What Makes Winning Basketball

Regressions 4214 Final Project Marhawe Asmerom

Executive Summary

Statistics is defined as the practice of collecting and analyzing numerical data in large quantities, to help infer greater proportions. Statistics is essential to predict what can be done in the future based on a numerical set. Within sports, it is used to help win games, especially in basketball. So, throughout the project, I will find the four predictors that will help basketball teams win and based on those predictors which teams should be in the playoffs. I faced issues like multicollinearity which is when there is too high of interaction between the variables. I had to remove insignificant variables from the model which then improved the problem. Removing those variables left me with a linear model that includes field goal percentage, three-point percentage, opponent two-point percentage, and opponent three-point percentage. These four predictors allowed me to calculate the projected win compared to actual wins, the results were pretty accurate based on the plot and R^2. The next thing was the generalized linear model for the four predictors allowed me to see what teams should make the playoffs based on their stats compared to the teams that are currently in playoff position. We find that thirteen of the sixteen teams currently in the playoffs are predicted correctly by the model and from the three that are incorrect, two of them should be in the playoffs but aren't and one team should not be in the playoffs but are in position. In the project, we solve the problem based on the data and conclude that certain predictors have a marginal effect on winning basketball games.

What Makes Winning Basketball

Regressions 4214 Final Project Marhawe Asmerom

Problem Context

Statistics play a huge role in current-day sports. Statistics in sports can determine or predict many things like sports betting, winning games, and evaluating player's performances. Many sports organizations have hired a team of data analysts to help them win games. The more games you win, the more cash inflow is received by the organization. One sport where statistics plays a major factor in basketball. Basketball use statistics for ranges of reasons from seeing which five players play best together on the court, what plays should be executed in late-game scenarios, and most importantly winning championships. Before you win a championship, you go through a regular season of eighty-two games (this year 72 games due to COVID) and based on games won compared to other teams, you have a chance to be part of the postseason. For my project, I am looking at statistics from this current 2020 NBA season to determine what predictors are essential for a team to make the playoffs. Based on these predictors, I will project what teams will make the postseason in the next two weeks. The overview of predictors that will be focused on is current wins, three-point attempts, total field goal percentage, two-point percentage, three-point percentage, opponent three-point percentage, current two-point percentage, assist to turnover ratio, and current playoff seeding. All the percentage stats are per game. Most of the stats came from Basketball Reference.com and I compiled a spreadsheet with the statistics that are essential to the project. I did not find any issues with the source, as it was accurately updated and had information that matched other sources including NBA.com. The one variable I compiled was assisted to turnover ratio which is the number of assists a team has compared to turnovers. Since the playoffs are in two weeks, I used the sixteen teams that are currently in playoff positions as the teams that will make it through. These predictors, I will be looking at each correlation they have with winning, and which predictors can be kicked out of

the model because the p-value is too high, making it not significant. There were no issues with the source, as it was accurately updated and had information that matched other sources including NBA.com. Another key topic of interest is the question of does defense wins games? I will be looking at the shooting percentage for teams compared to the shooting percentage they give up to their opponents and determine which has more of an outcome on winning basketball games.

Data Analysis

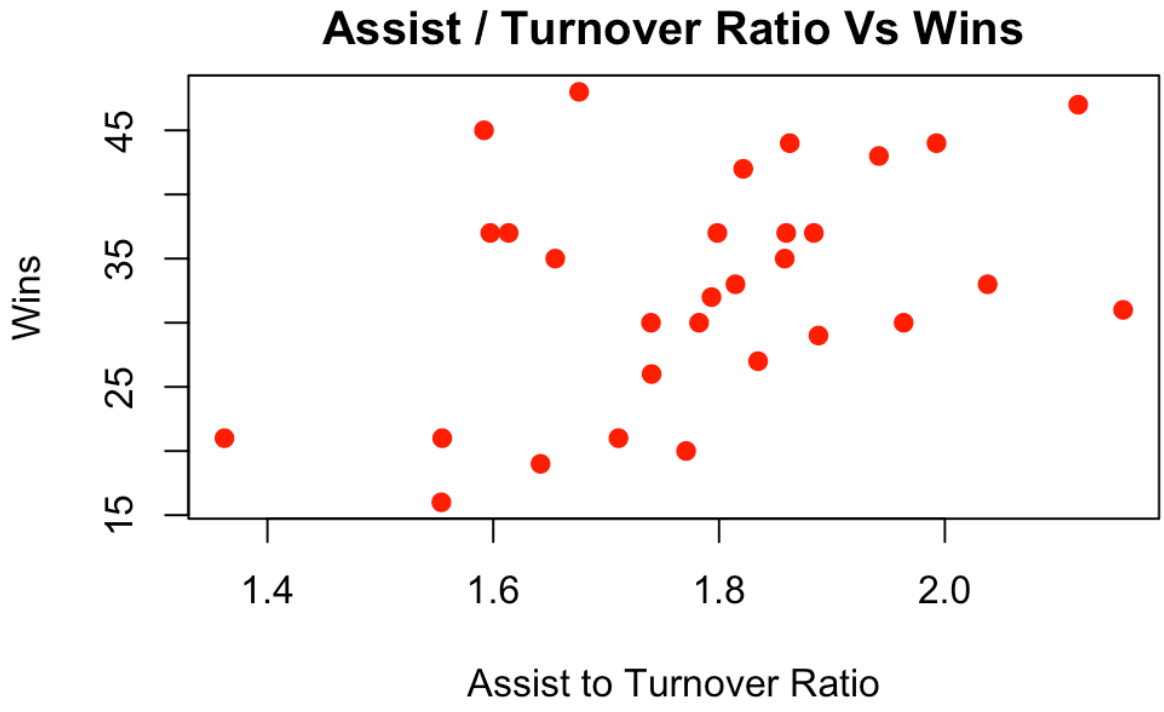
What predictor has the biggest impact on winning a basketball game? The dataset, includes information on two and three-point field goal percentage, opponent two and three-point percentage, three-point shot attempts, assists, turnovers, and games won. The first thing to look at is to create a model to help predict wins based on the predictor's dataset. Now once the model is determined, it is important to look at the summary for a couple of things, the significance of each variable, r – squared value, adjusted r-squared, and the increase of standard deviation. When looking at the significance of each variable in the R Console you notice that a couple of the numbers are greater than 0.05 which means it isn't significant. In your linear regression model, it is important to update the transformation to remove insignificant predictors so the first ones I will remove are three-point attempts and two-point percentage per game. Look at the difference in the outputs.

```
Call:
lm(formula = Wins ~ FG + ThreePt + ThreePA + TwoPt + OppThreePt +
    OppTwoPt + AST_TOV_RATIO, data = NBA)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7332 -2.2655  0.2018  2.2905  7.4119

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -12.00752   57.52703  -0.209   0.8366
FG            114.38658  426.74715   0.268   0.7912
ThreePt       192.32248  187.91725   1.023   0.3172
ThreePA        -0.04744    0.78401  -0.061   0.9523
TwoPt          74.74817  259.07559   0.289   0.7757
OppThreePt   -178.10392   80.50983  -2.212   0.0376 *
OppTwoPt     -122.33046   65.98347  -1.854   0.0772 .
AST_TOV_RATIO    7.45508    5.23379   1.424   0.1684
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.141 on 22 degrees of freedom
Multiple R-squared:  0.8356,    Adjusted R-squared:  0.7833
F-statistic: 15.98 on 7 and 22 DF,  p-value: 2.741e-07
```

```
Call:
lm(formula = Wins ~ FG + ThreePt + OppThreePt + OppTwoPt + AST_TOV_RATIO,
    data = NBA)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7393 -1.7305 -0.2972  2.1976  6.7538

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     -8.507     55.097  -0.154  0.87859
FG             220.103     60.296   3.650  0.00127 **
ThreePt        159.496     53.455   2.984  0.00645 **
OppThreePt    -180.400     69.462  -2.597  0.01581 *
OppTwoPt      -125.119     63.939  -1.957  0.06210 .
AST_TOV_RATIO    7.216      5.041   1.431  0.16520
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.021 on 24 degrees of freedom
Multiple R-squared:  0.8309,    Adjusted R-squared:  0.7957
F-statistic: 23.59 on 5 and 24 DF,  p-value: 1.534e-08
```

Another important note of information you get from the model output is R^2, it tells us how linear the relationship is between the predictors and winning, and since the number is higher than the third quartile, we can conclude the relationship is linear. Adjusted R-squared will tell us the amount of variance covered in the model which is 79.57 %. This shows that the model is fitted well, but not all variables are significant.

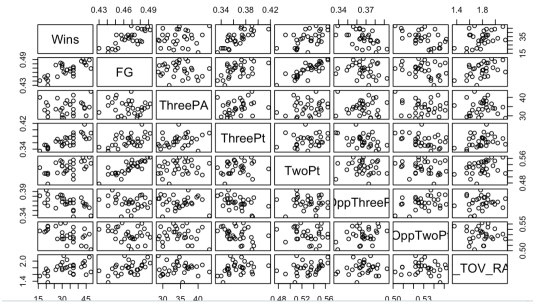
Now I want to focus on Assist / Turnover Ratio, even though it wasn't stat given from the dataset, I felt as if the stat was imperative to team success. An assist is a term that means to help your teammate score, so a direct pass leading to a basket where a turnover is quite the opposite, which is giving the opposing team the possession of the ball. It is likely the higher the assist to turnover ratio is the more wins a team has. Based on the graph below it seems like that assumption is correct, the teams with more wins have higher assist/turnover ratios.



What Makes Winning Basketball

Regressions 4214 Final Project Marhawe Asmerom

Another major issue when working with multiple predictor variables is multicollinearity. Multicollinearity can lead to inflation in the variance and insignificant results. With multicollinearity, it causes regression coefficients to be inestimable. So, what causes multicollinearity issues: high correlation among predictors. At first, you see that I start with seven predictors that can affect winning, which is a lot and it is proven through the R output by having a lot of predictors will cause multicollinearity. If you look at the plot below it shows that there is a high correlation between all the variables, which proves that the model is not fitted correctly. To improve the issue, you remove variables (Three-Point Attempts, Two Percentage) that are insignificant (did earlier in the report just repeating for multicollinearity purposes).

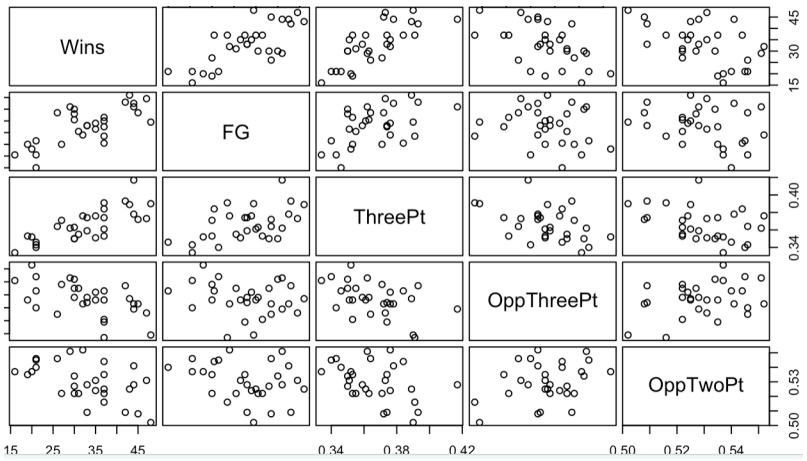



```
Call:
lm(formula = Wins ~ FG + ThreePt + OppThreePt + OppTwoPt, data = NBA)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1885 -1.8992 -0.2626  2.3840  7.3936

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -31.47      53.80  -0.585 0.563884
FG            254.00      56.60   4.487 0.000141 ***
ThreePt       176.07      53.27   3.305 0.002867 **
OppThreePt   -152.02      67.95  -2.237 0.034437 *
OppTwoPt     -118.36      65.09  -1.818 0.080993 .
```

What Makes Winning Basketball

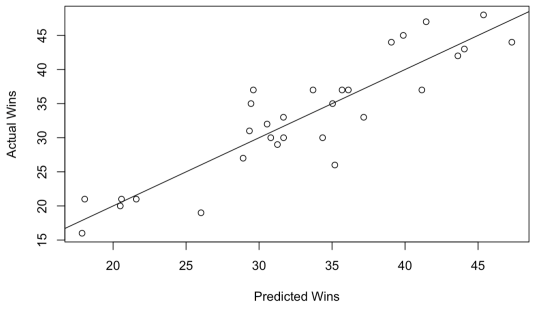
Regressions 4214 Final Project

Marhawe Asmerom



Now that I took out variables that will affect the model including assist to turnover ratio, I will focus on these four predictors – FG, ThreePt, OppThreePt, and OppTwoPt.There is less interaction, correlation (plot above) between the variables and now I can focus on the final part of the project, what teams will make the playoffs, and predicted wins in two weeks.

To find predicted wins, we use our four predictors to make a linear model that shows how many projected wins for the season. We plot both the predicted vs actual wins for the season and look at the relationship at the line. Another thing that will show the strengths between the two (predict vs actual) is R^2 = 0.8165. So, based on the plow below and the R^2, there is a strong association between the two, and it is pretty accurate.



What Makes Winning Basketball

Regressions 4214 Final Project Marhawe Asmerom

To find what teams make the playoffs, we use a generalized linear model of the four predictors (FG, ThreePt, OppThreePt, and OppTwoPt). The playoffs have 16 teams / 30 make it, so you either make it or you don't, hence it being binary. Now, we should compare two categories, the teams that are currently in the playoff picture (top 16 teams) and teams that should be in the playoffs based on the four predictors. The table shows out of the fourteen that shouldn't make it, two of them should be in the postseason. It also shows that out of the sixteen teams, that fifteen of them should be in the playoffs while one of them shouldn't.

		PlayoffsReal	PlayoffsEstimates
1	"Atlanta Hawks"	"Yes"	"Yes"
2	"Boston Celtics"	"Yes"	"Yes"
3	"Brooklyn Nets*"	"Yes"	"Yes"
4	"Charlotte Hornets"	"Yes"	"Yes"
5	"Chicago Bulls"	"No"	"Yes"
6	"Cleveland Cavaliers"	"No"	"No"
7	"Dallas Mavericks"	"Yes"	"Yes"
8	"Denver Nuggets*"	"Yes"	"Yes"
9	"Detroit Pistons"	"No"	"No"
10	"Golden State Warriors"	"Yes"	"Yes"
11	"Houston Rockets"	"No"	"No"
12	"Indiana Pacers"	"No"	"Yes"
13	"Los Angeles Clippers*"	"Yes"	"Yes"
14	"Los Angeles Lakers"	"Yes"	"Yes"
15	"Memphis Grizzlies"	"No"	"No"
16	"Miami Heat"	"Yes"	"No"
17	"Milwaukee Bucks"	"Yes"	"Yes"
18	"Minnesota Timberwolves"	"No"	"No"
19	"New Orleans Pelicans"	"No"	"No"
20	"New York Knicks"	"Yes"	"Yes"
21	"Oklahoma City Thunder"	"No"	"No"
22	"Orlando Magic"	"No"	"No"
23	"Philadelphia 76ers*"	"Yes"	"Yes"
24	"Phoenix Suns*"	"Yes"	"Yes"
25	"Portland Trail Blazers"	"Yes"	"Yes"
26	"Sacramento Kings"	"No"	"No"
27	"San Antonio Spurs"	"No"	"No"
28	"Toronto Raptors"	"No"	"No"
29	"Utah Jazz*"	"Yes"	"Yes"
30	"Washington Wizards"	"No"	"No"

What Makes Winning Basketball

Regressions 4214 Final Project Marhawe Asmerom

Conclusion

As you can see from the results, they seem pretty accurate for predicted wins and what teams should make the playoffs. We found what predictors from the seven we started with are most important to team's success (field goal percentage, three-point percentage, opponent two percentage, and opponent three-point percentage). It is important to note that even though assist/turnover ratio caused multicollearity, there seemed to be some positive trend with winning and a higher ratio. It seems like field goal percentage was the most significant and had the most impact on teams winning games. Surrounding your team with people that shoot the three pointer is imperative to win in the NBA. Another thing the data seems to show is that playing defense matters, as two of the final predictors are opponents two-point and three-point percentage. Holding teams to a lower field goal percentage will help lead you to the postseason. My model predicted correctly thirteen / sixteen teams to make the postseason, the one that the model got incorrect were Miami Heat, which are barely making the playoffs now; Indiana Pacers, who are a game away from making the playoffs; and lastly the Chicago Bulls who are playing well but not enough to win games. Statistics especially regressions help predict many things in basketball, and it can help teams win if they use the right metrics. Good metrics and solid statistics is what makes winning basketball.

What Makes Winning Basketball

Regressions 4214 Final Project — Marhawe Asmerom

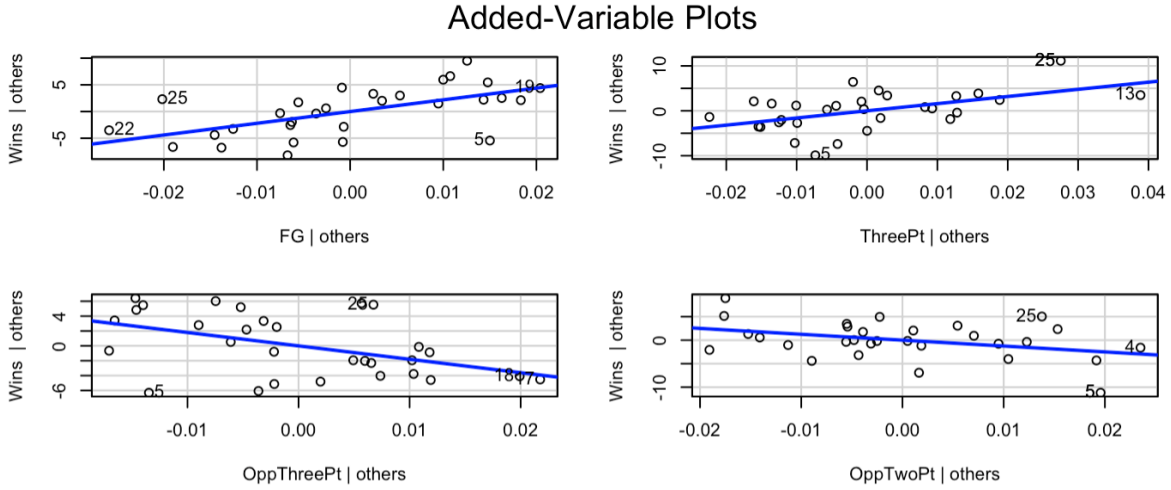
Appendix

Rk	Teams	Wins	FG	ThreePA	ThreePt	TwoPt	AST	TOV	OppThreePt	OppTwoPt	Playoffs
1	Atlanta Hawl	37	0.465	33.5	0.374	0.523	24.1	13.4	0.349	0.535	1
2	Boston Celtic	35	0.468	36.1	0.376	0.53	23.5	14.2	0.372	0.524	1
3	Brooklyn Net	43	0.492	36.5	0.389	0.564	26.6	13.7	0.367	0.525	1
4	Charlotte Ho	32	0.458	36.8	0.376	0.518	26.9	15	0.363	0.552	1
5	Chicago Bulls	26	0.477	33.6	0.364	0.546	26.8	15.4	0.355	0.546	0
6	Cleveland Ca	21	0.453	29.4	0.34	0.512	24.1	15.5	0.384	0.545	0
7	Dallas Maver	37	0.47	38.1	0.361	0.555	22.5	12.1	0.366	0.525	1
8	Denver Nugg	44	0.485	34	0.378	0.552	26.9	13.5	0.363	0.541	1
9	Detroit Pisto	19	0.45	33.1	0.353	0.511	24.3	14.8	0.366	0.535	0
10	Golden State	33	0.466	38.3	0.374	0.536	27.4	15.1	0.364	0.509	1
11	Houston Roc	16	0.441	40	0.334	0.53	23	14.8	0.381	0.537	0
12	Indiana Pace	30	0.471	34.1	0.363	0.536	26.9	13.7	0.368	0.522	0
13	Los Angeles (	44	0.482	34.5	0.417	0.526	24.4	13.1	0.359	0.528	1
14	Los Angeles I	37	0.473	31.4	0.353	0.543	24.6	15.4	0.351	0.522	1
15	Memphis Gri	33	0.466	31.1	0.359	0.521	26.9	13.2	0.368	0.528	0
16	Miami Heat	35	0.463	36.2	0.351	0.549	26.2	14.1	0.366	0.531	1
17	Milwaukee B	42	0.486	37.1	0.393	0.55	25.5	14	0.377	0.509	1
18	Minnesota T	20	0.446	37.2	0.352	0.511	25.5	14.4	0.393	0.537	0
19	New Orleans	30	0.48	30.5	0.35	0.549	26.2	14.7	0.382	0.534	0
20	New York Kn	37	0.457	29.9	0.391	0.492	21.3	13.2	0.337	0.516	1
21	Oklahoma Ci	21	0.441	35.6	0.343	0.508	22.2	16.3	0.36	0.546	0
22	Orlando Mag	21	0.43	31.8	0.346	0.477	21.9	12.8	0.373	0.54	0
23	Philadelphia	45	0.477	30	0.372	0.533	23.4	14.7	0.363	0.508	1
24	Phoenix Suns	47	0.489	34.7	0.373	0.564	26.9	12.7	0.356	0.531	1
25	Portland Trai	37	0.451	41.2	0.384	0.506	21.1	11.2	0.373	0.544	1
26	Sacramento	29	0.482	33	0.362	0.553	25.3	13.4	0.383	0.551	0
27	San Antonio	31	0.461	28.7	0.355	0.511	24.6	11.4	0.375	0.522	0
28	Toronto Rapt	27	0.45	39.6	0.371	0.513	24.4	13.3	0.378	0.522	0
29	Utah Jazz*	48	0.469	43.2	0.39	0.544	23.8	14.2	0.339	0.502	1
30	Washington	30	0.476	29	0.35	0.535	25.4	14.6	0.375	0.527	0

```
       Rk             Teams               Wins             FG            ThreePA
 Min.   : 1.00   Length:30          Min.   :16.0   Min.   :0.4300   Min.   :28.70
 1st Qu.: 8.25   Class :character   1st Qu.:27.5   1st Qu.:0.4540   1st Qu.:31.50
 Median :15.50   Mode  :character   Median :33.0   Median :0.4670   Median :34.30
 Mean   :15.50                      Mean   :32.9   Mean   :0.4658   Mean   :34.61
 3rd Qu.:22.75                      3rd Qu.:37.0   3rd Qu.:0.4770   3rd Qu.:37.02
 Max.   :30.00                      Max.   :48.0   Max.   :0.4920   Max.   :43.20
    ThreePt          TwoPt             AST             TOV          OppThreePt
 Min.   :0.3340   Min.   :0.4770   Min.   :21.10   Min.   :11.20   Min.   :0.3370
 1st Qu.:0.3523   1st Qu.:0.5122   1st Qu.:23.57   1st Qu.:13.22   1st Qu.:0.3608
 Median :0.3635   Median :0.5315   Median :24.60   Median :14.05   Median :0.3665
 Mean   :0.3665   Mean   :0.5299   Mean   :24.75   Mean   :13.93   Mean   :0.3669
 3rd Qu.:0.3760   3rd Qu.:0.5483   3rd Qu.:26.50   3rd Qu.:14.78   3rd Qu.:0.3750
 Max.   :0.4170   Max.   :0.5640   Max.   :27.40   Max.   :16.30   Max.   :0.3930
    OppTwoPt         Playoffs
 Min.   :0.5020   Min.   :0.0000
 1st Qu.:0.5220   1st Qu.:0.0000
 Median :0.5295   Median :1.0000
 Mean   :0.5298   Mean   :0.5333
 3rd Qu.:0.5393   3rd Qu.:1.0000
 Max.   :0.5520   Max.   :1.0000
```

What Makes Winning Basketball

Regressions 4214 Final Project — Marhawe Asmerom




```
Call:
glm(formula = Playoffs ~ ThreePt + FG + OppThreePt + OppTwoPt,
    family = binomial, data = NBA)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.64023  -0.16218   0.00209   0.23788   2.10936

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    73.22     114.57   0.639   0.5228
ThreePt       177.77      84.40   2.106   0.0352 *
FG            -22.86      86.31  -0.265   0.7911
OppThreePt   -222.80     167.05  -1.334   0.1823
OppTwoPt      -85.87      88.96  -0.965   0.3344
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 14.581  on 25  degrees of freedom
AIC: 24.581

Number of Fisher Scoring iterations: 8
```