made and allowed, rebounds, assists and other significant statistics recorded in basketball. For our sake we will limit our model to shots made and allowed. Kaggle measures the number of shots made and allowed by compiling the statistics from each team, from each game and summing them for a season total. Here is a sample from the 2005 season:

"name"	"o_2pm"	"o_3pm"	"d_2pm"	"d_3pm"	"wins"
"Atlanta Hawks"	2490	341	1717	236	30
"Boston Celtics"	2387	471	1754	212	24
"Chicago Bulls"	2566	480	1406	387	49

where o_2pm and o_3pm represent the number of two-point and three-point shots made respectively, and d_2pm and d_3pm represent the number of two-point makes allowed and three-point makes allowed respectively.

The design of the model will be as follows:

$$wins = \beta_0 + \beta_1(o\_2pm) + \beta_2(o\_3pm) + \beta_3(d\_2pm) + \beta_4(d\_3pm) \tag{1}$$

Where β_0 will have the units $wins$ and β_i for $i = 1, 2, 3, 4$ will have the units $\frac{wins}{shots}$. We will use the linear model function in R defined as $lm(wins \sim o\_2pm + o\_3pm + d\_2pm + d\_3pm)$. This creates a linear model object where $wins$ is our output and the other variables are our predictors.

2 Objectives

For our data set, we will use a few statistical tools in order to make commentary on the effects of our predictors. R has numerous functions and capabilities that will allow us to process

[1] `https://www.kaggle.com/open-source-sports/mens-professional-basketball#basketball_teams.csv`

the results of our model. We will use the function *anova*(), which will allow us to preform hypothesis tests. We will first define a temporary model that satisfies the hypothesis, then we will input our original and temporary models into *anova*(). The function then outputs the results of the test. Additionally, we can look at the correlation matrices using the *cor*() function in R. This will allow us to see the correlation coefficients from our predictors and our output. Lastly, we will use the *plot*() function on a linear model object. In R, this will show different residual analysis plots, and we will focus on the Residual vs Leverage plot.

2.1 Hypothesis testing

The first goal is to dispel the theories about the lack of defensive importance within the NBA. I want to reject the notion that current teams are only focused on offense and that defense is not as important. We can do this by performing the following hypothesis test:

$$H_0 : d\_2pm = d\_3pm = 0 \tag{2}$$

The hope is that we can reject this null hypothesis. This would tell us that wins are dependent on how well teams play offense and defense in tandem.

2.2 Correlation coefficient observation

Here we will observe the correlation matrices of each of our data sets. The goal here is to observe the effect each of our predictors has on each other as well as our output "wins". We aim to find the predictor that has the greatest correlation to our output. Our prediction is that two-point shots will have a greater influence than three-point shots on the older models, and vice versa for the newer models.

2.3 Residual vs leverage

Here we will observe the residual vs leverage plot. The goal here is to identify which teams had the greatest leverage on each model and if these teams correspond with our assertions that we made earlier. For example we aim to see the teams with the greatest leverage to have significant two-point shot values from the 80's model, while high leverage teams in the 2000's model will show importance for three-point shots.

3 Analysis

Within the analysis, first we will observe a few histograms of our data and our predictors. We want to make sure that none of them are skewed in any direction. If the data is skewed

then we will need to make a transformation. Next we will create the models for each of the data sets and compare at the t-values for the coefficients as well as the R^2 values. The t-values will tell us how significant the predictors are for our model, and the R^2 value tells us what percent of the data is explained by the predictors. We will then run the hypothesis test explained in section (2.1), compare the correlation coefficients as explained in section (2.2), and comment on various team's leverage as explained in section (2.3).

3.1 Pre-processing

We want our data to be approximately normal in distribution. If our predictors and result are normal then we do not need to transform the data before creating a linear model.

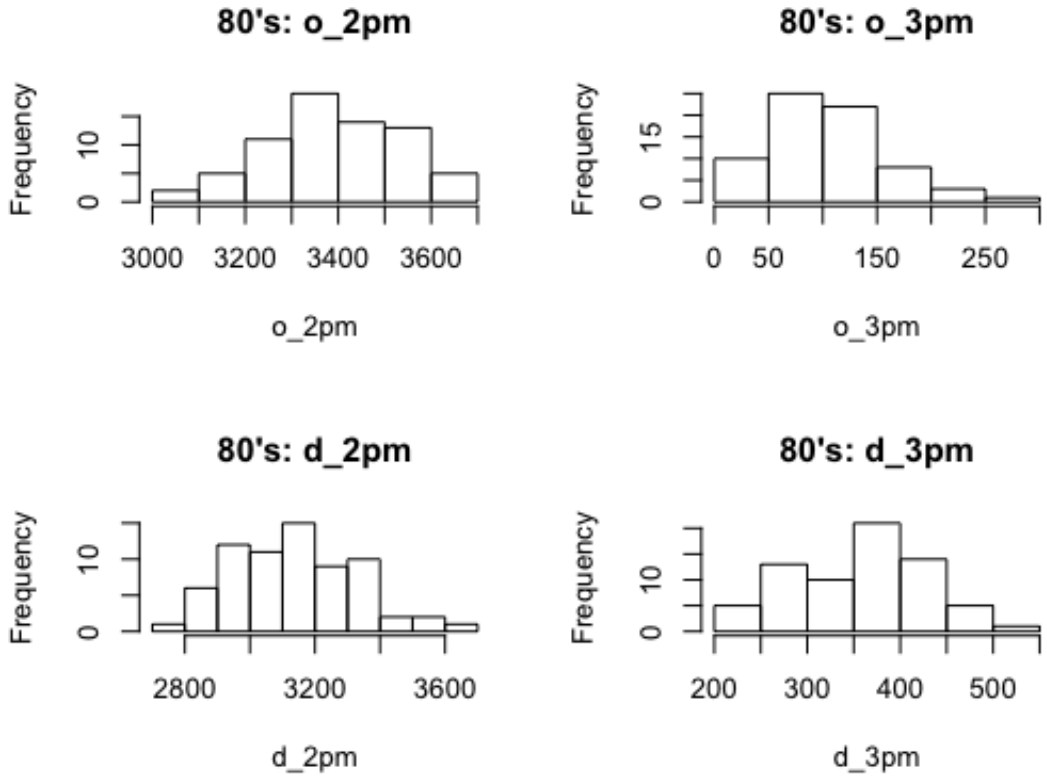

Figure 1: Histograms of predictor data from 1985-88 seasons. The histograms appear to be approximately normal.

From the figures (1) and (2) we see that indeed our predictors and outputs from the 1985-88 data set are approximately normal. Hence we will not need to transform the data before creating our model. We have similar results for the 1995-98 and 2005-08 sets as well.

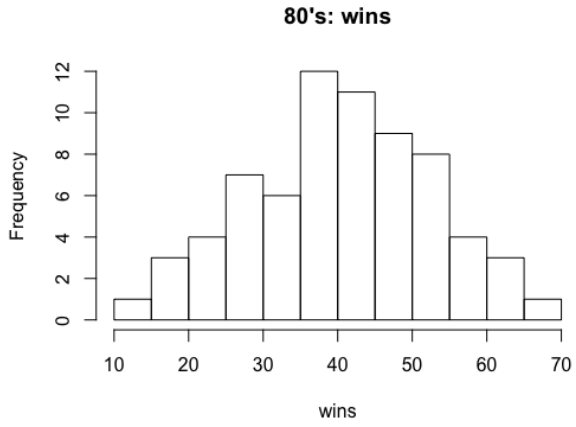

Figure 2: Histograms of wins from 1985-88 seasons. Again this is approximately normal.

3.2 Model t-values and R^2 values

Now we will compare the t-values and R^2 values of our model observed on the three data sets.

Model Comparison					
Model years	$Pr(>\|t\|)$ *o_2pm*	$Pr(>\|t\|)$ *o_3pm*	Pr(>|t|) *d_2pm*	Pr(>|t|) *d_3pm*	R^2
1985-88	$1.62e-10$	0.00186	$1.44e-06$	0.48015	0.5418
1995-98	$1.09e-13$	$1.20e-13$	$< 2e-16$	$9.26e-13$	0.7725
2005-2008	$1.00e-12$	$< 2e-16$	$2.12e-12$	$< 2e-16$	0.7423

We observe that in the 1995-98 and 2005-08 models that the coefficients for all of our predictors have highly significant values. Additionally both of these data sets have high R^2 values. The values suggest that the models represent more than 70% of the data, which is again significant. For the 1985-88 model, we see that the *o_3pm* and *d_3pm* coefficients are much less significant than the *o_2pm* and *d_2pm* coefficients. In fact, we see that the *d_3pm* coefficient is not significant at all. This further illustrates the idea that three-point shots have become more important as basketball has evolved.

3.3 Hypothesis test

Here we will evaluate the hypothesis from equation (2). Using the technique described in section (2.1), we produce the following results:

Model Comparison				
Model years	DF	Sum of Sq	F	Pr(>F)
1985-88	2	2363.7	16.116	$2.146e-06$
1995-98	2	8993	86.812	$< 2.2e-16$
2005-2008	2	5129.2	68.23	$< 2.2e-16$

We see that the F values are significant enough where we indeed reject the null hypothesis. Hence, we see that the defensive coefficients are significant for winning NBA basketball games. This shows us that defense is valued, even as three-point shots have gained popularity in more recent years.

3.4 Correlation coefficients

We now will investigate the correlation matrices of the predictors, along with the output *wins*, to see how the predictors interact with each other.

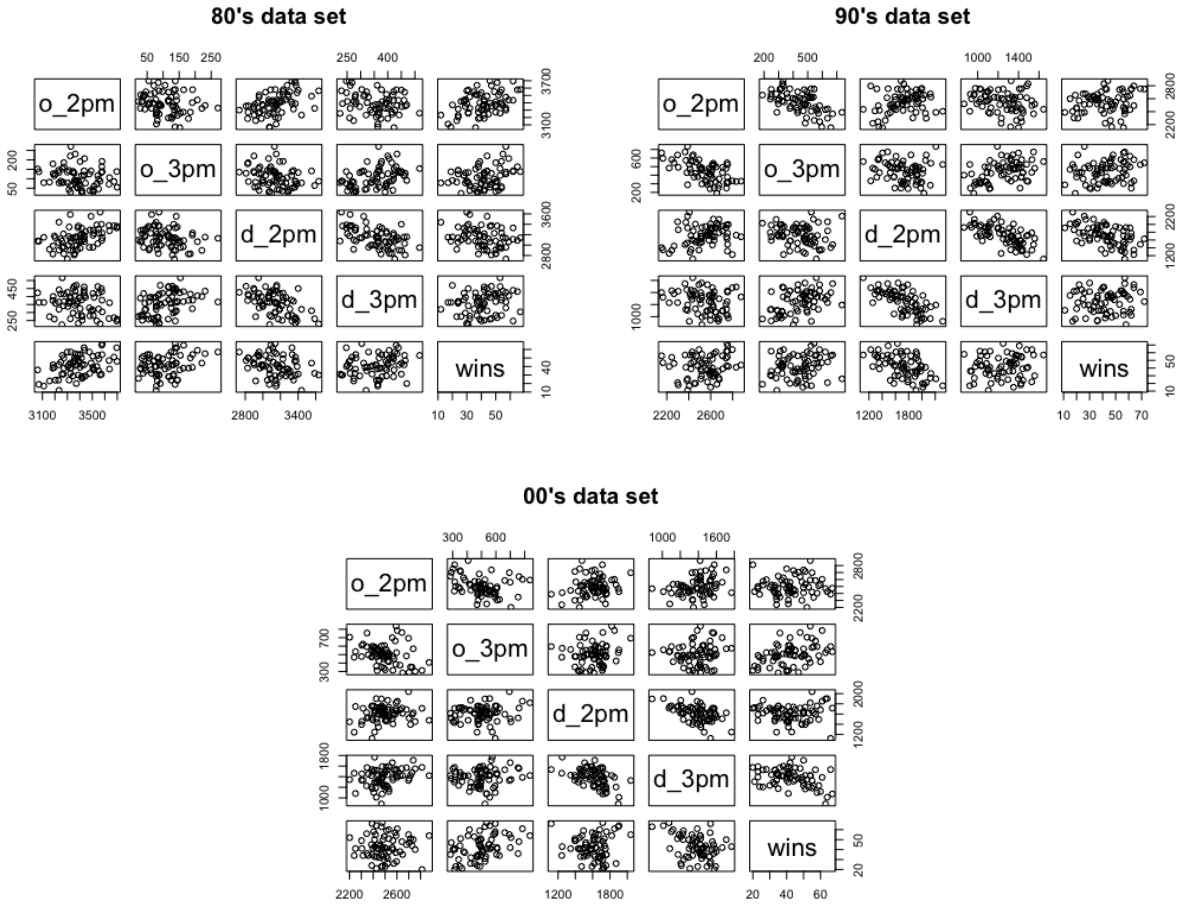

Figure 3: Pairs plots of each data set.

80's data set correlation matrix					
	o_2pm	*o_3pm*	*d_2pm*	*d_3pm*	*wins*
o_2pm	1.0000	-0.1844	0.4535	-0.2141	0.4232
o_3pm	-0.1844	1.0000	-0.2430	0.4049	0.2789
d_2pm	0.4535	-0.2430	1.0000	-0.5261	-0.2845
d_3pm	-0.2141	0.4049	-0.5261	1.0000	0.1968
wins	0.4232	0.2789	-0.2845	0.1968	1.0000

90's data set correlation matrix					
	o_2pm	*o_3pm*	*d_2pm*	*d_3pm*	*wins*
o_2pm	1.0000	-0.5719	0.2448	-0.2238	0.1742
o_3pm	-0.5719	1.0000	-0.1574	0.4438	0.1903
d_2pm	0.2448	-0.1574	1.0000	-0.6662	-0.5275
d_3pm	-0.2238	0.4438	-0.6662	1.0000	0.1178
wins	0.1742	0.1903	-0.5275	0.1178	1.0000

90's data set correlation matrix					
	o_2pm	*o_3pm*	*d_2pm*	*d_3pm*	*wins*
o_2pm	1.0000	-0.2908	0.1746	0.2961	0.0423
o_3pm	-0.2908	1.0000	0.1491	0.1046	0.4059
d_2pm	0.1746	0.1491	1.0000	-0.4937	0.0099
d_3pm	0.2961	0.1046	-0.4937	1.0000	-0.3641
wins	0.0423	0.4059	0.0099	-0.3641	1.0000

For the 80's data set, we observe that the largest correlation coefficient for *wins* is *o_2pm*. This makes sense considering the time period, where three-point shots are relatively new to the NBA. We see that this linear relationship is accurately represented in the upper left figure (4). We can see there is a strong linear relationship between *wins* and *o_2pm* in this set. For the 90's data set, we observe that the largest correlation figure for *wins* is *d_2pm*. This is interesting, and somewhat surprising. Again in (4), in the upper right figure, we observe a clear linear relationship between *wins* and *d_2pm*. Finally, in the 00's data, the two strongest correlation figures for *wins* are *o_3pm* and *d_3pm*. This seems to follow the trend, where three-point shots are more valued than two-point shots in the common basketball era. Referring to figure (4) in the lower figure, we see a distinct linear relationship between wins and three-point figures.

3.5 Leverage analysis

Finally, we will observe the leverage plots for all three models and discuss the most influential teams in the plots.

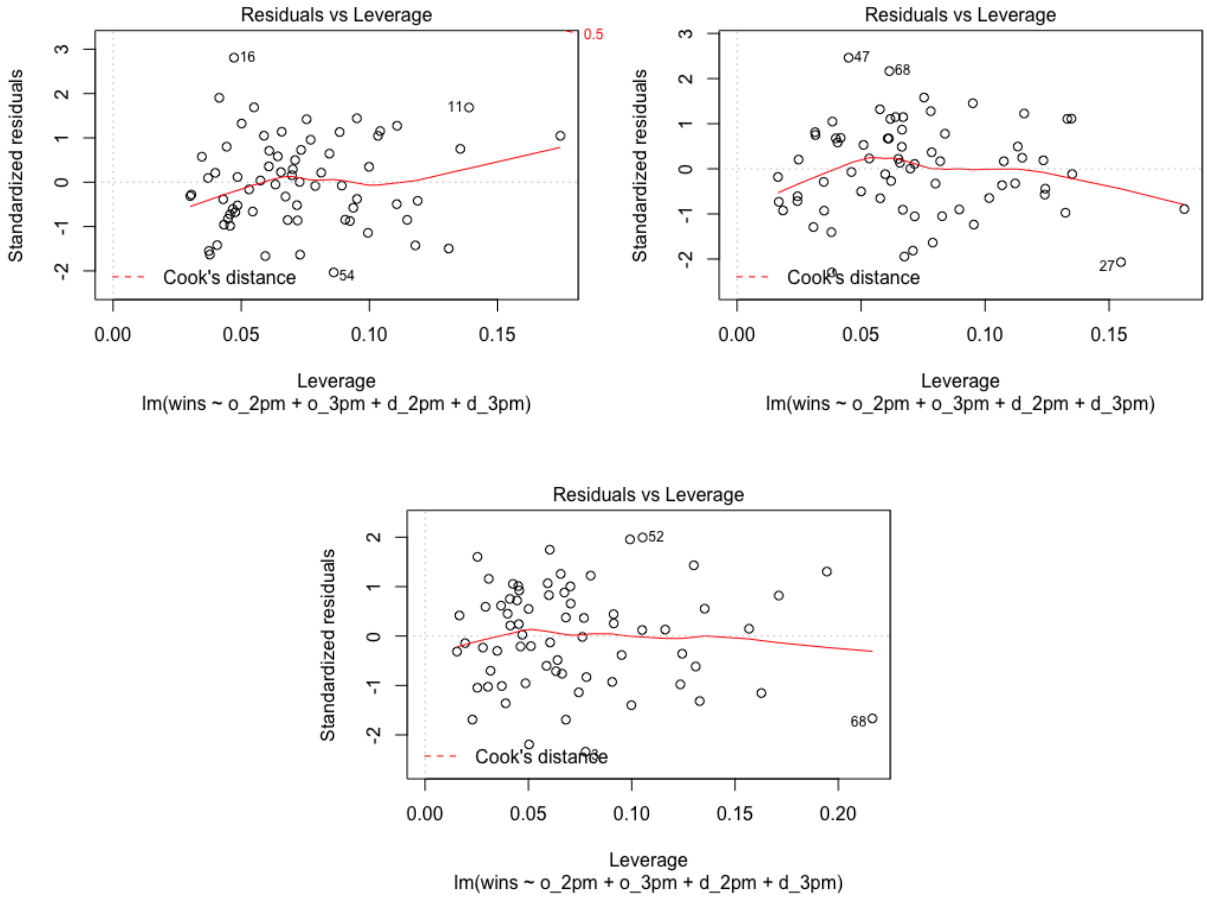

Figure 4: Leverage plots for each set. Upper left is 80's, upper right is 90's and lower is 00's

Here we observe that there are no major outliers that are in the Cook's distance that would cause issues. R highlights three data points in each set. Here are the three teams highlighted from the 1985-88 data set.

80's data set leverage highlights						
	name	*o_2pm*	*o_3pm*	*d_2pm*	*d_3pm*	*wins*
11	Los Angeles Clippers	3324	64	3588	261	32
16	Philadelphia 76ers	3384	51	3333	282	54
54	Golden State Warriors	3372	91	3157	470	20
	Mean values	3390.145	107.8261	3141.058	356.913	41

Here we see that both the Clippers and the Warriors are on the lower end of wins, while Philadelphia has well above the average wins. All three of these teams have low o_3pm values and high d_2pm values which explains why these values are highlighted.

Next we observe the teams highlighted in the 1995-98 data set.

90's data set leverage highlights						
	name	o_2pm	o_3pm	d_2pm	d_3pm	$wins$
27	Cleveland Cavaliers	2221	483	1258	1230	42
47	Atlanta Hawks	2550	337	1944	1017	50
68	Utah Jazz	2744	249	1684	1122	62
	Mean values	2545.101	443.2174	1722.014	1248.594	41.898

From these teams, we see that all of them are above the average win total. All three have lower d_3pm values than the average. The Jazz have the highest win total yet the lowest o_3pm value. Atlanta has a large d_2pm, which had the largest correlation to $wins$ in this data set. However, Atlanta still has fewer wins than Utah.

Finally we observe the highlighted teams in the 2005-08 data set.

00's data set leverage highlights						
	name	o_2pm	o_3pm	d_2pm	d_3pm	$wins$
3	Chicago Bulls	2444	560	1345	1428	41
52	Denver Nuggets	2738	569	1624	1710	50
68	Utah Jazz	2872	407	1477	1419	54
	Mean values	2499.377	504.6812	1610.246	1383.275	42.492

Again, in this data set, all three teams have higher win totals than the average. Denver and Chicago have slightly higher than average o_3pm values. Yet both have fewer wins than Utah, which has a remarkably lower o_3pm value. Additionally, all three teams have higher than average d_3pm values.

4 Conclusions

From the analysis of the models of the three data sets, we see that there is indeed an evolution occurring. Referring to the correlation, we see that two-point shots were in fact more significant than three-point shots, in both offense and defense, in the 1985-88 season compared to the 2005-08 seasons. Further suggestion of this trend is evident in the averages of makes and makes allowed as time progresses. Referring to the tables in section (3.5)

we observe that three-point makes and makes allowed increase with time, while two-point makes and makes allowed decrease. Clearly, with the addition of the three-point line, the style of play has changed. In the common NBA era we are experiencing a greater focus on three-point shooting as the data suggests.

In regards toward defense, restricted to this model and data set, there is no evidence suggesting a difference in the "importance of defense". In all three sets, we observe approximately the same total number of makes allowed (about 3300 per team) and approximately the same number of average wins. Additionally, the hypothesis test rejected the null hypothesis (from equation (2)) in each of the models. The test set our defense predictor coefficients to zero, which would imply that defense does not affect the win result. As expected, we showed that incorporating the defensive values does allow the model to more accurately predict the data. Hence, contrary to popular belief, defense is still an important factor for NBA teams.

The weaknesses of this experiment stem from the limited amount of data that we are considering. Limiting ourselves to only four predictors as well as only modeling on three consecutive seasons may not accurately predict all of the trends that are being experienced throughout the NBA. There are a number of other statistics such rebounds, assists, steals, etc. that are recorded in the NBA that we are ignoring. Obviously these other statistics will have an effect on the outcome of basketball games in addition to scoring.

To further extend the study of this data set, one may consider adding additional predictors to the model. Perhaps other elements of basketball better predict the outcome of games rather than purely scoring. Additionally, it would be interesting to model more recent data sets, to see how the style of play has changed in the last ten years. Is it similar to the 2005-08 data set, or are three-point shots dwarfing all other statistics in the current NBA? Something that may be considered, and worth further study, would be the incorporation of a four-point shot. How would that affect the style of play? Is a four-point shot more efficient than a three-point shot? How far away and how much a shot is worth will obviously change tactics and strategy, so is it ridiculous or not to have a four-point shot? Until such a time, we will still cheer on our three-point sharp shooters just like fans from the 80's cheered on their short-shot masters.