Modeling Shot Probability in the NBA

Ahmed Cheema, Charlie Henehan, Chris Stuckart

Abstract

In this study, we investigate the relationship between the outcome of NBA field goal attempts and several predictor variables describing the circumstances of each individual shot. Using comprehensive data from the 2014-15 NBA regular season, we apply a multiple logistic regression model to determine which factors impact a shot's success rate. In our analysis, we find that the location of the game, the period in the game, the time elapsed since the shooter initially touched the ball, the distance of the shot, the type of shot, and the distance from the closest defender are all significant predictors of whether a shot is successful. Through the use of a multiple logistic regression model with these input variables, we are able to evaluate individual NBA players and gain a better understanding of how these factors impact shooting efficiency as a whole

I. Background

Basketball is a team sport centered around shooting a ball into a hoop to score points. The team that scores more points wins, so shooting efficiency is intertwined with team success. For decades, teams have looked to maximize their chances of winning games and in recent years, these attempts have taken the form of in-depth statistical analysis (Ross). Nearly every NBA team employs data analysts who comb through data to draw insights to aid performance.

Basketball teams look to maximize their offensive efficiency by seeking the best shots possible. No two shots are alike -- every basketball shot is impacted by a variety of different variables, so controlling for these variables to produce good shot attempts can contribute to a more efficient offense overall (Gabor). Similarly, teams look to prevent their opponent from attempting efficient shots.

The goal of this study is to add to the league's analytics revolution by determining which variables actually impact a shot's outcome. In doing so, we hope to gain a better understanding of the factors that affect shooting efficiency. We also seek to demonstrate the applications of this approach to allow NBA teams to evaluate individual players based on the difficulty of their shots and to shape their gameplans around pursuing statistically efficient shots.

II. Data Exploration

The dataset explored in this study consists of 128,069 field goal attempts from the 2014-15 NBA regular season. In total, the dataset contains 51.1% of the total field goal attempts from the 2014-15 NBA season, as no shots after March 4th, 2015 are included.

The data includes nine candidate predictor variables and one response variable, all of which are described in Table 1.

Table 1: Description of variables

Name	Type	Description
location	Binary	Indicator for whether the field goal was attempted at the shooting team's home stadium
shot_number	Discrete	The shooter's n^{th} field goal attempt in the game
period	Discrete	The quarter of the game in which the field goal was attempted
dribbles	Discrete	The number of times the shooter dribbled the ball prior to the shot
touch_time	Continuous	The time elapsed in seconds since the shooter gained possession of the ball
shot_distance	Continuous	The distance between the shooter and the basket at the time of the shot
shot_type	Binary	Indicator for whether the field goal was attempted behind the three point line
def_dist	Continuous	The distance between the shooter and the nearest defensive player at the time of the shot
fgm	Binary	Indicator for whether the field goal attempt was successful

The data was collected through the use of SportVU, a video tracking system the NBA used in the 2014-15 season. Our research question is whether these variables impact shot probability, and, if they do, how we can use these to model the likelihood of a shot being made in the NBA.

Upon exploring the distributions of the variables, multiple impossible values were found in the *touch_time* variable. Some observations had a negative *touch_time*, but a player cannot

possess the basketball for a negative amount of time. Some other observations had a *touch_time* of over 24, which is also not possible because a single basketball possession can last a maximum of 24 seconds. The 316 rows containing these erroneous values were discarded, leaving us with 127,753 observations.

III. Initial Model Selection and Diagnostics

One issue that can negatively impact a multiple regression model's performance is collinearity among the predictor variables. A correlation matrix of the eight candidate predictor variables is shown in Table 2. The correlation coefficient between the *dribbles* and *touch_time* variables is equal to 0.931, which is indicative of extreme collinearity. The model will only include one of these two variables in order to minimize collinearity.

We will use backward elimination to select the variables for the model. If the corresponding p-values for any variable is greater than 0.05, the model will be fit again excluding that variable. This process will repeat itself until all of the p-values corresponding to variables in the model are below the significance level of 0.05. We will conduct model selection twice, once with the *dribbles* variable, and once with the *touch_time* variable. Both models will start with seven variables and the model with the higher adjusted R-squared will be selected.

The model using the *touch_time* variable has an adjusted R-squared of 4.01%, while the model with *dribbles* finished with an adjusted R-squared of 3.94%. Thus, the latter model was discarded and the summary for the model with *touch_time* is shown in Table 3.

The creation of diagnostic residual plots on this model demonstrated the lack of linearity between the continuous predictors and logit(p), which violates a key condition for fitting a logistic regression model. In order to circumvent this issue and improve model performance, logarithmic transformations were applied to the *touch_time*, *shot_dist*, and *def_dist* variables.

IV. Final Model Results and Diagnostics

The logistic regression model fit to the logarithmically transformed data exhibited greater performance, with an adjusted R-squared value of 4.54%, a 13.2% increase in variance explained from the initial model. The final model summary is shown below in Table 4.

The p-value for every variable is well below the significance level of 0.05. We can thus conclude that there is a relationship between each variable and the outcome of a field goal attempt.

The key assumptions for fitting a multiple logistic regression model can now be assessed. First, we must confirm that each outcome is independent of the other outcomes. As shown in Figure 1, there is no discernible trend on a broad scale between the order of collection and the model residuals. We can also look at a smaller subset of the data, as in Figure 2, to see that there is no relationship between the order in which the observations were collected and the model residuals. We can thus conclude that the outcomes are independent.

We must also assess the relationship between logit(p) and the predictor variables. For each observation, we can compute logit(p) as ln(p/(1-p)). The scatterplots in Figure 3 depict the linear relationship between logit(p) and the *shot_dist*, *def_dist*, *touch_time*, and *shot_number* variables. In order to study the residual structure of variables with limited levels, we use box plots as shown in Figures 4, 5, and 6. We can confirm that the residual distribution shape and

variability remains relatively constant for different levels of the *location*, *period*, and *shot_type* variables. Thus, we have satisfied the necessary conditions for multiple logistic regression.

V. Discussion

The p-values and estimated coefficients of the logistic regression model (Table 4) have multiple implications on the nature of shooting in the NBA. The distance of a field goal attempt unsurprisingly has a negative relationship with its outcome, while the distance between the shooter and the nearest defender has a positive relationship. Touch time is negatively related to shooting success rate, suggesting that a player holding onto the ball for a long period of time does not lead to efficient shots. The positive coefficient of the *location* variable along with the significant p-value also suggests that NBA players perform best on their home court.

The logistic regression model can be used to quantify shooting ability by comparing an individual player's actual output to the model's predicted output. In order to demonstrate a potential application of the model, we calculated each player's actual effective field goal percentage (an adjustment of traditional field goal percentage which weighs three-point makes 1.5x higher than two-point makes) to the model's predicted effective field goal percentage.

Table 5: Most efficient shooters relative to expectation

Player	FGA[1]	eFG%[2]	XeFG%[3]	Difference[4]
Kyle Korver	235	67.6%	49.3%	18.2%
Stephen Curry	470	58.4%	48.7%	9.7%
DeAndre Jordan	279	71.5%	62.2%	9.4%
Chris Paul	425	53.5%	44.3%	9.2%
JJ Redick	299	56.9%	49.8%	7.1%

[1] Total field goal attempts
[2] Effective field goal percentage
[3] Expected effective field goal percentage
[4] eFG% - XeFG%

The five most efficient shooters relative to expectation are shown in Table 5. Instead of simply assessing how efficient various players at shooting the ball, we can contextualize their efficiency relative to expectation. DeAndre Jordan may have a higher eFG% than Steph Curry, but he also has a far higher expected eFG% (XeFG%) because he attempts more shots close to the basket.

While our investigation successfully modeled shot probability with an adjusted R-squared value of 4.54%, it did have some limitations that leave room for improvement. The NBA has changed dramatically since the collection of the data analyzed in this study, so research on current data would be more meaningful. Since 2015, the average effective field goal percentage has increased from 49.6% to 53.7%. Furthermore, 39.4% of field goal attempts are three-pointers now versus the 26.8% three-point rate in 2015 (Sports Reference LLC). It is unclear how these shifts would impact the trends we found, but it would certainly be worth exploring.

There are many additional variables which future research can utilize to further improve the model. For instance, the difference in height between the shooter and the nearest defender can be considered. The *def_dist* variable on its own is limited because a 6'0 defender will not be able to contest a shot from two feet away as well as a 7'0 defender.

Analyses similar to this study have been conducted before, but we attempted to add to the literature by analyzing previously unexplored variables. While this study was not the first to incorporate granular shooting data, it served as an insightful examination on how various factors impact shooting accuracy and how different players perform relative to expectations formed based on those factors.

References

Becker, Dan. *NBA shot logs*. (Version 1) Kaggle, 2016. Web. 3 Mar 2021. https://www.kaggle.com/dansbecker/nba-shot-logs.

Gabor Csataljay, Nic James, Mike Hughes & Henriette Dancs (2013) Effects of defensive pressure on basketball shooting performance, International Journal of Performance Analysis in Sport, 13:3, 594-601, DOI: 10.1080/24748668.2013.11868673

Ross, Terrance F. “This Isn't Your Dad's NBA: Thank Big Data.” The Atlantic, Atlantic Media Company, 25 June 2015, www.theatlantic.com/entertainment/archive/2015/06/nba-data-analytics/396776/.

Sports Reference LLC. "NBA League Averages." Basketball-Reference.com - Basketball Statistics and History. https://www.basketball-reference.com/leagues/NBA_stats_totals.html.

Appendix

Table 2: Correlation matrix of candidate predictor variables

	location	shot_number	period	dribbles	touch_time	shot_dist	shot_type
shot_number	-0.003						
period	0.003	0.655					
dribbles	-0.015	0.141	0.055				
touch_time	-0.013	0.147	0.047	0.931			
shot_dist	-0.002	0.012	0.030	-0.083	-0.087		
shot_type	0.005	0.003	0.049	-0.170	-0.186	0.741	
def_dist	0.004	-0.038	-0.010	-0.154	-0.167	0.523	0.414

Table 3: Initial logistic regression model summary

	Estimate	Std. Error	Z-Value	P-Value
(Intercept)	1.284	0.065	19.91	0.00E+00
location	0.014	0.006	2.40	1.62E-02
shot_number	0.130	0.030	4.28	1.87E-05
period	-0.075	0.020	-3.71	2.05E-04
touch_time	-0.405	0.025	-16.33	6.09E-60
shot_dist	-1.604	0.025	-62.95	0.00E+00
shot_type	0.058	0.010	5.77	7.97E-09
def_dist	2.837	0.072	39.28	0.00E+00

Table 4: Final logistic regression model summary

	Estimate	Std. Error	Z-Value	P-Value
(Intercept)	0.773	0.029	26.72	0.00E+00
location	0.029	0.012	2.52	1.16E-02
shot_number	0.007	0.002	4.27	2.00E-05
period	-0.025	0.007	-3.75	1.75E-04
touch_time	-0.150	0.010	-14.99	1.15E-16
shot_dist	-0.722	0.011	-67.16	0.00E+00
shot_type	-0.145	0.018	-8.29	9.02E-51
def_dist	0.650	0.015	44.12	0.00E+00

Figure 1: Residuals vs. Order of Collection

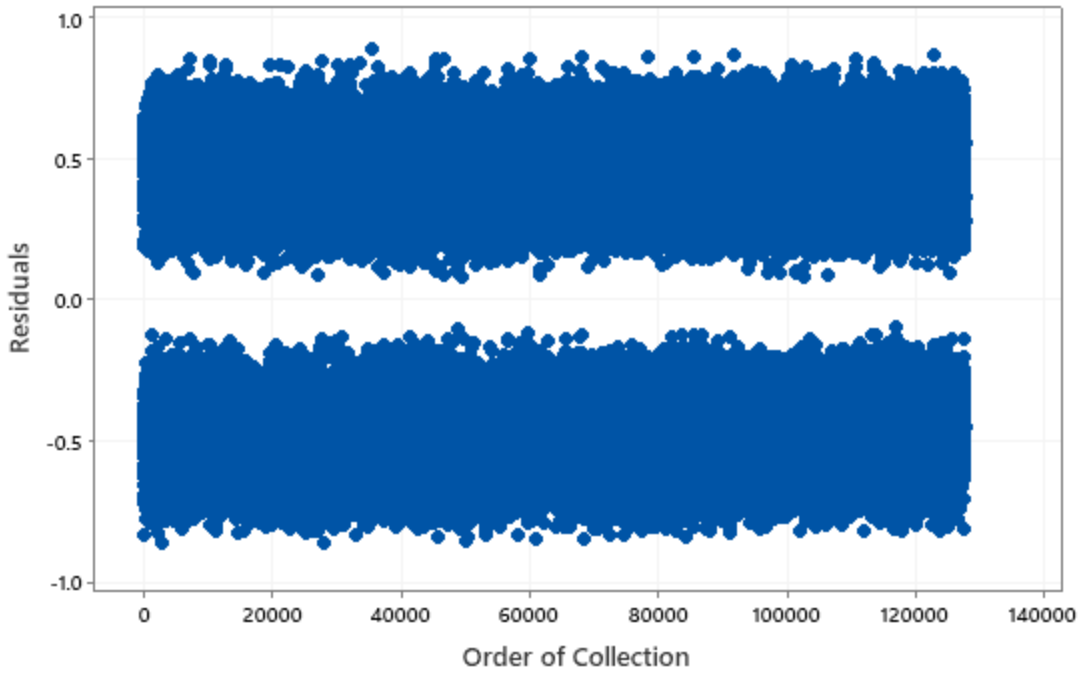

Figure 2: Residuals vs. Order of Collection (First 100 Observations)

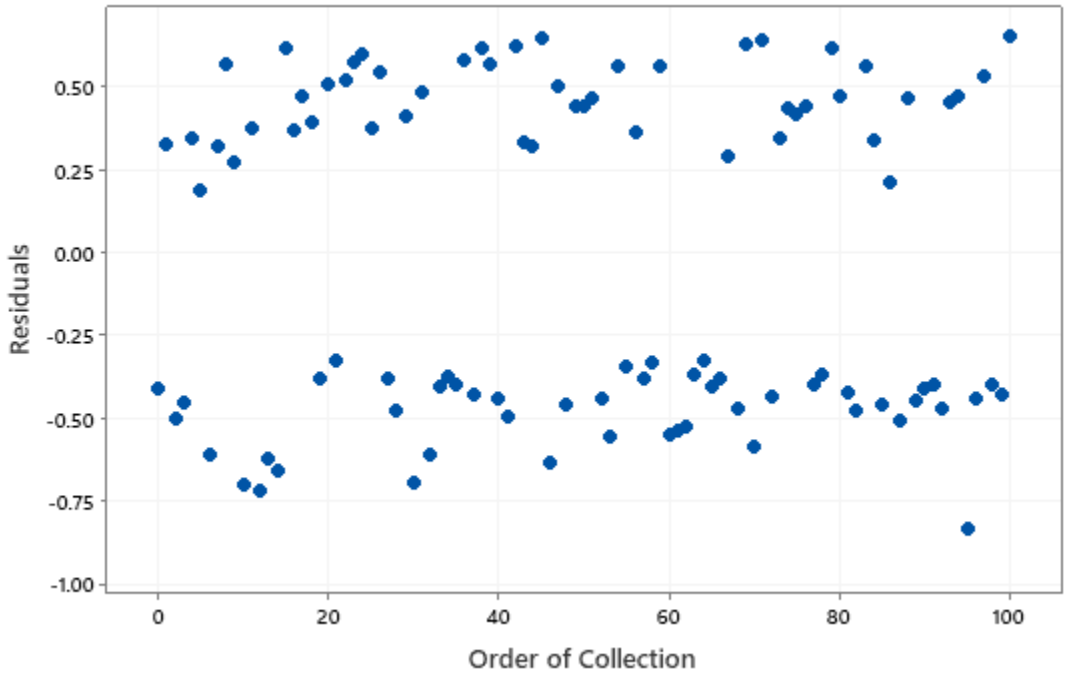

Figure 3: Relationship between logit(p) and numerical variables

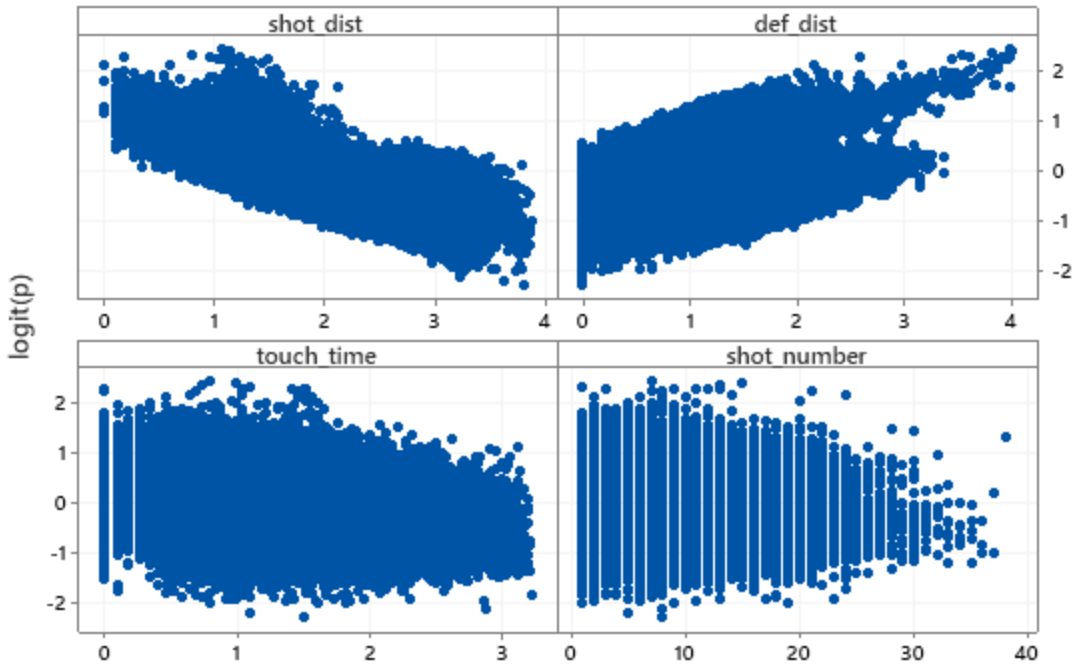

Figure 4: Residual distribution vs. location

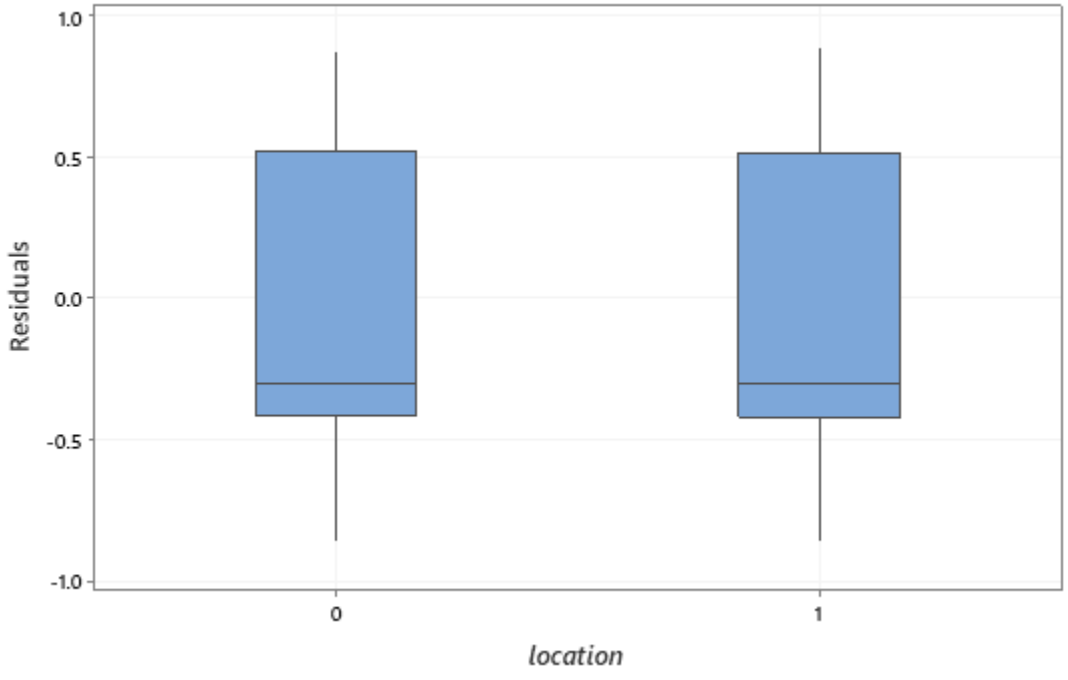

Figure 5: Residual distribution vs. period

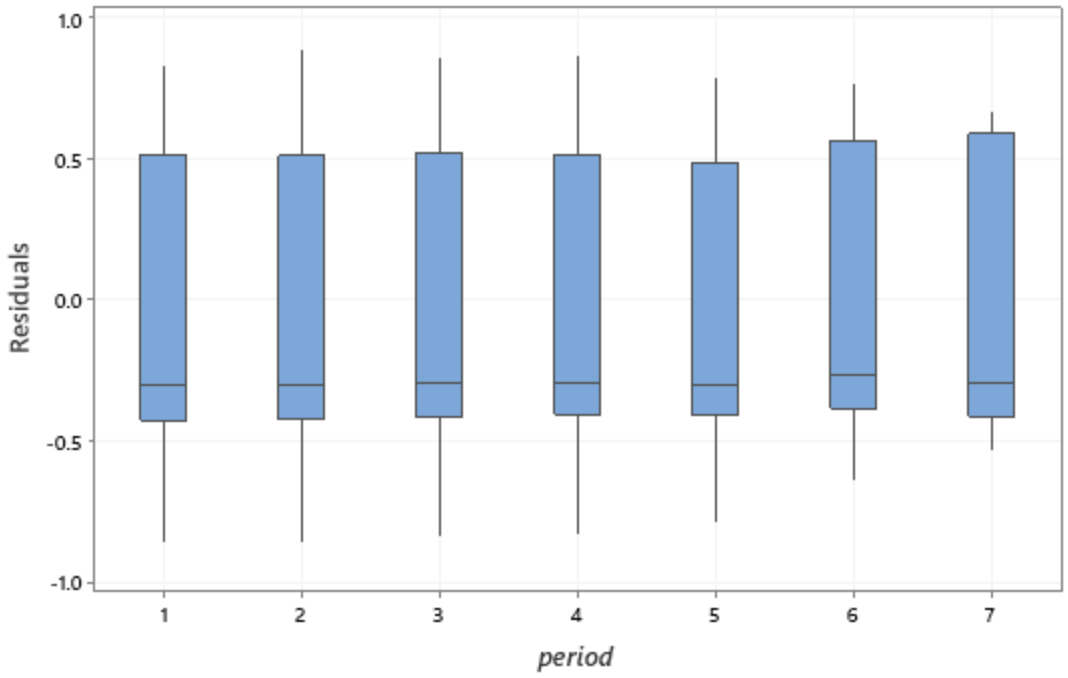

Figure 6: Residual distribution vs. shot_type