Modeling Shot Efficiency in the NBA

Stat guys: Lewis Eatherton, Chris Yang, Charlie Bonetti

10/28/20

Your written report goes here! Before you submit, make sure your code chunks are turned off with `echo = FALSE` and there are no warnings or messages with `warning = FALSE` and `message = FALSE`

Introduction

For our STA 210 Final Project, our group is interested in investigating how certain factors may have an influence on a basketball players' shot efficiency during a game in the National Basketball Association (NBA). We are all avid NBA fans and express great curiosity in what makes a good shot versus a bad shot. This is an exciting opportunity to explore the presence or absence of certain variables which surprisingly, or unsurprisingly, have an effect on shot success.

Our interest in this subject matter was piqued by an 2015 article titled "Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball" published by Frane Erčulj and Erik Štrumbel, which looked into how different shot types affected shot success across different levels of competition. It was insightful to learn that there were no discernable differences between different situational variables (e.g. player type, shot location) on shot type and shot success between levels. Since the effect of situation variables on shot success seemed to be constant throughout all levels of competition, we became motivated to identify what exactly the situation variables might be, and to what degree of influence they had on an individual's shot.

As a result, we formulated our research question: "Do certain situational variables during a game have an effect on an NBA player's shot success?" Based on our prior knowledge of the game of basketball, we recognize that both players and coaches deem certain shots as "good shots" and some as "bad". Yet, what exactly characterizes these shots to fall under those two categories? Furthermore, through our own intuition after watching countless games and playing the sport ourselves, we hypothesize that some situational variables (e.g. shot distance or the distance of the closest defender) will have a greater effect on an NBA player's shot success than other situational variables. It seems intuitive as the time left on the clock and distance to the defender decrease, the likelihood of making a shot increases. Seeking clarity in how these factors contribute to a successful shot are the key objectives of our analysis.

Data

In order to identify answers to our questions, we will explore data from the 2014-2015 National Basketball Association (NBA) season. The data was originally collected via NBA's rest API, which scrapes historical data of past games off of the NBA website. All of the data was either recorded by player tracking via computer vision technique or by manual entry at an NBA scoring table.

Each observation in our data set is a unique shot that was taken in an NBA basketball game during the 2014-2015 season. Features of each observation include qualities such as the game location, matchup, player name, and shot result. There are a total of 128,069 observations within the data set. The response variable we want to record is shot percentage. This value determines the likelihood of a successful shot when all predictor values are taken into consideration.

The response variable we are interested in investigating is:

- `FGM`: Field Goal Made; whether the shot went in or not

The predictor variables that we expect to have an influence on the response variable are:

- `Shot_number`: The shot number for that player in that specific game (8 would mean it was the player's 8th shot in that game)
- `Shot_clock (sec)`: How many seconds were left on the shot clock when the player shot the basketball
- `Shot_distance (ft)`: How far away the player was from the basketball when shooting
- `Dribbles`: The number of dribbles the player took before shooting
- `Touch_time`: How long the player touched the ball for before shooting (after being passed to)
- `PTS_TYPE`: Whether the shot was 2 or 3-pointer
- `Closest_Defender`: Who the closest defender was to the shooting player
- `Close_Def_Distance (ft)`: How far away the defender was from the player when the ball was shot
- `Player_name`: Who shot the ball

```
Rows: 128,069
Columns: 21
$ GAME_ID   <dbl> 21400899, 21400899, 21400899, 21400899, ...
$ MATCHUP   <chr> "MAR 04, 2015 - CHA @ BKN", "MAR 04, 201...
$ LOCATION   <chr> "A", "A", "A", "A", "A", "A", "A", "A", ...
$ W   <chr> "W", "W", "W", "W", "W", "W", "W", "W", ...
$ FINAL_MARGIN   <dbl> 24, 24, 24, 24, 24, 24, 24, 24, 24, 1, 1...
$ SHOT_NUMBER   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 2, 3, 4, 1...
$ PERIOD   <dbl> 1, 1, 1, 2, 2, 2, 4, 4, 4, 2, 2, 4, 4, 4...
$ GAME_CLOCK   <time> 01:09:00, 00:14:00, 00:00:00, 11:47:00,...
$ SHOT_CLOCK   <dbl> 10.8, 3.4, NA, 10.3, 10.9, 9.1, 14.5, 3....
$ DRIBBLES   <dbl> 2, 0, 3, 2, 2, 2, 11, 3, 0, 0, 8, 14, 2,...
$ TOUCH_TIME   <dbl> 1.9, 0.8, 2.7, 1.9, 2.7, 4.4, 9.0, 2.5, ...
$ SHOT_DIST   <dbl> 7.7, 28.2, 10.1, 17.2, 3.7, 18.4, 20.7, ...
$ PTS_TYPE   <dbl> 2, 3, 2, 2, 2, 2, 2, 2, 3, 3, 3, 2, 2, 3...
$ SHOT_RESULT   <chr> "made", "missed", "missed", "missed", "m...
$ CLOSEST_DEFENDER   <chr> "Anderson, Alan", "Bogdanovic, Bojan", "...
$ CLOSEST_DEFENDER_PLAYER_ID <dbl> 101187, 202711, 202711, 203900, 201152, ...
$ CLOSE_DEF_DIST   <dbl> 1.3, 6.1, 0.9, 3.4, 1.1, 2.6, 6.1, 2.1, ...
$ FGM   <dbl> 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0...
$ PTS   <dbl> 2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, 2, 0...
$ player_name   <chr> "brian roberts", "brian roberts", "brian...
$ player_id   <dbl> 203148, 203148, 203148, 203148, 203148, ...
```
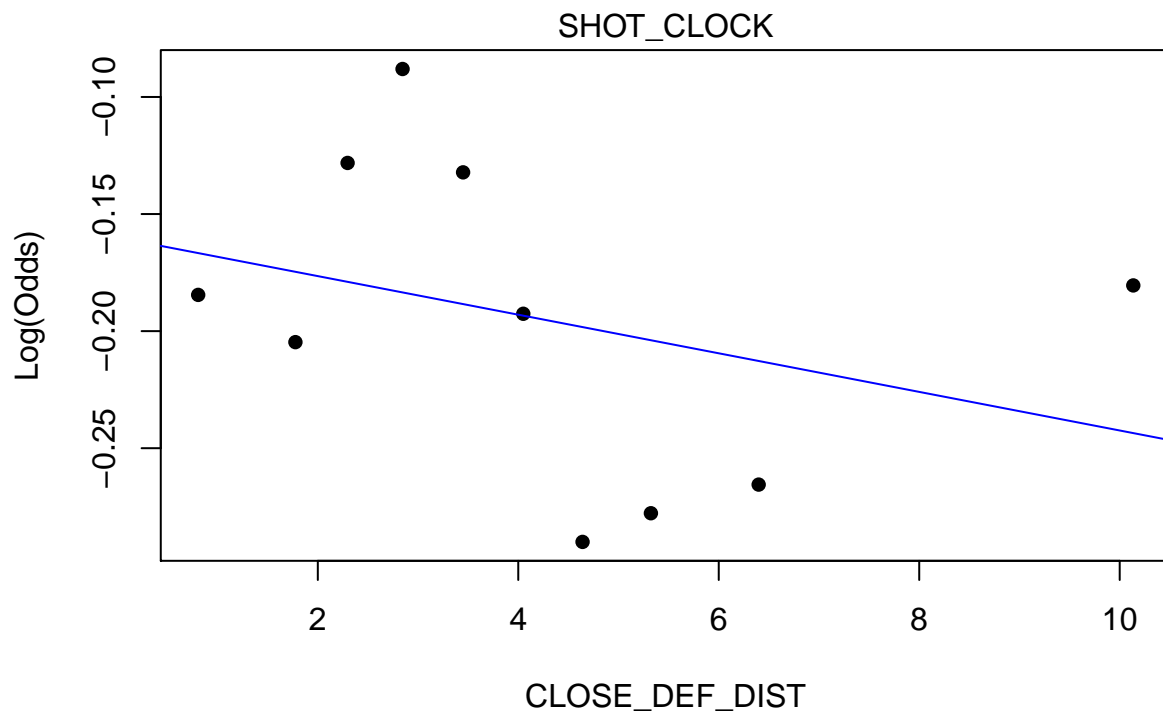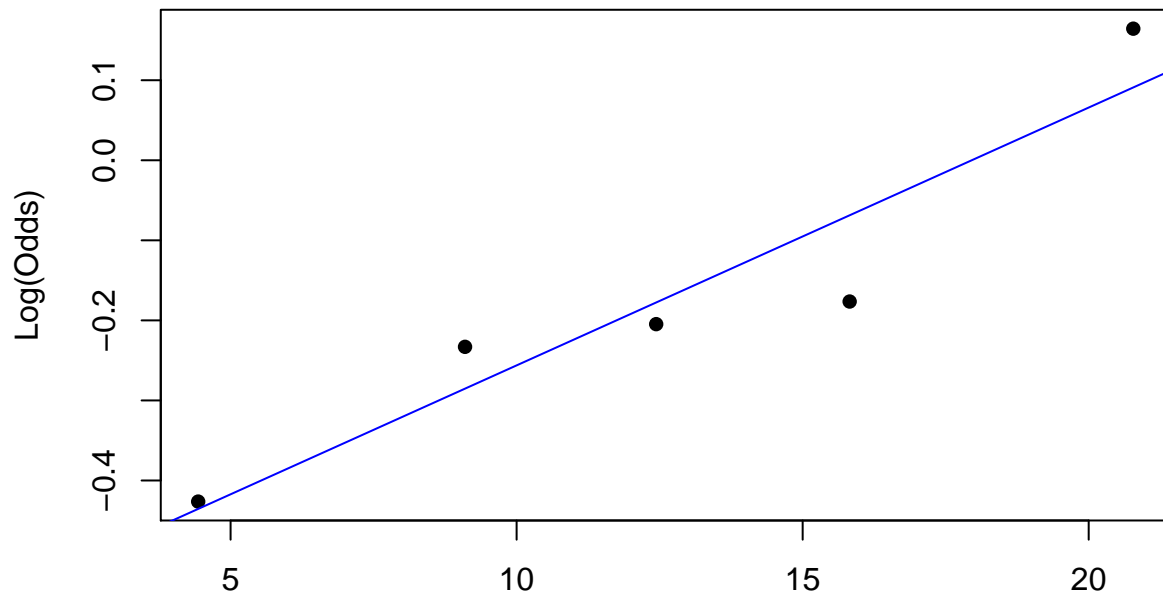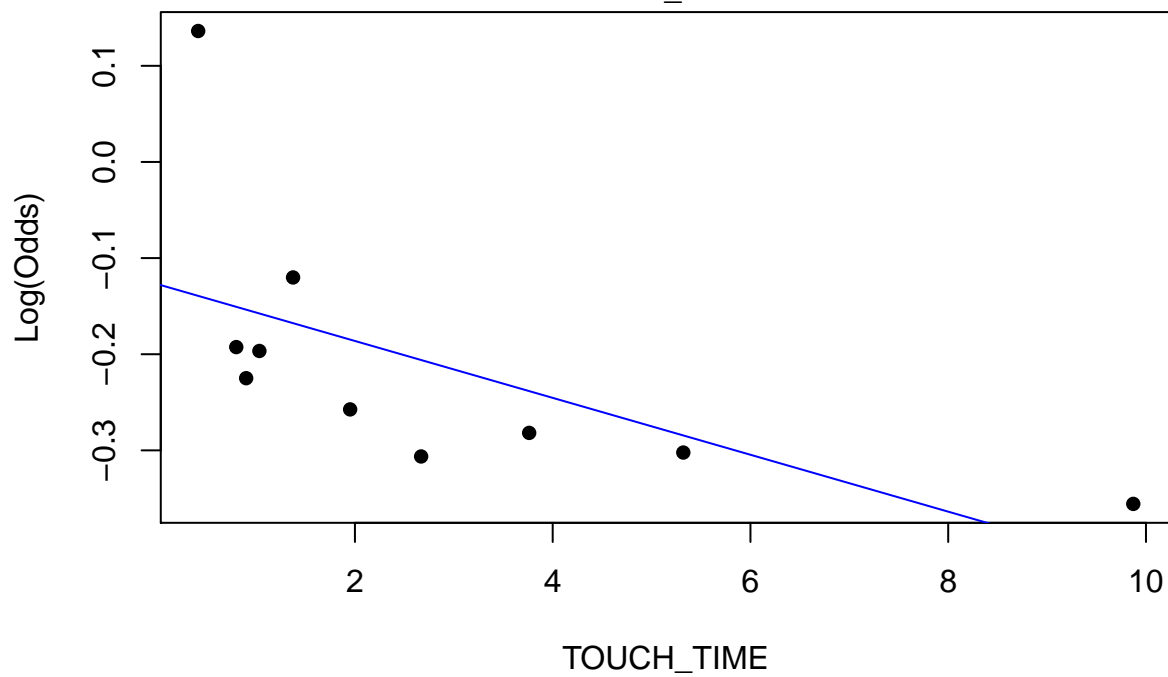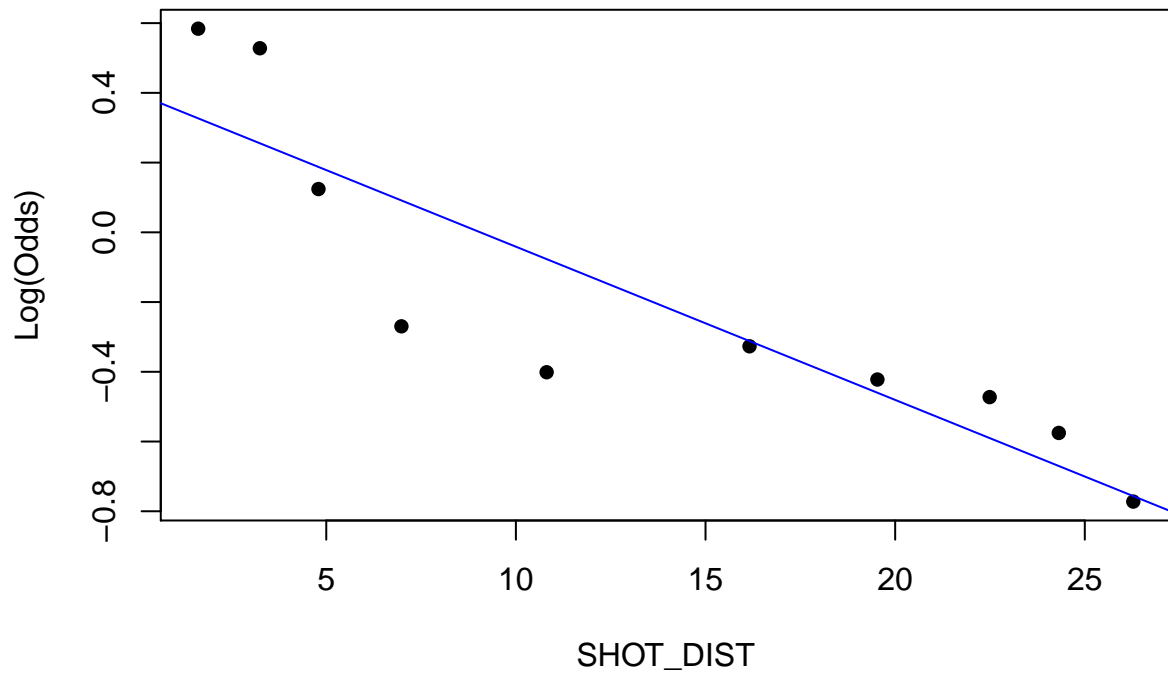
```
## # A tibble: 470 x 5
## # Groups:   CLOSEST_DEFENDER [470]
##    CLOSEST_DEFENDER  Make      n  prop emp_logit
##    <chr>             <fct> <int> <dbl>     <dbl>
## 1 Acy, Quincy       1       118 0.428    -0.292
## 2 Adams, Jordan     1        16 0.533     0.134
## 3 Adams, Steven     1       215 0.444    -0.224
## 4 Adrien, Jeff      1        40 0.548     0.192
## 5 Afflalo, Arron    1       191 0.417    -0.335
## 6 Ajinca, Alexis    1       114 0.465    -0.139
## 7 Aldemir, Furkan   1        33 0.465    -0.141
```

4

```
##  8 Aldrich, Cole       1         142 0.532        0.128
##  9 Aldridge, LaMarcus 1         302 0.461       -0.156
## 10 Allen, Lavoy        1         141 0.449       -0.205
## # ... with 460 more rows
```


CLOSE_DEF_DIST
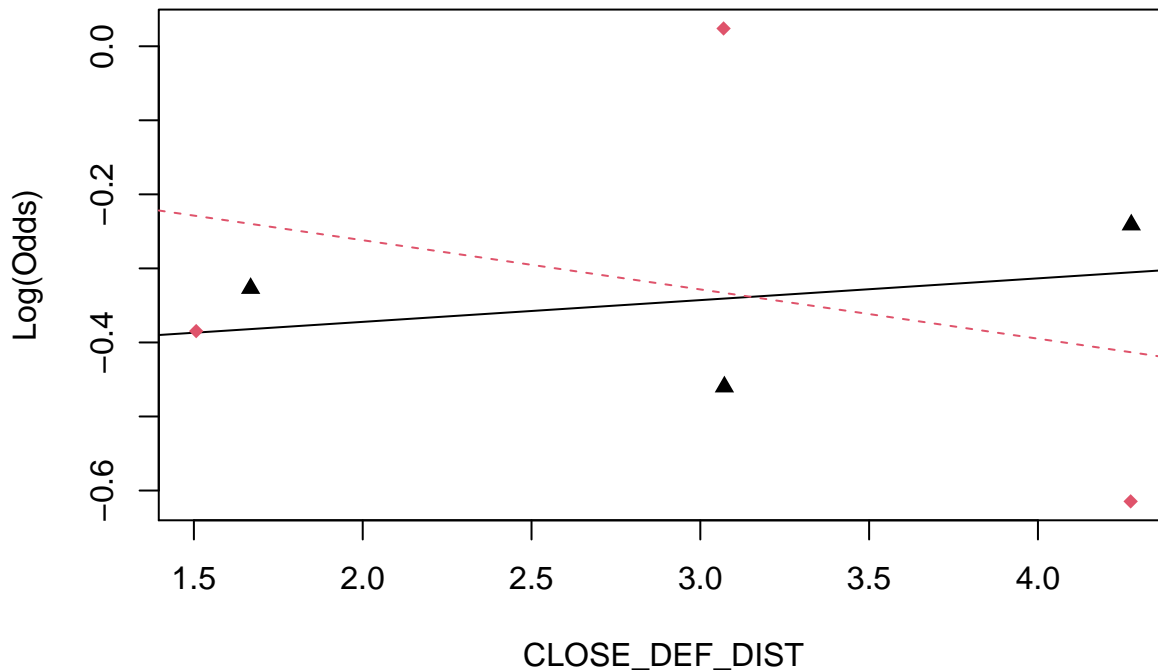
Linearity is not totally satisfied... graphs aren't linear

independence and randomness seems fine accorind to how data was collected

we should also look into collinearity

Creating Model

```
## Single term deletions
##
## Model:
## Make ~ SHOT_CLOCK + DRIBBLES + TOUCH_TIME + SHOT_DIST + CLOSE_DEF_DIST
##               Df Deviance    AIC
## <none>            162325 162337
## SHOT_CLOCK     1   162532 162542
## DRIBBLES       1   162354 162364
## TOUCH_TIME     1   162434 162444
## SHOT_DIST      1   167561 167571
## CLOSE_DEF_DIST 1   163764 163774

## Single term deletions
##
## Model:
## Make ~ SHOT_CLOCK + DRIBBLES + TOUCH_TIME + SHOT_DIST + CLOSE_DEF_DIST +
##     CLOSEST_DEFENDER
##               Df Deviance    AIC
## <none>            855.93 869.93
## SHOT_CLOCK     1   858.72 870.72
## DRIBBLES       1   855.93 867.93
## TOUCH_TIME     1   855.94 867.94
```

5

```
## SHOT_DIST         1    872.34 884.34
## CLOSE_DEF_DIST    1    865.25 877.25
## CLOSEST_DEFENDER  1    855.98 867.98

## Single term deletions
##
## Model:
## Make ~ SHOT_CLOCK + TOUCH_TIME + SHOT_DIST + CLOSE_DEF_DIST +
##       CLOSEST_DEFENDER
##                  Df Deviance    AIC
## <none>              855.93 867.93
## SHOT_CLOCK        1  858.81 868.81
## TOUCH_TIME        1  855.99 865.99
## SHOT_DIST         1  872.34 882.34
## CLOSE_DEF_DIST    1  865.25 875.25
## CLOSEST_DEFENDER  1  855.98 865.98

## Single term deletions
##
## Model:
## Make ~ SHOT_CLOCK + TOUCH_TIME + SHOT_DIST + CLOSE_DEF_DIST
##                Df Deviance    AIC
## <none>            855.98 865.98
## SHOT_CLOCK      1  858.90 866.90
## TOUCH_TIME      1  856.02 864.02
## SHOT_DIST       1  872.45 880.45
## CLOSE_DEF_DIST  1  865.36 873.36

## Single term deletions
##
## Model:
## Make ~ SHOT_CLOCK + SHOT_DIST + CLOSE_DEF_DIST
##                Df Deviance    AIC
## <none>            856.02 864.02
## SHOT_CLOCK      1  859.04 865.04
## SHOT_DIST       1  872.60 878.60
## CLOSE_DEF_DIST  1  865.68 871.68

## # A tibble: 2 x 5
##   Resid..Df Resid..Dev    df Deviance p.value
##       <dbl>      <dbl> <dbl>    <dbl>   <dbl>
## 1       645       856.    NA    NA      NA
## 2       643       855.     2   0.813   0.666
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.737	0.294	-2.508	0.012
SHOT_CLOCK	0.026	0.015	1.733	0.083
SHOT_DIST	-0.048	0.012	-4.013	0.000
CLOSE_DEF_DIST	0.243	0.079	3.067	0.002

talk about final model outcome and how we came to it

Discussion

Talk about our results, the limitations of these results, and what'd we do differently...

Work Citation:

(1) Erčulj F, Štrumbelj E (2015) Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball. PLoS ONE 10(6): e0128885. https://doi.org/10.1371/journal.pone.0128885