Statistical Research In Basketball

The use of logistic regression to find the most efficient playing strategy

Mattia Da Campo

June 12, 2020

Abstract

We present evidence, based on play-by-play data from the 2009/10 to the 2018/19 season of the National Basketball Association (NBA), that basketball scoring can be modeled by mathematics. We use logistic regression to test hundreds of different in-game scenarios that could affect the likelihood of a player making a basket, such as the amount of time left in the game, the score differential, and the defensive coverage. We demonstrate that certain shots are more efficient than others, and show that the playing tendencies of teams and players have drastically changed in the last 10 years.

Contents

1 Introduction

This research is motivated by my immense passion for both basketball and mathematics. This passion was nourished by incredible people who have been supporting me especially during these four years at Seattle University. This paper is the result of continuous and unconditional assistance, encouragement, and love by the Math Department, my coaches and my family.

1.1 What is basketball?

The goal of basketball is simple, its rules are well defined and the results are easily quantifiable. Basketball is played 5-on-5, and each team tries to put the ball in the basket positioned at 10ft from the floor. Not every shot is worth the same. In fact, if a player shoots the ball behind the 3 point line (the "*arc*") and makes it, the shot will be worth 3 points. A shot made from inside the arc is worth 2 points. If a player is fouled in the attempt of shooting, he will have two unguarded shots from the free throw line. A game consists of 4 quarters of 12 minutes each, and for every possession a team has 24 seconds to complete their action. The team that at the end of the game has the most points wins the game.

1.2 Question

As we will see, there is an enormous amount of high-quality and detailed data, and that's why basketball is a great and rich laboratory for statistical and mathematical study. This research is motivated by the following question: can basketball scoring be described by mathematics?

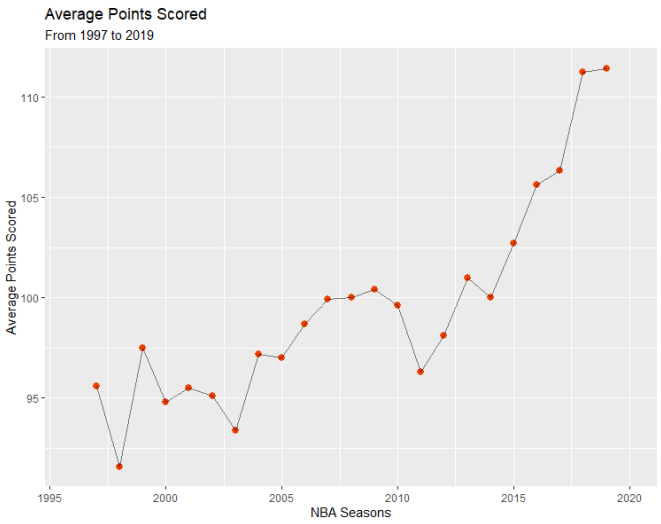

Figure 1.1: Average points scored per game from 1995 to 2019

In Figure 1.1 are shown the average points scored per game in the last twenty-three years. My hypothesis is that such an increase in points scored is given by a change in shot selection by teams and players.

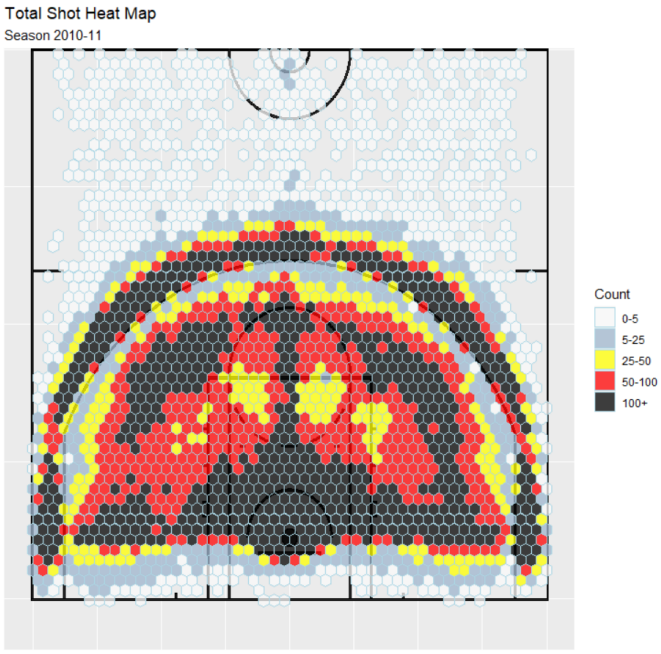

Figure 1.2: Heat Map 2010-11 Season

In Figure 1.2 and 1.3 we have two heat maps from the 2010 and 2018 season. We divide the court in bins of size 1 foot squared, and count how many shots are taken inside each bin. Notice that the distribution of shots has shifted either close to the basket or outside the three point line. In this paper we will mathematically describe why this change happened. Understanding the effect of shot selection and in-game decisions on a team's performance is essential. Team managers and coaches can use this information to make decisions about which players to hire, for example, or determine optimal policies regarding playing style and team tendencies.

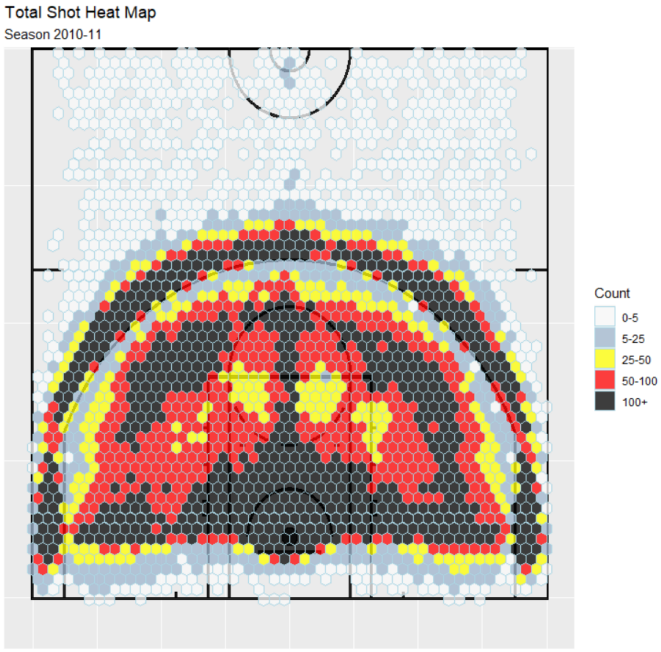

Figure 1.3: Heat Map 2018-19 Season

1.3 Method

To answer this question, we analyze play-by-play data from the 2009/10 to the 2018/19 season of the National Basketball Association (NBA), as well as shooting data from 1999 to 2020 and additional data from the 2019/2020 College Division I Men's Basketball season. At first we proceed with an initial exploratory analysis, confirming that two important variables such as total points scored and margin between teams follow a normal distribution. Then, we develop a logistic regression model, where our null hypothesis is that shot distance, game situation and shot characteristics do not influence the probability of making a basket. We will find that we have enough evidence to reject the null hypothesis. We will arrive at the conclusion that many variables have a negative relationship with making the shot, and we will conclude this section with the making of a web application based on the logit model, that given shot characteristics returns the probability of making the shot. Given the fact that not every shot is worth the same number of points, and utilizing the model described above, we will find which shots are more efficient and will show how this insight has initiated a substantial change in playing style in the last twenty years.

2 EDA (Exploratory Data Analysis)

Play-by-play data records every event that happens in a basketball game, such as shots, rebounds, turnovers, substitutions and more. There are 30 teams and each one of them plays a total of 82 regular season games. That makes a total of 1230 total games per season. On average, each team attempts 84 shots per game. That makes more than 200,000 shots per season. Given that we are analyzing 10 seasons of NBA basketball, we will work with more than 2 million observations. This kind of data is provided by *BigDataBall*[2], where the reader can find detailed data for any major sport.

First, let's observe total points scored per game. In Figure 2.1 we see the distribution of average points scored per game

distribution. The red line is the normal distribution curve with mean 101 points and standard deviation 10. Following the 68-95-99 rule we can say that on average a team has 95% chance of scoring between 81 and 121 points.

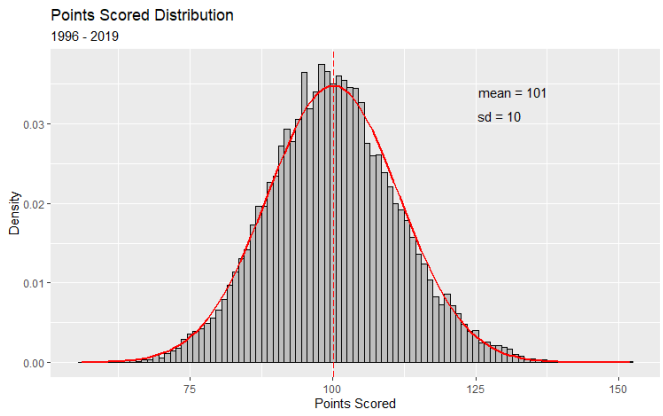

Figure 2.1: Average Points Scored Distribution

We can do the same process with the margin between two teams after 48 minutes. In Figure 2.2 we take the score difference between home and away team after 48 minutes of game. We notice two things different than before. There is a spike at 0 points and the mean is 2.1. The first one is because 11% of NBA games end in overtime. When a team is down two points with less than a minute in the game, it will shoot a two pointer 70% of the time. The mentality here is to be safe first, and worry about winning the game later. The latter is because of what we call 'home court advantage', which is given by many reasons, such as home crowd/fans and not traveling the day before the game.

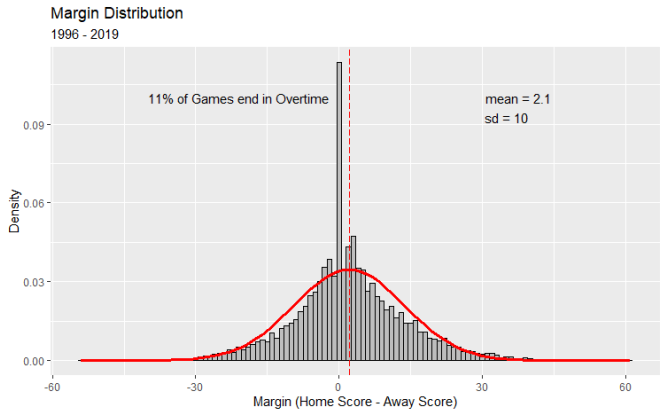

Figure 2.2: Margin between home team and away team after 48 minutes

3 Logistic Regression

3.1 Method

I learned logistic regression in my econometrics studies. We studied mortgage lending decisions at different banks and we found traces of discrimination between white, black, and Hispanic applicants. This type of regression is also used in machine learning, many medical fields and social sciences, where there is the need to find the relationship between a categorical dependent variable and some independent variables, also called 'predictors'. In our case, the dependent variable will be if the basketball shot is made or not.

Let

$$z = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{3.1}$$

where z is a transformation variable of our predictors x_i and their respective coefficients β_i. For example, a one unit increase in x_i will represent a β_i-unit increase in z . The relationship between this variable z and the outcome is shown in Figure 3.1 .

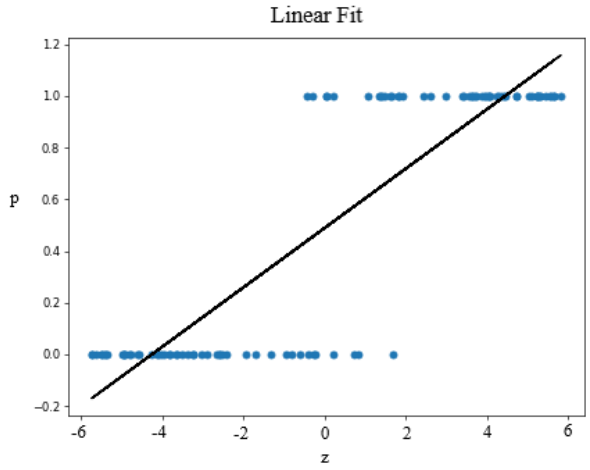

Figure 3.1: Fitting a line to a categorical variable will return probabilities greater than 1 and smaller than 0.

Notice that if we try to fit the line $p(z) = z$ then

$$\lim_{z\to\infty} p = \infty \qquad \lim_{z\to-\infty} p = -\infty \tag{3.2}$$

And we know this is not possible, since p has to be bounded between 0 and 1 .

To solve this, we use what we call a *sigmoid function* (Figure 3.2), which has the form

$$p(z) = \frac{1}{1+e^{-z}} \tag{3.3}$$

Now we have that

$$\lim_{z\to\infty} p = 1 \qquad \lim_{z\to-\infty} p = 0 \tag{3.4}$$

However, the relationship between the transformation variable z and the outcome is not linear. In order to make it so, we need two steps. First, we take the odds of the *Sigmoid* function. If we let P_A be the probability of the event A to happen. Then

$$odds_A = \frac{P_A}{1-P_A} \tag{3.5}$$

which is the probability of the event happening over the probability of not happening. For example, if on average I make a free throw 80% of the times, the odds of making the shot will be $\frac{0.8}{0.2}$ which is equal to 4 to 1 odds of making the shot.

Now let's take the odds of the Sigmoid function.

$$odds = \frac{p(z)}{1-p(z)} = \frac{\frac{1}{1+e^{-z}}}{1-\frac{1}{1+e^{-z}}} = e^z \tag{3.6}$$

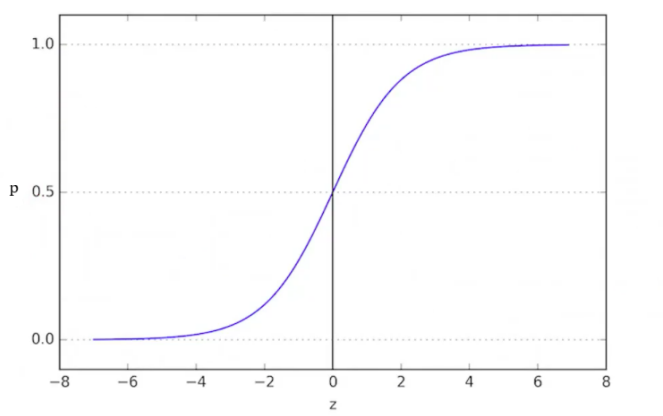

Figure 3.2: We use a Sigmoid Function so the probabilities are bounder between 0 and 1

Looking at equation 3.6 we would expect an exponential relationship between the predictors and the odds of making a shot. In Figure 3.3 we compare the expected odds against the actual odds of making a shot. In this study we will mainly focus on shots between 0 and 30 ft, so we are satisfied with this relationship.

Figure 3.3: Real odds of making a shot are taken from the real shooting percentages, compared to what we would expect from the logistic model

The next step is taking the *logit transformation* of the odds which we define as the natural log of the odds. We already know that $odds = e^z$ then

$$logit = \ell = \ln(odds) = \ln\left(e^z\right) = z \tag{3.7}$$

In Figure 3.4 we see the linear relationship for both the expected and actual logit.

Then, the logit model will have the form:

$$logit = \ell = \ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{3.8}$$

Where ℓ is the log-odds and β_k are the parameters of the model, and x_k the values taken by our independent variables. A explained in '*An introduction to Logistic Regression Analysis and Reporting*'(Peng et al., 2002 [10]), the value of the coefficients β determines the direction of the relationship between x and the logit of the dependent variable (Y) . When β is greater than zero, larger (or smaller) x values are associated with larger(or smaller) logits of Y . On the other hand, if β is less than zero, larger(or smaller) values of x are associated with smaller(or larger) logits of Y .

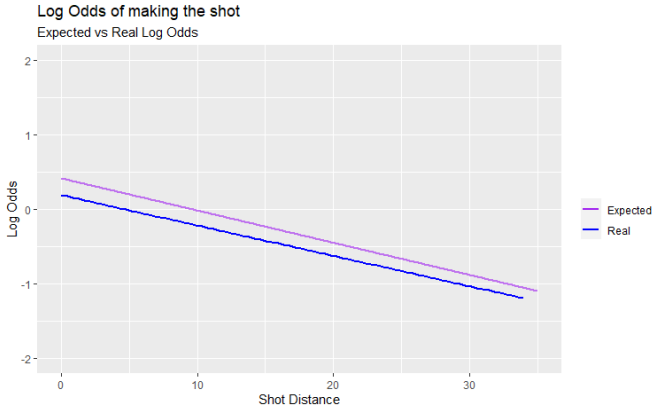

Figure 3.4: Real log-odds of making a shot are taken from the real shooting percentages, compared to what we would expect from the logistic model

3.2 Define Variables

A number of different aspects of a shot could impact a player's chances of success. For example, at the end of a close game, a player may be nervous and the pressure of the game might impact his performance. Also, from personal experience the distance of the closest defender has a huge impact on shooting accuracy. I chose to work with variables I thought were more valuable in determining the probability of making the shot. As we will see later in the paper, there are other aspects of the game I decided to not include for different reasons.

We associate the following varables to each shot:

- **Shot made** → *binary*
  - made shot = 1
  - missed shot = 0
- **Shot distance** → *numeric*
  - distance from the basket in ft (min = 0, max = 96)
- **Crunch situation** → *binary* | =1 if less than 4 minutes left in the game & margin <5 pts
- **Shot Clock** → factorized
  - Early Shot when 18-24 seconds
  - Regular Shot when 6-18
  - Late Shot when <6 seconds
- **Closest Defender** → *factorized*
  - Wide open if closest defender >10ft
  - Open if closest defender 6-10 ft
  - Contested if 3-6 ft - Forced if 0-3 ft
- **Dribbles** → *factorized*
  - 0 dribbles
  - 1-2 dribbles
  - 3-5 dribbles
  - 6+ dribbles
- **Touchtime** → *factorized*
  - less than 1 second before shot
  - 1-3 seconds
  - 3-5 seconds - 5+ seconds

Why did we decide to not use continuous variables but to factorize them? In economics the law of marginal utilities states that as consumption increases the marginal utility derived from each additional unit declines.

In Figure 3.6 we see a contour plot of the distance of the closest defender. We notice diminishing marginal effects as

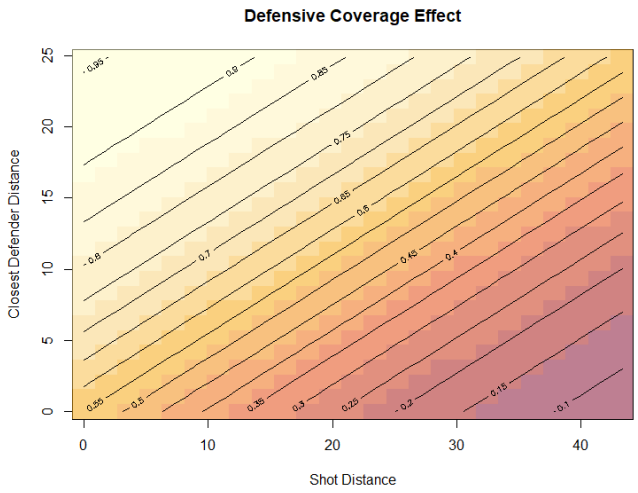

Figure 3.5: Contour Plot of shot distance against defensive coverage. The lines is black are the predicted probability of making the shot. Notice that as distance of closest defender increase the impact of the probability becomes smaller and smaller.

the distance increases. The difference in probability between 0 and 5 ft is 15 percentage points, while the difference between 20 and 25 is less than 5. From the perspective of a basketball player, having the defender attached to you or at 5 ft makes a big difference because you are able to get the shot off. On the other side, if the defender is at 25 ft rather than 20 ft, the shooter would barely notice the difference. A similar case could be made for the other factor variables.

3.3 Models

3.3.1 Reduced Model

Let's start with considering a reduced model with shot distance and crunch situation as predictors(Table 3.1).

	Dependent variable:	
	FGM	
	(1)	OR
SHOT DISTANCE	-0.043^{***}	0.958
	(0.001)	
CRUNCH	-0.112^{***}	0.894
	(0.037)	
Constant	0.406^{***}	
	(0.011)	
Observations	119,320	
Log Likelihood	−80,177.650	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3.1: Reduced model. For space reasons, the full model table can be found at the end of the paper.

From the sign of the coefficients we can say that there is a negative relationship with making the shot. That is, as distance increases the logit will decrease. We use odds ratio (OR) to compare the relative odds of making the shot across two different groups.

We define

$$OR = \frac{odds_1}{odds_0} \tag{3.9}$$

where $odds_1$ are the odds of the group we are checking and $odds_0$ the odds of the reference group. Let's compare the odds of making a crunch shot against a regular shot. Let's hold shot distance constant. We know that $odds = e^{\beta_0 + \beta_1 \cdot CRUNCH}$ from equation 3.6 where β_0 is the intercept, β_1 is the coefficient for crunch shot, and CRUNCH is a categorical variable that can take the value of 0 or 1. Then

$$OR = \frac{odds_{CRUNCH}}{odds_{REGULAR}} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1} \tag{3.10}$$

Thus, to calculate the odds ratio of a variable with respect to the reference group, we will exponentiate the respective logit coefficient. An odds ratio of 0.894 means that on average, controlling for shot distance, shots taken in crunch time situations have 10% lower odds of going in. For a continuous variable like shot distance, we know that on average, for every extra foot we go away from the basket, the odds of making the shot decrease by 4.2% . However, differences in predicted probabilities may be more meaningful than odds ratios.

Let's find the predicted probability of making a shot taken from 15 feet in a crunch time situation. We need our z value. We have that

$$z = \beta_0 + 15 \cdot \beta_1 + 1 \cdot \beta_2 \tag{3.11}$$

where β_0 is the constant and β_1 and β_2 the coefficients for shot distance and crunch situation respectively. Then,

$$z = 0.406 - 15 \cdot 0.043 - 0.112 = -0.351 \tag{3.12}$$

Then let's substitute z into the Sigmoid function. We have

$$p(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{0.351}} = 0.413 \tag{3.13}$$

On average, a 15 ft shot in crunch time has a predicted probability of going in of 41% . The effects of crunch time are shown in Figure 3.6

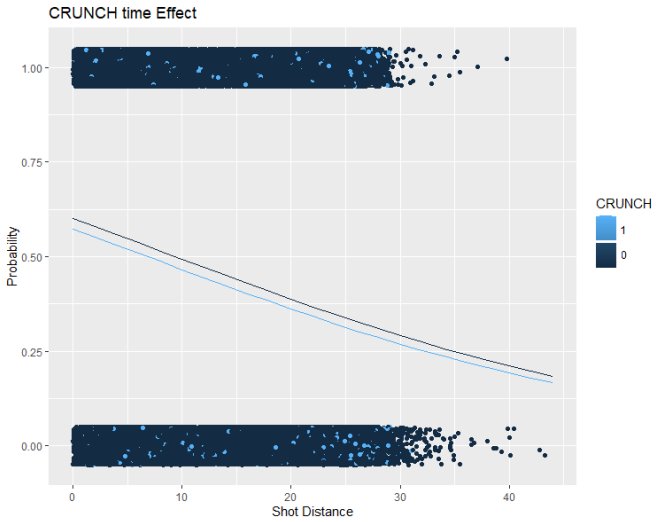

Figure 3.6: The predicted effect of taking a shot in a crunch situation compared to a regular situation

3.3.2 Full model

Every time we wish to add an independent variable to our model we need to check for the assumption of *little or no multicollinearity*. Multicollinearity problems consist of including in the model different variables that have similar predictive relationships with the outcome. We first create a correlation table (Figure 4.1) to see which variables could cause problems.

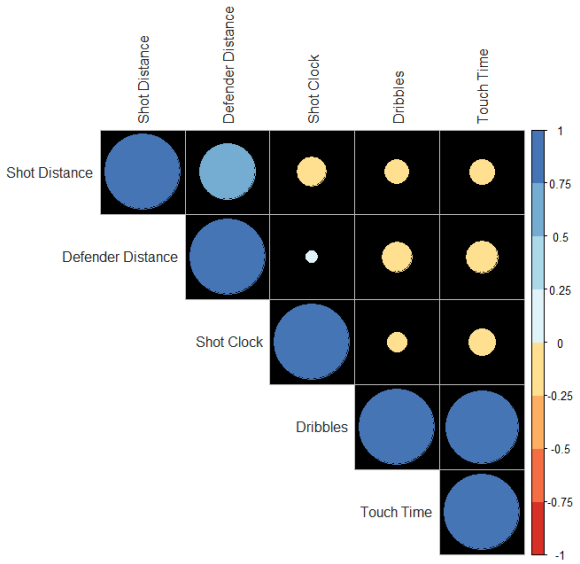

Figure 3.7: Correlation Table between variables in the model.

We see that dribbles and touchtime are highly correlated, which makes sense because for each extra dribble taken the touch time before the shot increases. We need to check if these two variables have similar predictive power in the model. We calculate the *Variance Inflation Factor (VIF)*.

High values signify that it is difficult to assess the contribution of predictors to a model. A value greater than 5 is generally considered bad and the solution is to not insert *touchtime* in the model (Tables 3.2 and 3.3).

The full logit model is shown in Table 7.1. Notice that for our factor variables we don't include one of each. The excluded variables will represent our reference variable, which will be a shot not in a crunch situation, between 6 and 18 seconds on the shot clock, wide open and with no dribbles. The signs are as expected: there is a negative relationship between shot distance and the logit of making a shot; same thing with number of dribbles and defender distance. In addition, shooting and early shot is associated with higher likelihood of making a shot. Also, every coefficient is statistically significant at the 0.01 level. Looking at odds ratios we can say that on average, holding other variables constant, an open shot (defender between 6-10 ft) has odds 11% lower of going in respect to a wide-open shot (10+ ft). For a forced shot (0-2 ft) the odds are 59% lower than a wide open one. Let's find the probabilities as we did before. Let's say two shots are taken from 15ft. For the first one (1) the defender is at 8 ft (open shot) and for the second one (2) at 2 ft (forced shot). Then

$$z_1 = 1.221 - 15 \cdot 0.061 - 1 \cdot 0.117 = 0.189$$
$$z_2 = 1.221 - 15 \cdot 0.061 - 1 \cdot 0.880 = -0.574$$

When we plug in z_1 and z_2 into the Sigmoid function we get

$$p_1 = 0.547$$
$$p_2 = 0.36$$

The defensive coverage effect is shown in Figure 4.2. We could do the same comparison for other variables in the model such as shot clock and dribbles. Some examples of these variables' effects on the probability of making the shot are shown at the end of the paper (here and here). To calculate the exact probabilities of different combinations of shot characteristics, we would have to go through the same process as before. If for example we let shots be between 0 and 40 ft, then we have 2 choices for crunch situations, 3 choices for shot clock, 4 choices of defensive coverage and other 4 choices for number of dribbles. That makes a total of 3840 combinations. This is the reason why I developed a tool that lets you find the probability of making the shot for any of those combinations in less than 10 seconds. This application was built using the R studio Shiny App functionality and can be found at the following link and also in the reference section. I invite the reader to open it and verify our calculation above for an open and contested 15 feet shot.

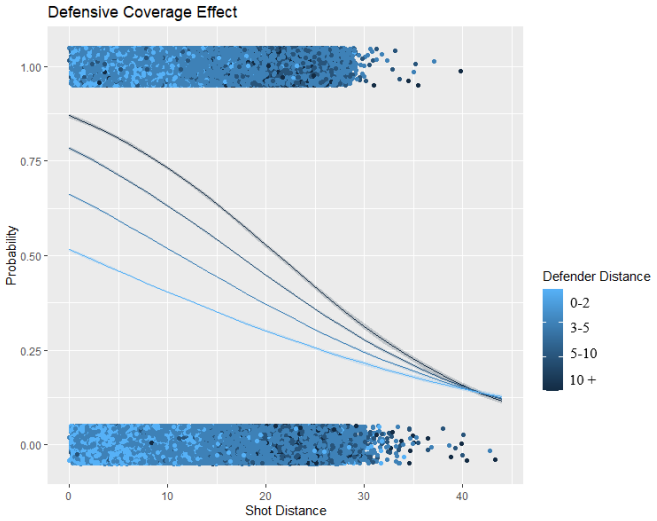

Figure 3.8: The effects of different defensive coverage. If the player is under the basket and the closest defender is at more than 10 feet away the expected probability is about 0.90. However, if the defender is between 0 and 2 ft, the probability drastically drops at 0.50

4 Goodness of Fit

4.1 MLE

One more thing I wanted to understand during my research was how the logic coefficients $(\beta_0, \beta_1, \ldots, \beta_k)$ are calculated. A logit model uses a *maximum likelihood estimation (MLE)*. This method is different from the ordinary least square estimation in linear regression, where the model finds the best fit by minimizing the distance between the observed value and the regression line. This method would not be possible, since as we saw in Section 3.1 we cannot fit a line on a categorical variable.

Variables	Shot Distance	CRUNCH	Shot Clock	Dribbles	Touch time	Defender Distance
VIF	1.597	1.003	1.072	7.389	7.536	1.592

Table 3.2: Variance Inflation Factor : all variables included

Variables	Shot Distance	CRUNCH	Shot Clock	Dribbles	Defender Distance
VIF	1.595	1.004	1.038	1.039	1.590

Table 3.3: Variance Inflation Factor : touch time excluded

MLE selects the set of parameter estimates that give the highest probability of obtaining the observed result. If we let

$$p(y = 1|\vec{x}) \tag{4.1}$$

be the probability of making the shot given predictors x, then we can say that

$$p(y = 0|\vec{x}) = 1 - p(y = 1|\vec{x}) \tag{4.2}$$

Then the likelihood function is given by

$$\prod_{i=0}^{119{,}320} p(y_i = 1|\vec{x})^y \cdot [1 - p(y_i = 1|\vec{x})]^{1-y} \tag{4.3}$$

Notice that the single terms of Eq 4.3 take the value of $p(y_i = 1|\vec{x})$ if $y = 1$ and the value of $1 - p(y_i = 1|\vec{x})$ if $y = 0$. For example, if the predicted probability of making a shot was 0.75 and the shot was made, we would multiply by 0.75. However, if the shot was missed, i.e. the model didn't predict well, we will multiply by 0.25. Since we have more than a hundred thousand observations, the likelihood will be close to 0. For mathematical convenience, we usually take the logarithm of the likelihood. Since the likelihood function is bounded between 0 and 1, the log likelihood will always be a negative number. The MLE finds the values of $(\beta_0, \beta_1, \ldots, \beta_k)$ that maximizes the log likelihood of a model.

4.2 Estimating goodness of model

How do we know that our logit model is accurate? Should we be confident in it when estimating probabilities of making the shot? We will conduct three different tests to measure the goodness of our model.

1. Likelihood Ratio Test
2. McFadden Pseudo R^2
3. TT test

The first two tests will use the likelihood function explained in Section 4.1 to assess the improvement of the full model over the reduced model (LR test) and to find an index analog to the R^2 known in linear regression. The third test is something less canonical, something I came up with to measure the precision of the model and possibly improve it. Following the unwritten rule of mathematics that everything has to have a name, I called it with my nickname.

4.2.1 Likelihood Ratio Test

Absolute values of Log Likelihoods are not meaningful on their own, since they depend on the sample size. However, for a given sample size the smaller the absolute value the better the fit. The Likelihood Ratio test is a hypothesis test that assesses the difference in predictive power between two nested models. In our case, we will test if the full model significantly improves the reduced model. We define the LR test value (γ_{LR}) as:

$$\gamma_{LR} = 2 \cdot [\log(L)|\hat{\beta_{Full}} - \log(L)|\hat{\beta_{Red}}] \tag{4.4}$$

Where $\log(L)|\hat{\beta_{Full}}$ is the log likelihood achieved at the maximum likelihood estimates of the full model, and $\log(L)|\hat{\beta_{Red}}$ is the log likelihood achieved with the reduced model. Professor Wilks in his paper *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*((Wilks, 1938) [11]) proves that this difference follows a χ square distribution with q degrees of freedom, where q is the number of variables left out in the model. Referring to Table 7.1 at the bottom, we can see the log likelihoods of the two models. Then we have

$$\gamma_{LR} = 2 \cdot (-78{,}917.420 + 80{,}177.650)$$

$$\gamma_{LR} = 2520.46 \quad q = 8$$

A critical value of 2350 with 8 degrees of freedom is statistically significant at the 0.01 level. Therefore, we have enough evidence to reject the null hypothesis that the full model does not improve the reduced one. Now we know that adding variables improved the predictive power of our model. Let's keep going with our testing.

4.2.2 McFadden Pseudo R^2

In linear regression there is an index called R^2 which assesses what percentage of the variation in the dependent variable can be explained by the predictors in the model. We can transform the log-likelihood function into an analogous index. It's called McFadden Pseudo R^2, from the professor who invented it, and it is defined as:

$$\rho^2 = 1 - \frac{\log(L)|\hat{\beta_{Full}}}{\log(L)|\hat{\beta_0}} \tag{4.5}$$

Where $\log(L)|\hat{\beta_{Full}}$ is the log likelihood achieved at the maximum likelihood estimates, and $\log(L)|\hat{\beta_0}$ is the log likelihood obtained on the null hypothesis being true, i.e. considering only the intercept.

The McFadden pseudo R^2 is only one of numerous Pseudo R squares we could choose. I preferred McFadden because it's straightforward and it lets us use the Likelihood function. I encourage the reader to find more about this using the reference page at the end of the paper(reference). It is important to know that these R squared tests come with limitations. In fact, each one of them will give different results and for this

reason they have to be treated carefully. We cannot interpret his index as the classic linear regression R^2. Even if it appears to be a stable relationship between the two indices (Figure 4.1)professor McFadden in the footnotes of his paper *Quantitative Methods for analyzing travel behavior for individuals* ((McFadden, 1977)[8]) writes that "values of ρ^2 tend to be considerably lower than those of the R^2." We can find these values in Table 7.1. The reduced and full model have a Pseudo R^2 value of respectively 0.33 and 0.54. At page 35 in McFadden paper the professor explains that values of .2 to .4 for ρ^2 represent an excellent fit.

Therefore, we are satisfied with a value of 0.54, that following Figure 4.1 could be compared to a standard R^2 value of 0.8. Now, let's proceed with our last test.

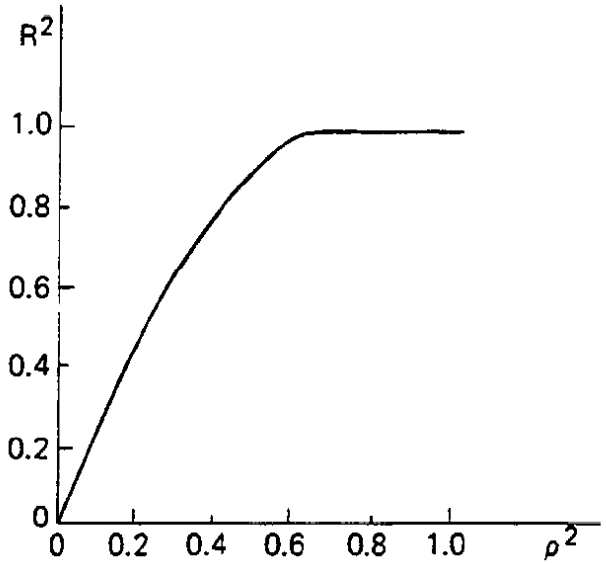

Figure 4.1: From McFadden's paper 'Quantitative methods for analyzing travel beavior for individuals'. The relationship between the two indexes appears to be stable, even though not entirely linear. In fact, ρ^2 tend to be lower.

4.2.3 TT test

We want to test the ability of the model to forecast observed responses. In other words, to find how precise the model is at predicting shot probabilities. On average a shot goes in 45% of the times, so let's classify a shot as made if its predicted probability is at least 0.45. Let's reference Table 4.3, where we have 4 random observations from our data set. We can assess the precision of the model by comparing the prediction and the actual result. For example, the model predicted correctly the first observation but not the other three. Doing this process with every observation in our data set (Table 4.4) we get *sensitivity = 0.54* and *specificity = 0.62* , where sensitivity is the true positive rate, that is how many times the model predicted a made shot correctly, and specificity is the true negative rate, that is how many times the model predicted a missed shot as missed. We can combine these two values and find that the precision of the model is 0.59, that is a correct prediction was made 59% of the time. The major limitation of this process is that we are using the same threshold of 0.45 for every type of shot. This will lead to losses in precision, since for example shots from 1 ft are made 63% of the time and from 35 feet only 15% of the time.

	MISSED	MADE
PREDICTED MISSED	29490	19232
PREDICTED MADE	17556	23212

Table 4.1: Four random observations and respective prediction. With a threshold of 0.45 the model predicted the result right 1 time out of 4

Let's find the best threshold for every shot distance. As a motivating example, we consider only shots taken from 4 feet (Figure 4.2

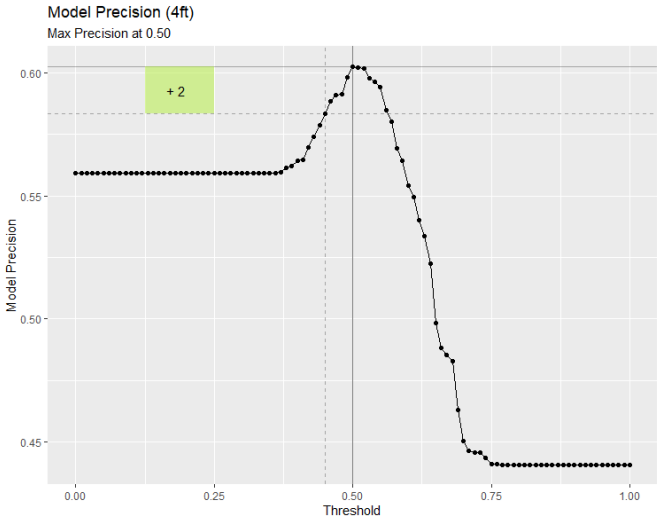

Figure 4.2: Precision graph for 4 feet shots. On the x axis we have every possible threshold we can choose from 0 to 1. On the y axis we have the precision we would get if we used a specific threshold. From this we can determine which threshold returns the highest precision. In this case: 0.50

Before we calculated the precision of the model when our threshold was 0.45. Now, we try every possible threshold from 0.01 to 1, and plot it against the respective precision. Doing so, we can figure out which threshold gives the highest precision. Figure 4.2 tells us we should use 0.50 instead of 0.45, this will bump out precision for 4 ft shots by two percentage points. It is interesting what happens at the straight lines. Among 4 feet shots, there is not a predicted probability less than 0.37. If we choose any threshold between 0 and 0.37 we will classify every shot as made, since every predicted probability will be higher than our threshold. By doing so, we will get back the percentage of real makes. On the other side, there is not a predicted probability higher than 0.75. Therefore, if we choose a threshold of 0.75 or more, every shot will be predicted as a miss, since there is not a predicted probability higher than our threshold. The results of doing this process for every shot distance are shown in Figure 4.3.

	FALSE	TRUE
PREDICTED MISSED	40631	8091
PREDICTED MADE	26691	14077

Table 4.2: Four random observations and respective prediction. With a threshold that varies with shot distance the model predicted the result right 2 time out of 4

The graph tells us which threshold we should choose. For

Figure 4.3: The same process as before is done for each shot distance, and the threshold which returns the highest precision is plotted. In black we see the real shooting percentages for reference.

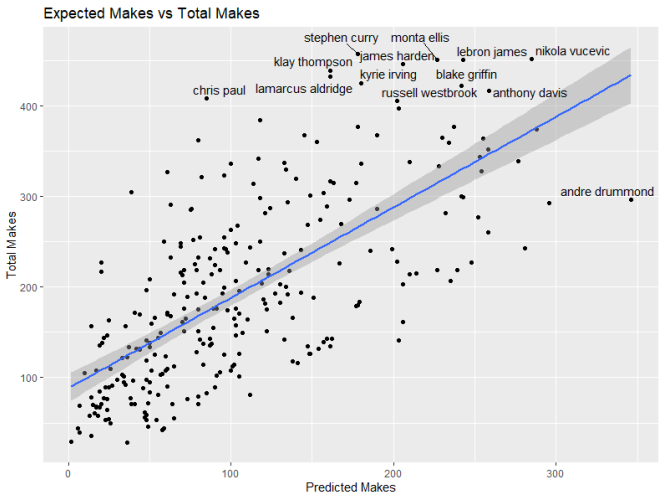

Figure 4.4: The same process as before is done for each shot distance, and the threshold which returns the highest precision is plotted. In black we see the real shooting percentages for reference.

reference, the actual percentages were plotted in black. Let's analyze the same four observations as before, but this time let's use the threshold plot we just built. Observation 1 and 2 remain the same. Observation 3 however will have a different prediction. We just saw that 4 feet have a threshold of 0.50. Therefore, we will predict observation 3 as missed and improve our model.

Doing this for every observation, we get *sensitivity* = *0.64* and *specificity* = *0.60*, *precision* = *0.61*. Even though precision only improved by two percentage points, we see that sensitivity, i.e. the ability of the model to predict shots made, improved by ten percentage points. We are happy with this improvement. However, one question could come to mind.

4.2.4 Why is the model not perfect?

The highest precision we can reach with this model is 61%. Why can't we do better? Limitations of the model play a huge part. Particularly, the model doesn't reflect an important feature of basketball players, which is talent. It turns out that the model is great at predicting the probability of making a shot of average players, but not for very good (or bad) players.

In Figure 4.4 we compare predicted makes and actual makes of every NBA player during the 2014/2015 season. Predicted makes are calculated with the same procedure of Section 4.2.3. There is a linear relationship between them, but we have some outliers. It's interesting that the players at the top are some of the best players in the league. In fact, each one of them has been selected for the all-star game, where only the best 24 players in the NBA can play, at least once. Let's consider *Chris Paul*. The Los Angeles Clippers point guard made more than 400 hundred baskets, the model only predicted less than 100. What happened? Basketball fans will know that Chris Paul is famous for hitting difficult shots. The model predicted those shots as misses, because the average NBA player would miss them most of the time. Chris Paul is not an average player, he's elite! And his talent let him make tough shots. The same thing can be said for other players such as Stephen Curry, James Harden, and Lebron James, which have multiple MVP awards and will be remembered as some of the best of all time.

5 Playing Strategy

5.1 Not every shot is worth the same

At the beginning we said that not every shot is worth the same amount of points. Shots taken behind the three point line are worth 50% more than those taken inside the arc. We will show how this fact has influenced the playing strategy of teams and players during the last ten NBA seasons. Only for this section, we will not consider three point shots from corners. In fact, a corner 3 point shot in only 22 feet from the basket, while a regular 3 pointer is 23.75 (Figure 5.1)

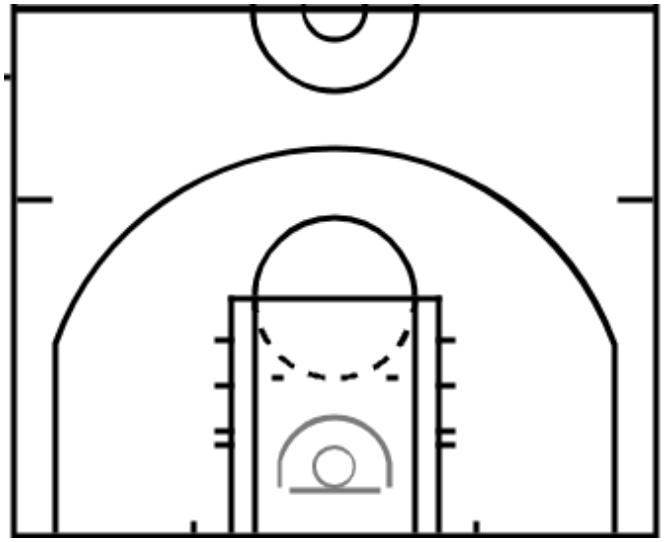

Figure 5.1: NBA court, notice that the three point line is not a perfect arc, but for this section we will not consider corner threes

Given that in our data shot distance is given in the form of integers, we define

$$EP = \text{expected points}$$

Distance	Crunch	Dribbles	Defender Distance	Shot Clock	Result	Probability	Prediction
10	0	0	8	16	made	0.60	1
23	1	5	2	3	made	0.16	0
4	0	0	1	21	missed	0.49	1
25	1	0	5	12	made	0.37	0
						Correct	1/4

Table 4.3: Four random observations and respective prediction. With a threshold of 0.45 the model predicted the result right 1 time out of 4

Distance	Crunch	Dribbles	Defender Distance	Shot Clock	Result	Probability	Prediction
10	0	0	8	16	made	0.60	1
23	1	5	2	3	made	0.16	0
4	0	0	1	21	missed	0.49	0
25	1	0	5	12	made	0.37	0
						Correct	2/4

Table 4.4: Four random observations and respective prediction. With a threshold of 0.45 the model predicted the result right 1 time out of 4

$$EP = \begin{cases} 2 \cdot p(y=1|\vec{x}) & \text{for } 0 \leq \text{distance} \leq 24 \\ 3 \cdot p(y=1|\vec{x}) & \text{otherwise} \end{cases} \tag{5.1}$$

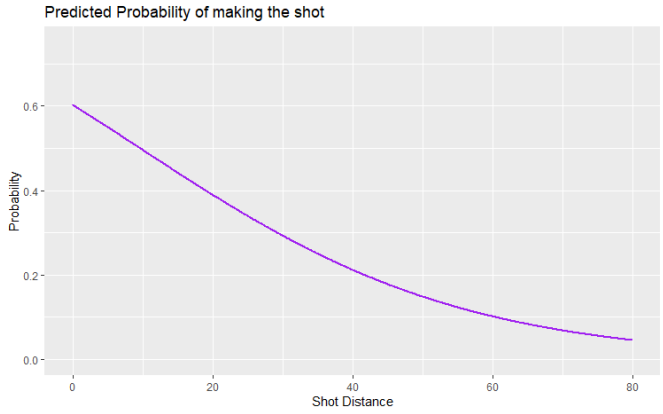

Figure 5.2: Predicted probability as a function of shot distance

In Figure 5.2 we see predicted probability of making the shot in relation to shot distance. If we apply Equation 5.1 we get Figure 5.3. We can tell that at some point shooting a two becomes less efficient than shooting a three. Let's find out when. The model predicts:

$$2 \cdot \frac{1}{1+e^{-\beta_0+\beta_1 \cdot x}} = 3 \cdot \frac{1}{1+e^{-\beta_0+\beta_1 \cdot 24}} \tag{5.2}$$

where β_0 is our constant coefficient and β_1 is the coefficient for shot distance. If we solve this with some algebra, we find that at 7.3 feet from the basket, a two pointer becomes less efficient than a three. For the scope of this section, when we talk about three pointers we intend shots between 23 and 25 feet. After that, shots between 7 and 10 feet might still be more efficient. Furthermore, attempts from 24 feet generate 1.10 points per shot. Shots between 0 and 7.3 feet generate on average 1.13 point per shot, while shots between 7.3 and 22 feet generate on average 0.89 points per shot. Therefore, 0 to 7.3 and 23 to 25 are the most efficient spots to shoot from. Someone could ask why not shooting only in the paint? If a team did that, the defense would adjust and only stay in the

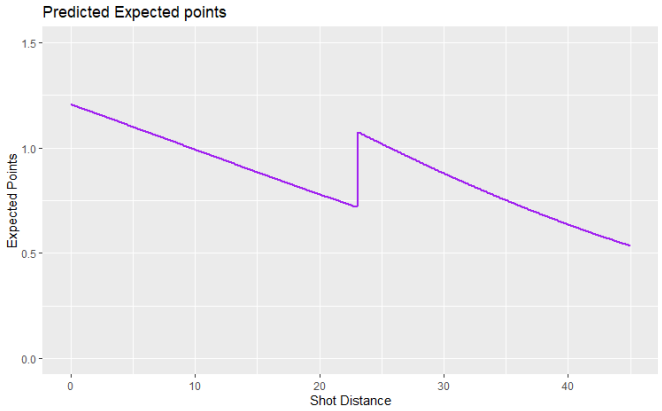

Figure 5.3: Predicted expected points as a function of shot distance

paint. At that point, the distance of closest defender would always be less than 2 feet and we saw from the model how that affects the probability of making the basket. As my coach says, "the fear of the shot sets up the drive, and the fear of the drive sets up the shot". The solution is a right balance between paint shots and threes.

5.2 Change in Behavior

At the beginning we saw that average points scored per game have been increasing, and our hypothesis was that this increase is due to a change in shot selection and playing strategy. In the previous sections we saw that this hypothesis is supported by a mathematical explanation. However, we still have to verify what really happened during the last ten seasons. Remember, we are trying to prove that teams have reacted to the fact that some shots are more efficient than others, in particular that they have been using three point shots more than midranges (10-20 ft).

In Figure 5.4 are shown the shooting rates of the last 10 NBA seasons. We notice a few things. During the 2015-2016 season, three point shooting rate surpassed the mid range shooting rate. At this pace 50% of the shots will be threes by the 2025 NBA season. This boost could be ex-

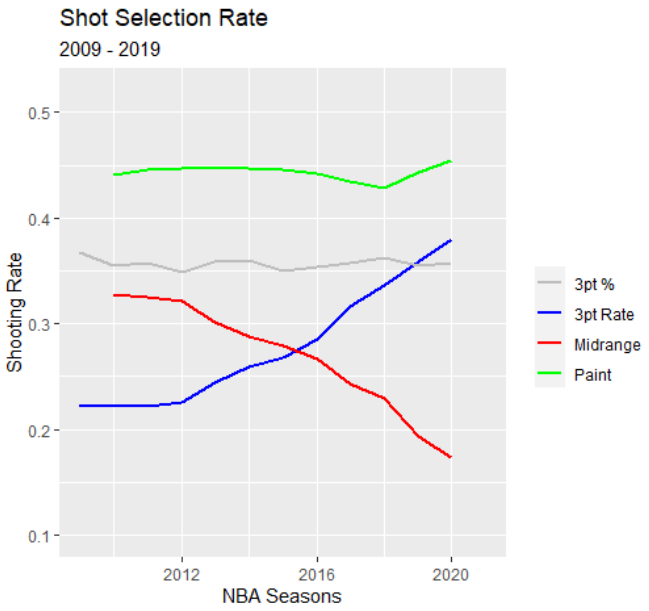

Figure 5.4: caption here

plained by players getting better at shooting threes, but this is not the case. The three point percentage remains constant through the years. Also, notice how paint shots, from 0 to 10 ft, remain constant as well. The only thing that changed is shot selection!

Understanding the change in playing behavior may be easier with Figure 5.5

I invite the reader to start from the left side, with the 2009-2010 season, and follow the bellies of the violin plot, which become smaller and smaller at the mid range level.

5.3 Real Life Examples

5.3.1 Team

When analyzing this shift in shot selection I thought it could be useful to have a tool that lets us visualize the shot selection of different seasons, teams and players. That's why I developed a second application. We already saw two examples of this app in action at the beginning of the paper in Figure 1.2 but now let's focus on teams and players. Note that the following are screenshots taken from the app, and I invite the reader to play around with it and search their favorite player (link).

Let's consider the Houston Rockets and compare the 2009-10 and 2018-2019 seasons. We already explained how the heatmap works in Section 1.2, we divide the court in bins of size 1 foot squared, and count how many shots are taken inside that bin. In Figure 5.6 we have the 2009-10 Houston Rockets. Notice how the midrange shot is still used in abundance. The Rockets took 1819 midrange shots that season on a total of 8945 shots, 20% of the shots. Now let's take a look at the 2018-19 Rockets in Figure 5.7. It looks like every mid range shot from 10 year ago now is taken either in the paint or outside the three point line. The Rockets shot 743 midranges on a total of 10354 shots, 7% of the shots. In addition, the reason why the Rockets took 16% more shots ten years later is beyond the scope of this research, but they could be accelerating their offense to shoot more quickly before the defense has a chance to get set. This could be explained by the fact that an early shot clock has a positive relationship with the logit of making a shot, as shown by the coefficient in table 1.3 and by the shot clock effect plot in Figure 7.1

5.3.2 Players

We know that teams have changed their shot selection, and of course this transition has to start from the individual players. Let's keep using our application and analyze how Stephen Curry, a two times MVP and three times NBA champion, has changed his shot selection through the years. We compare his rookie season in 2009-10 to his last championship season in 2017-2018 using density plots that work exactly as the violin plots we saw in Figure 5.5 In 2009 (Figure 5.8Stephen Curry was fresh out of college, his playing style was not evolved yet and he was taking a considerable number of mid range shots: 458 out of 1343 total shots or 34% . During his most recent championship campaign (Figure 5.9) Curry shot only 316 midranges out of 2215 shots, which makes 14% . His 3 point shooting rate increased by almost 50% as he almost completely eliminated shots between 10 and 20 feet.

However, there are some exceptions. In Section 5.1 we showed that shots taken outside 7.3 feet become less efficient, but we made this assumption by considering the predicted shooting percentage from our logit model. There are players that are so good at shooting midranges that they do not have to follow our model to be efficient. Let's take for example Kevin Durant, arguably the best player in the world (Figure 5.10).

Durant shot an astonishing 52% between 10 and 20 feet last season, compared to the 41% league average. This means that on average Durant generates 1.04 points per shot from the midrange area, which is far above the rest of the league (see Figure 5.3). Knowing that Durant shot 35% from three, which makes on average 1.05 points per shot, we could argue that he could only take midrange shots and maintain the same level of efficiency.

To put this into perspective, these numbers are close to Michael Jordan most efficient season ever. During his career, Jordan won 5 MVP awards and 6 championships. He averaged 30.12 points per game, which makes him the most prolific scorer (per game) ever, and he was part of the 1996-97 Chicago Bulls team which won 72 games. Unfortunately, I was unable to get data from that season.

In Figure 5.11 we have a shooting map made by NBA analyst and best-selling author Kirk Goldsberry. Michael Jordan's was so good at every aspect of the game that no model could describe what he used to do. Almost 60% of his shots were midranges, which is almost unbelievable compared to the 2019 Houston Rockets 7% . It would be interesting to see how Jordan's playing style would adapt in today's playing style. My guess is that he would continue to be one of the greatest players of all time and be one of the exceptions, together with Kevin Durant and other elite players.

5.4 Practice what you preach

After my freshman year of college basketball Seattle University adopted a completely different playing style. My coach recognized the efficiency of certain shots. Every single day we practiced taking only great shots, either close to the basket or outside the three point line. We believed so much in this that we got to a point where players taking midrange

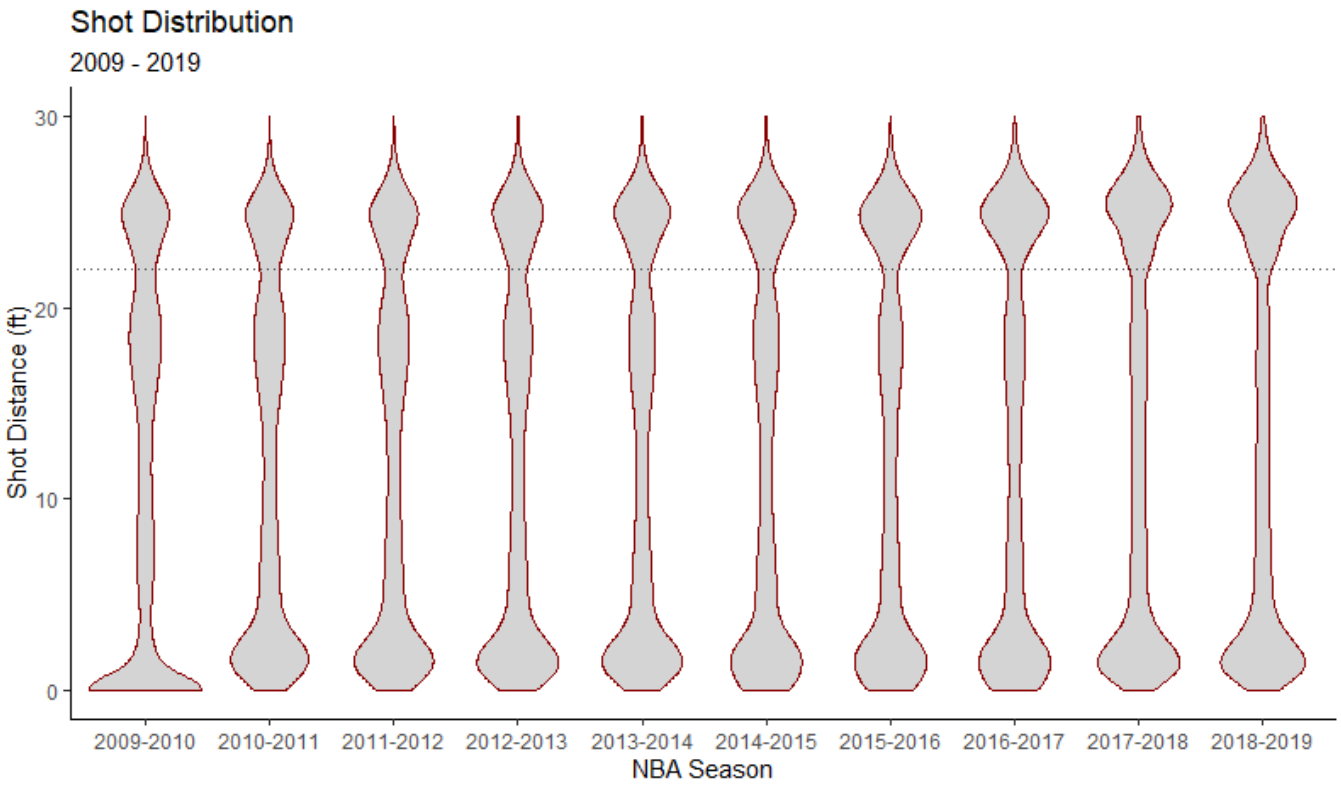

Figure 5.5: Violin plot of each season from 2009-10 to 2018-2019.

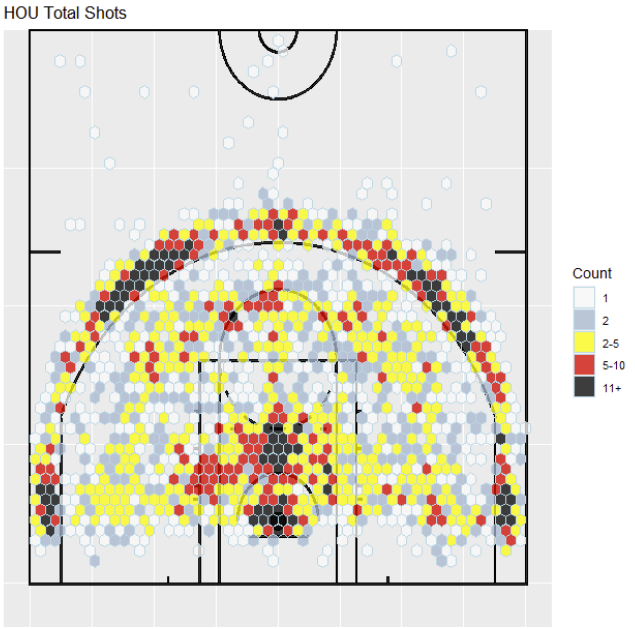

Figure 5.6: Houston Rockets Season 2009-2010

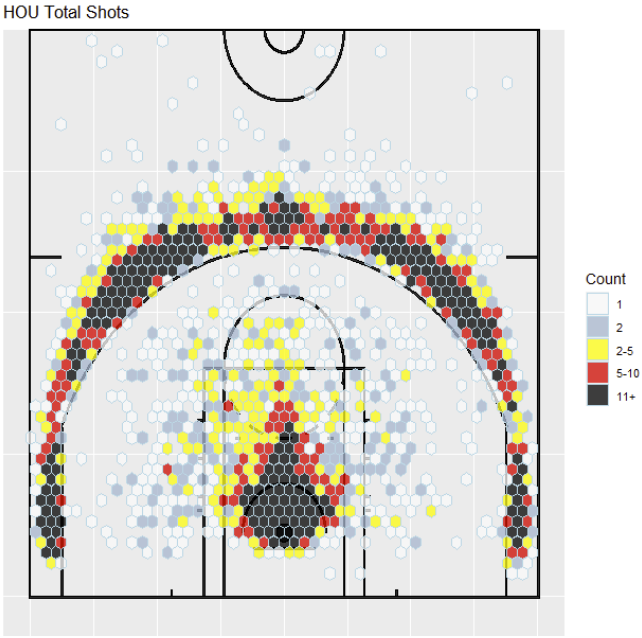

Figure 5.7: Houston Rockets Season 2018-2019

shots were substituted because they were not following the team philosophy. In Figure 5.12 we see Seattle U shot selection during the 2019-20 season. It looks similar to the Houston Rockets in Figure 5.7 where there are very few mid range shots compared to paint shots and three pointers. Figure 5.13 shows my personal shot selection. This almost extreme playing strategy increased my points scored per game by 60%. Many other factors could have helped my scoring ability, but this is an incredible result. It may look that during the season I was doing mathematical research to find the most efficient shots!

6 Conclusion

The model presented today comes with some limitations. As briefly mentioned before, it does not take in account players' talent. This could be solved by factoring and making each player a dummy variable. I tried but having hundreds of names did not help, and also many of the coefficients be-

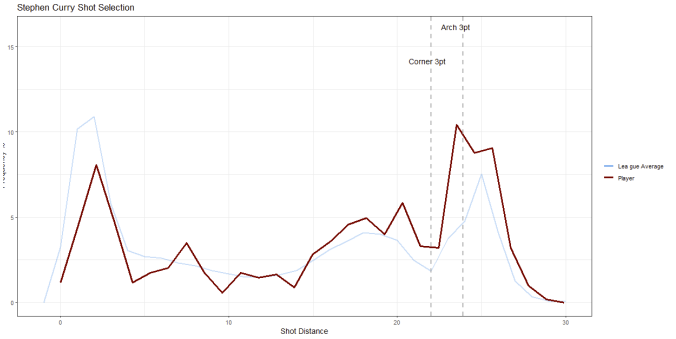

Figure 5.8: Stephen Curry Season 2009-2010.

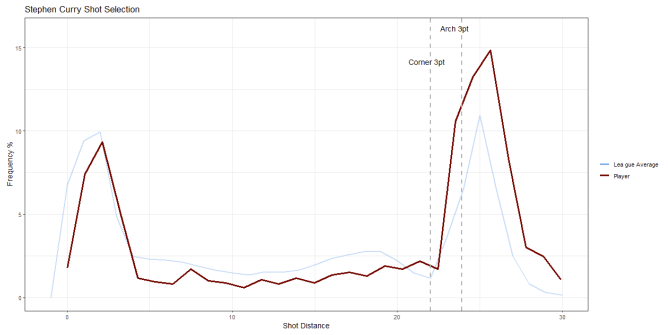

Figure 5.9: Stephen Curry Season 2016-2017

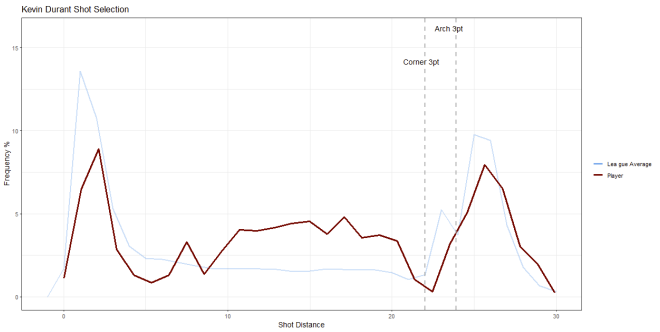

Figure 5.10: Kevin Durant Season 2018-2019

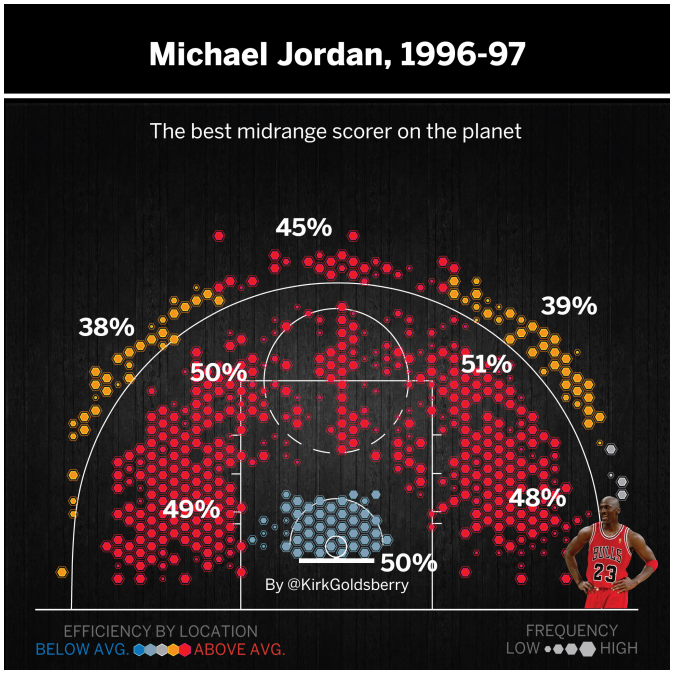

Figure 5.11: Houston Rockets Season 2018-2019

came not statistically significant. We could consider just a few players, but then the problem is choosing who and how many. A solution could be to create just one categorical variable, that takes the value of 1 if the player is in the top 25 and 0 if he's not.

Further, another issue is that we don't account for the type of shot, for example if it was a jump shot, a hook shot, a dunk, etc. When this was tried the model started having statistical significance issues, and I preferred to stick with the variables we encountered in the paper. I believe all this could be great material for future research, together with many topics I did not have the time (or the space) to cover. Just to give a few examples, we did not discuss in depth the effects of a faster pace on the game. Teams are shooting more threes and playing faster, and this seems to be a deadly combination that will lead to a steady increase in average points scored per game. In addition, proving that shots outcome are independent from each other and the 'hot hand phenomenon' does not exist would give me enough material for a second Senior Research. Also, we haven't covered the exciting topic of late game decisions. We saw that teams tend to shoot more twos when down two points at the end of games, but we stopped there. An extensive research on this particular aspect of the game is something I would love to do in the future.

7 Additional Material

In this section we provide extra material and articles in case the reader wants to expand covered covered in the paper. In Figure 7.1 we have the effects of shooting an early shot, regular shot, or late shot. Notice that from 0 to around 25 feet, a shot in the first 6 seconds of the shot-clock is more likely to go in. This is confirmed by the coefficient for early shot, which is the only one that has a positive relationship with the logit of making the shot.

In 7.2 we have the effect plot of two variables: defender distance and number of dribbles. The best combination we have is a shot from under the basket, with zero dribbles and the nearest defender at more than ten feet. However, we are still not certain that the shot will go in. Even the best players on the planet will miss the easiest shots in basketball sometimes. To conclude, I wanted to say that if the reader has any doubts, questions, or comments to please email me at my school address. I'll be more than happy to discuss my work with you.

One thing I wanted to share is how I managed to draw a basketball court in R. The process is incredibly complex, and I wouldn't have been able to do it without the help of data analyst *Ewen Gallic*. The reader can find his guide here [4].

Even though I coded my apps from scratch, I got my inspiration from a different NBA Shot Visualization tool programmed by Peter Beshai. It's way more advanced than mine and the reader can find here [6].

As mentioned in the paper, I learned logistic models in my Econometrics class taught by professor Hiedemann. The course is part of the Ecnomics studies, but I strongly reccommend it as elective even for math majors.

Finally, for anyone who's interested in learning more about

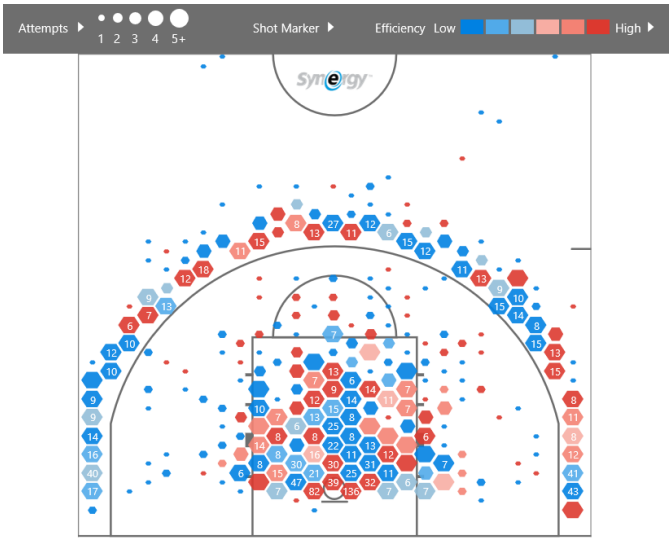

Figure 5.12: Seattle University Season 2019-2020

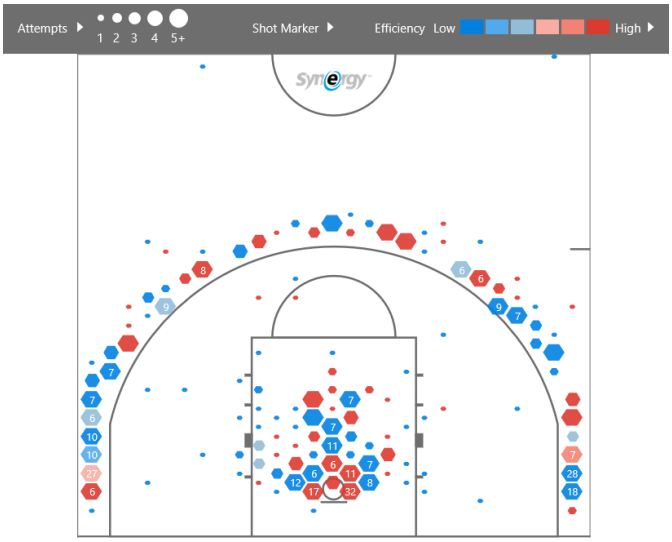

Figure 5.13: Mattia Da Campo (Seattle University) Season 2019-2020

R squared in non-linear regressions, I recommend looking into this guide made by the UCLA statistics department. [9]

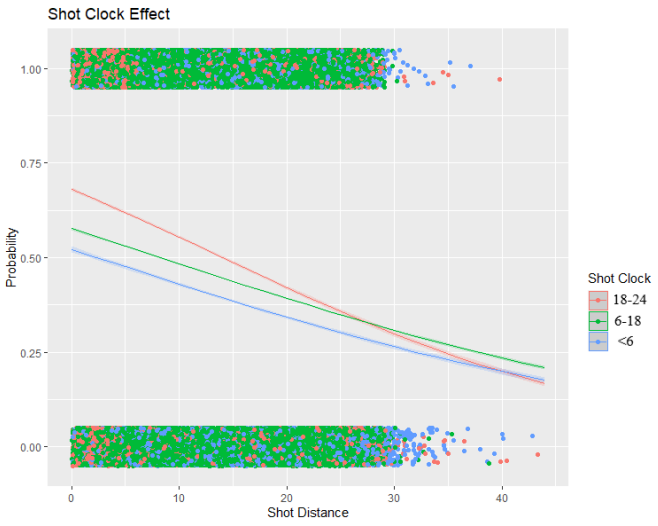

Figure 7.1: Shot Clock effects on the probability of making the shot.

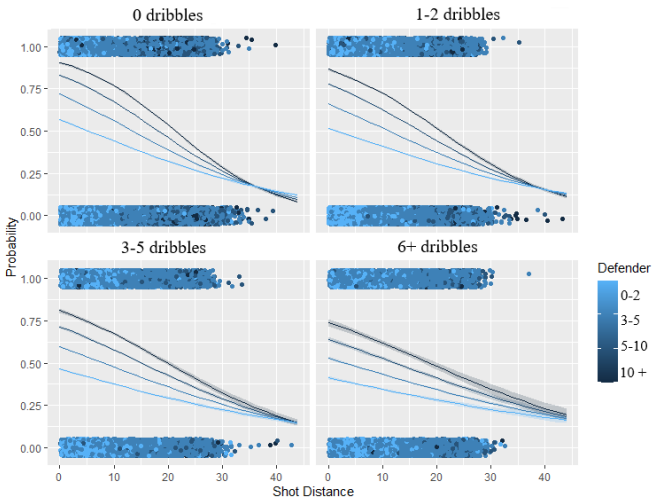

Figure 7.2: Defensive coverage and number of dribbles effects.

References

[1] The probability app can be found at https://mattiadacampo.shinyapps.io/Basketball1/

The shot charts app can be found at https://mattiadacampo.shinyapps.io/ChartsFinal/

[2] Historical & In-Season NBA Data in Excel [Reliable+Enriched]. *BigDataBall*. Retrieved 8 November 2019, from https://www.bigdataball.com/datasets/nba/?ap_id=nbastuffer

[3] De Saá Guerra, Y., Martín Gonzalez, J., García-Manso, J. (2013). Basketball scoring in NBA games: An example of complexity. *Journal Of Systems Science And Complexity*, 26(1), 94-103. https://doi.org/10.1007/s11424-013-2282-3

[4] Gallic, E. (2014). *Drawing a basketball court with R*– Ewen Gallic. Egallic.fr. Retrieved from https://egallic.fr/en/drawing-a-basketball-court-with-r/

[5] Friendly, M. (2015). *Visualizing GLMs for binary outcomes*. Ddar.datavis.ca. Retrieved from http://ddar.datavis.ca/pages/extra/titanic-glm-ex.pdf

[6] Beshai, P. (2015). Buckets: *NBA Shot Visualization*. Buckets.peterbeshai.com. Retrieved from https://buckets.peterbeshai.com/app/#/compareView/x,x?playerSelector=true&selectedPlayer=5

[7] Analytics,M.(2015). *Evaluating Logistic Regression Models*. R-bloggers. Retrieved from https://www.r-bloggers.com/evaluating-logistic-regression-models/

[8] McFadden, D. (1977). *Quantitative Methods for Analyzing Travel Behavior*. Footnotes page 35. Cowles.yale.edu. Retrieved from http://cowles.yale.edu/sites/default/files/files/pub/d04/d0474.pdf

[9] *What are pseudo R-squareds?*. Statistics.ats.ucla.edu. (2011). Retrieved from http://statistics.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm

[10] Peng, C., Lee, K., Ingersoll, G. (2002). *An Introduction to Logistic Regression Analysis and Reporting*. The Journal Of Educational Research, 96(1), 3-14. https://doi.org/10.1080/00220670209598786

[11] Wilks, S. (1938). *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*. The Annals Of Mathematical Statistics, 9(1), 60-62. https://doi.org/10.1214/aoms/1177732360

	Dependent variable:			
	FGM			
	(1)	OR	(3)	OR
SHOT DISTANCE	-0.043^{***}	0.958	-0.061^{***}	0.941
	(0.001)			
CRUNCH	-0.112^{***}	0.894	-0.085^{**}	0.919
	(0.037)			
EARLY SHOT			0.194^{***}	1.215
			(0.016)	
LATE SHOT			-0.170^{***}	0.844
			(0.017)	
OPEN			-0.117^{***}	0.890
			(0.036)	
CONTESTED			-0.389^{***}	0.678
			(0.034)	
FORCED			-0.880^{***}	0.414
			(0.038)	
1-2 DRIBBLES			-0.272^{***}	0.762
			(0.015)	
3-5 DRIBBLES			-0.318^{***}	0.728
			(0.019)	
6+ DRIBBLES			-0.286^{***}	0.752
			(0.020)	
Constant	0.406^{***}		1.221^{***}	
	(0.011)		(0.038)	
Observations	119,320	119,320	119,320	119,320
Log Likelihood	−80,177.650	−80,177.650	−78,917.420	−78,917.420
McFadden R^2	0.338		0.546	
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 7.1: Reduced and Full model. This table is discussed in the sections above