Will Matteson
180.356
Final Project
NBA MVP
Table of Contents

R Code is separated into three parts
Web Scraper is located at https://github.com/WillMatteson/NBABioData

## An Attempt at Modeling the NBA MVP Race

### Introduction

The 21st century has seen a sudden and intense quantification of professional sports. Academics and media companies have applied computational techniques to athletic analysis, and the industry has begrudgingly followed suit. These new metrics and models are of great use to both the casual sports fan and decision-making bodies within sports organizations. As an avid sports fan and wanna-be sports statistician, the 2016-2017 NBA MVP race is the perfect target for an econometrics project. This paper will humbly try to accomplish a few things:

1. Can the MVP award be mathematically understood[1]
2. What are the most important components of MVP selection
3. Can MVP candidates be predicted

### Method

After experimenting with several web scrapers I realized that I needed to write my own. The script and sample data are uploaded to a public git repository[2]. In order to analyze the MVP award, I decided to scrape NBA data from the 1996-1997 season until present. Each of the 21 Seasons contains the regular season performance of every player, with per game averages as well as the rank of each performance category for that season. I used basketball reference to then append the amount of MVP votes a player received in each season, whether they won the award that season, and whether they had won the award before[3,4].

With all the years combined (except the most recent season, in which there have been no formal MVP votes awarded yet), my dataframe contained over 8,000 observations. The majority of players do not get any votes whatsoever, so I realized that my model needed to reflect the value of high performance at the margins of the categories. The difference between the 5th and 10th highest scorers is much more important than the difference between the 50th and 55th highest scorers in MVP determination.

Because of the importance of marginal production at first I tried to use rank data and nonlinear models to represent the MVP award. These models generally yielded poor results and were very expensive to compute. In particular, nonlinear model selection with a large amount degree of predictors is expensive to accomplish, so I had to explore other pathways.

I then realized that I could greatly reduce the total number of computations and solve my marginal production issues through restricting my sample. I chose a

---

[1] All terminology are explained in glossary

[2] https://github.com/WillMatteson/NBABioData

[3] http://www.basketball-reference.com/awards/awards_2016.html

[4] votes cut off for players who got <2% of tally

position agnostic measure (minutes per game > 32), which brought my total number of observations down to 1,500 without biasing my data. My reasoning was that MVP candidates are important players that have high usage rates, restricting my sample as such only removed players who were almost guaranteed not to get votes.

My initial guesses at relevant predictors were as such for my rank choices and raw choices.

Raw:

```
 [1] "W_PCT"      "FG_PCT"     "FG3_PCT"    "FT_PCT"     "REB"        "AST"        "TOV"
 [8] "STL"        "BLK"        "PTS"        "PLUS_MINUS" "DD2"        "TD3"
```

Rank:

```
 [1] "W_PCT_RANK"      "FG_PCT_RANK"     "FG3_PCT_RANK"    "FT_PCT_RANK"
 [6] "AST_RANK"        "TOV_RANK"        "STL_RANK"        "BLK_RANK"
[11] "PLUS_MINUS_RANK" "DD2_RANK"        "TD3_RANK"
```

My results were not ideal -- (See Graphs 1 and 2), so I decided to generate interaction terms and also conduct some additional tests to see what the most appropriate variables were to include.

From here I decided to carry out a Best Subset Regression in order to determine an optimal model. I decided to use the raw points data as the rank data behaved poorly in preliminary tests. A barrier in this part was my level of computing power. I also decided to interact my wins term with my other predictors in order to capture the win adjusted value of positive performance.

Graph 3 shows the results of my subset selection. Model performance improved while parameters amount increased. Upon examining the $C_p$ maximizing model coefficients I realized that they were ill suited for prediction.

```
  (Intercept)           FGM        FG_PCT       FG3_PCT        FT_PCT           AST
-1879.6723859   313.7121052  4443.9765092   181.1516176   248.8061060   -26.6591389
          TOV           DD2           TD3             W           Won    FGM:FG_PCT
   53.8634052    -3.3509811   -16.1921333    56.5161813   344.5429564  -753.2581793
 FG3M:FG3_PCT    FT_PCT:FTM       W_PCT:W         FGM:W      FG_PCT:W     FG3_PCT:W
 -134.4388723   -47.3903187     3.1172243    -9.3143549  -130.6316484    -5.9340712
     FT_PCT:W         FTM:W         AST:W         TOV:W         STL:W         BLK:W
  -14.0484886    -1.6727645     0.8418238    -1.6411597     0.4517989     0.3568325
        DD2:W         TD3:W  FGM:FG_PCT:W FG3M:FG3_PCT:W FT_PCT:FTM:W
    0.1044640     0.5213267    22.6415674     4.8766112     3.6749606
```

The coefficient on FG_PCT is far too high and other variables have curious directionality, so I then turned to the $R^2$ maximizing model.

```
  (Intercept)         W_PCT           FGM        FG_PCT       FG3_PCT        FT_PCT
-1.830858e+03 -6.927920e+01  3.036989e+02  4.432191e+03  1.759251e+02  2.332372e+02
          AST           TOV           BLK           DD2           TD3             W
-2.749362e+01  5.326302e+01 -1.801844e+01 -3.097110e+00 -1.481753e+01  5.553013e+01
          Won    FGM:FG_PCT  FG3M:FG3_PCT    FT_PCT:FTM       W_PCT:W         FGM:W
 3.432949e+02 -7.308692e+02 -1.282147e+02 -4.627180e+01  4.391634e+00 -9.092896e+00
     FG_PCT:W     FG3_PCT:W      FT_PCT:W         FTM:W         AST:W         TOV:W
-1.305547e+02 -5.867489e+00 -1.357275e+01 -1.659103e+00  8.654692e-01 -1.615810e+00
        STL:W         BLK:W         DD2:W         TD3:W  FGM:FG_PCT:W FG3M:FG3_PCT:W
 4.514243e-01  7.913106e-01  9.735695e-02  4.895907e-01  2.214879e+01  4.714602e+00
 FT_PCT:FTM:W
 3.630983e+00
```

This model was preferable, but I realized that I still had some issues. There were too many complicating variables, causing possible overfitting. Also, it was possible for my model to produce negative fitted values. For that reason I then decided to experiment with a Poisson regressions, which seemed especially appropriate as votes are counting variables.

With the Poisson model all of my predictors had strong significance but my model suffered from high Residual Deviance[5]. The errors were also skewed right. I suspect that there is some over fitting, but given my computational resources and timeline I was unable to conduct a programmatic model selection process to filter predictors out.

I decided that this was the most realistic model, as the model chosen through the best subsets selection had unrealistic predictor coefficients didn't wasn't very practical. After running the just elapsed regular season (2016-2017) data through the Poisson model the projected winner was Russell Westbrook. Although vote data has not been formally announced, he is the borderline census MVP. The fitted value for the data is too high, but their votes amount relative to each other is very accurate. As I am just trying to predict the winners, the magnitude of predicted votes are less important than their order. The high residual deviance seems to be more a reflection of the vote inflation, rather than a sign that the model is not informative.

| | Most Votes | Least Votes (Cut off at 32 mpg) |
|---|---|---|
| 1. | Russell Westbrook | Justice Winslow |
| 2. | James Harden | Courtney Lee |
| 3. | Kevin Durant | Ricky Rubio |
| 4. | LeBron James | Evan Fournier |
| 5. | Stephen Curry | Kentavious Caldwell-Pope |

## Conclusion

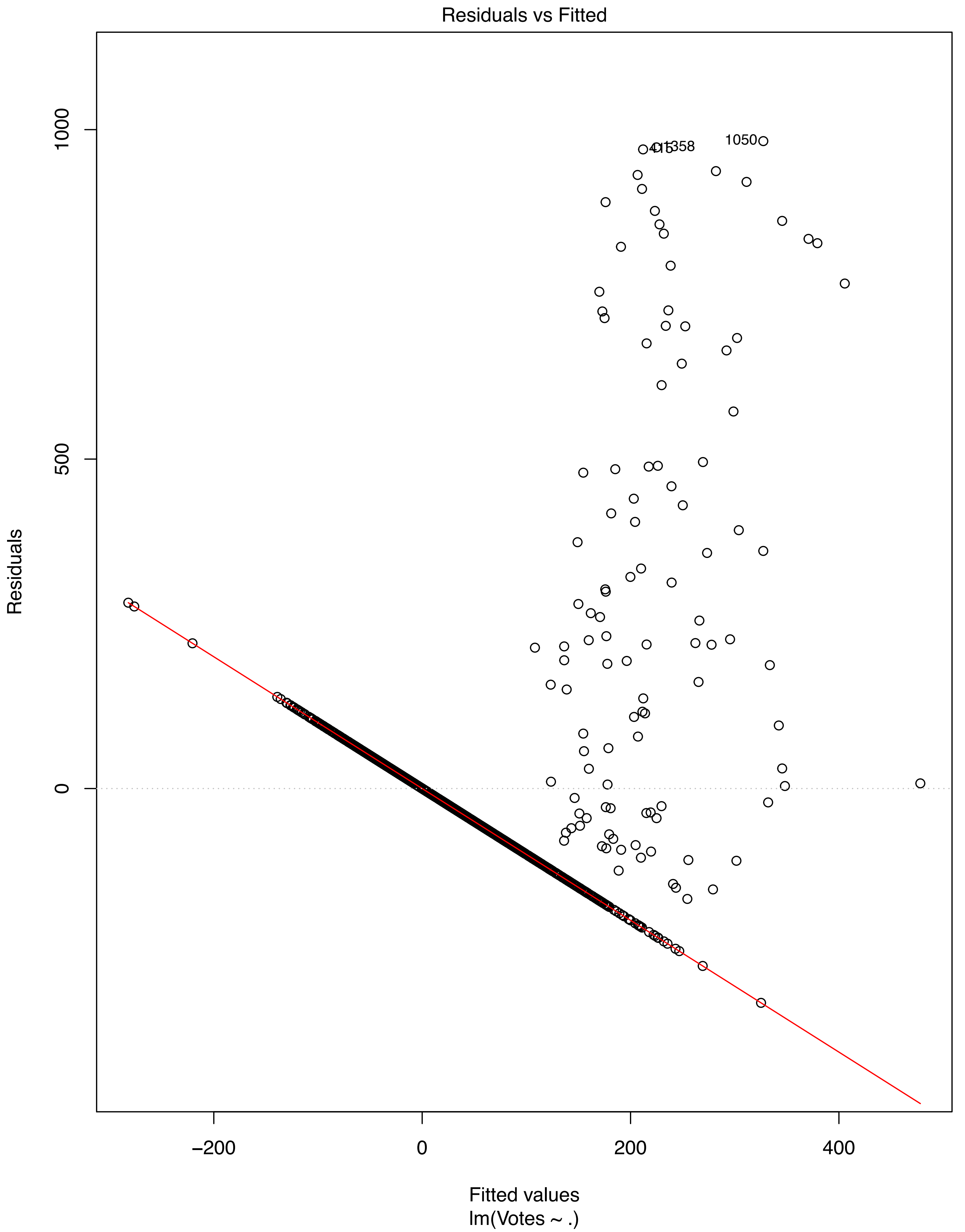
This project was much more difficult than I anticipated.  There are inherent issues with using votes as a response variable, when they aren't necessarily independent of each other (voters are notoriously collaborative and often vote in blocs). I think there is a lot of further analysis to be done about the main influences in MVP determination that could be accomplished with more rigorous model selection.

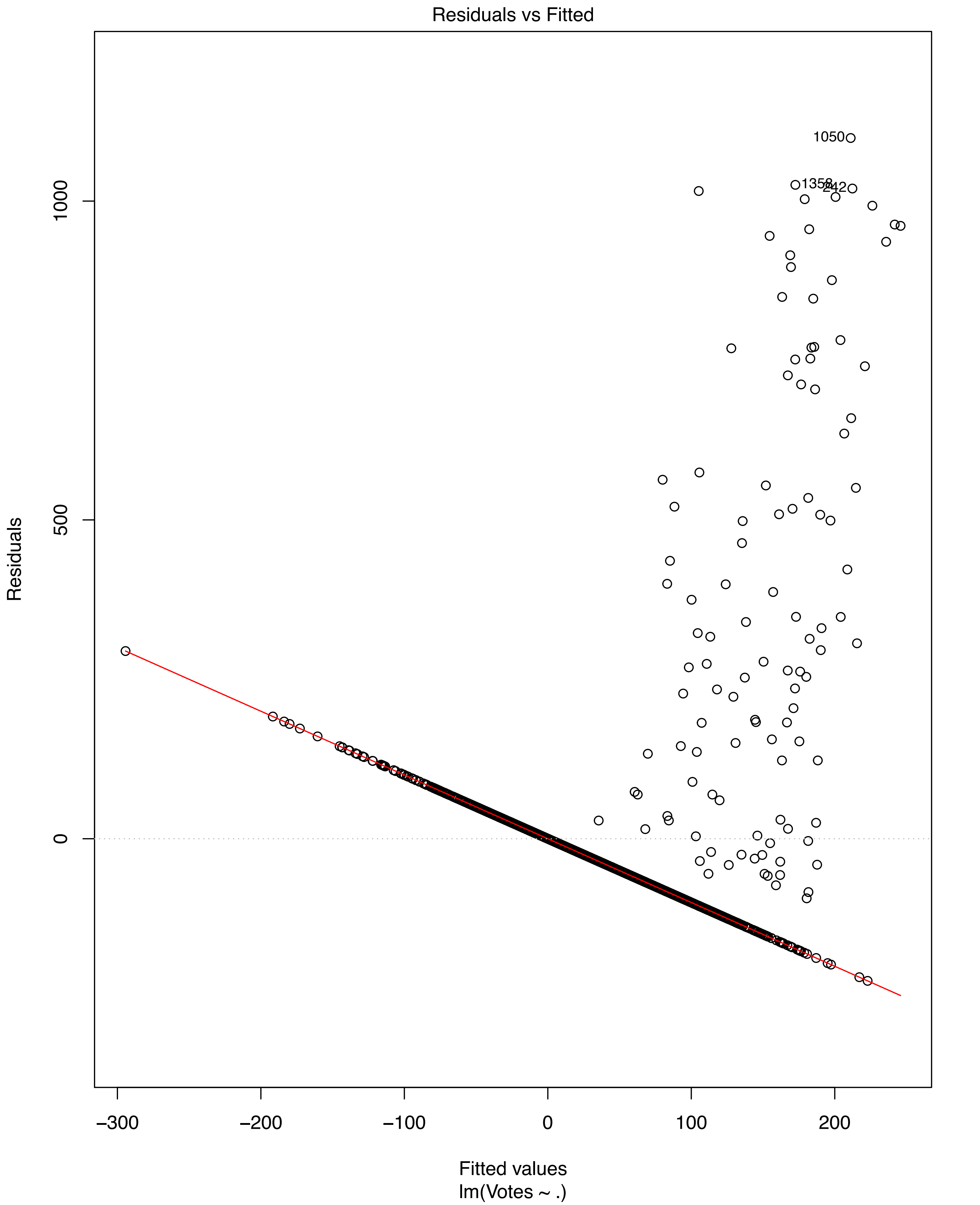
Despite the difficulties of this project I do feel that there are some worthwhile findings. I was surprised by the importance of including win interaction terms. Including the terms was much more explanatory than I had anticipated. Defensive statistics like steals and blocks were much less important than I had expected. I had thought that winning the award in previous years diminishes  (or at least sets the bar higher for) future votes. I suspect that I would need to find a more sophisticated way to interact winning a previous MVP with future performance in

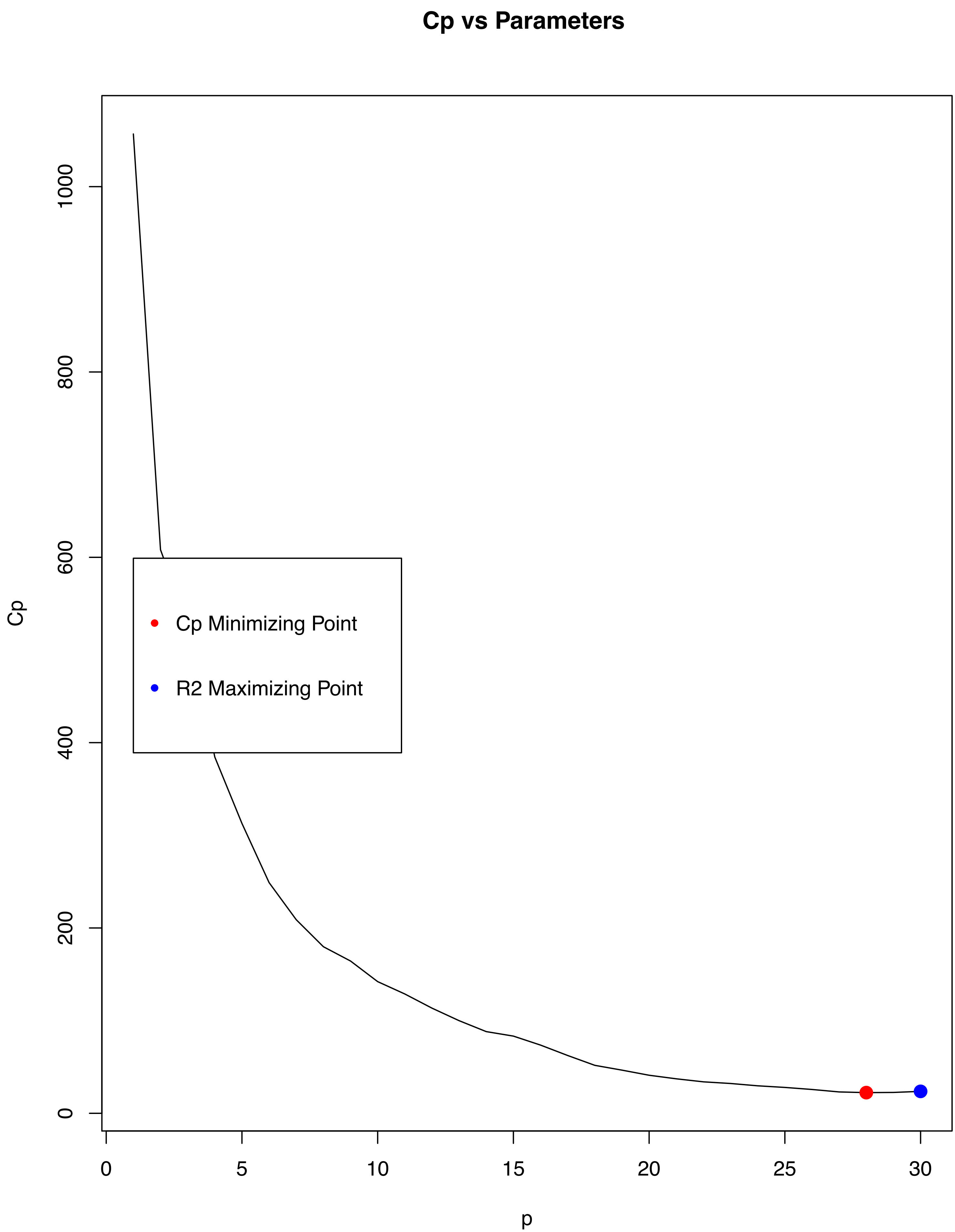
[5] see poissonOutput.txt

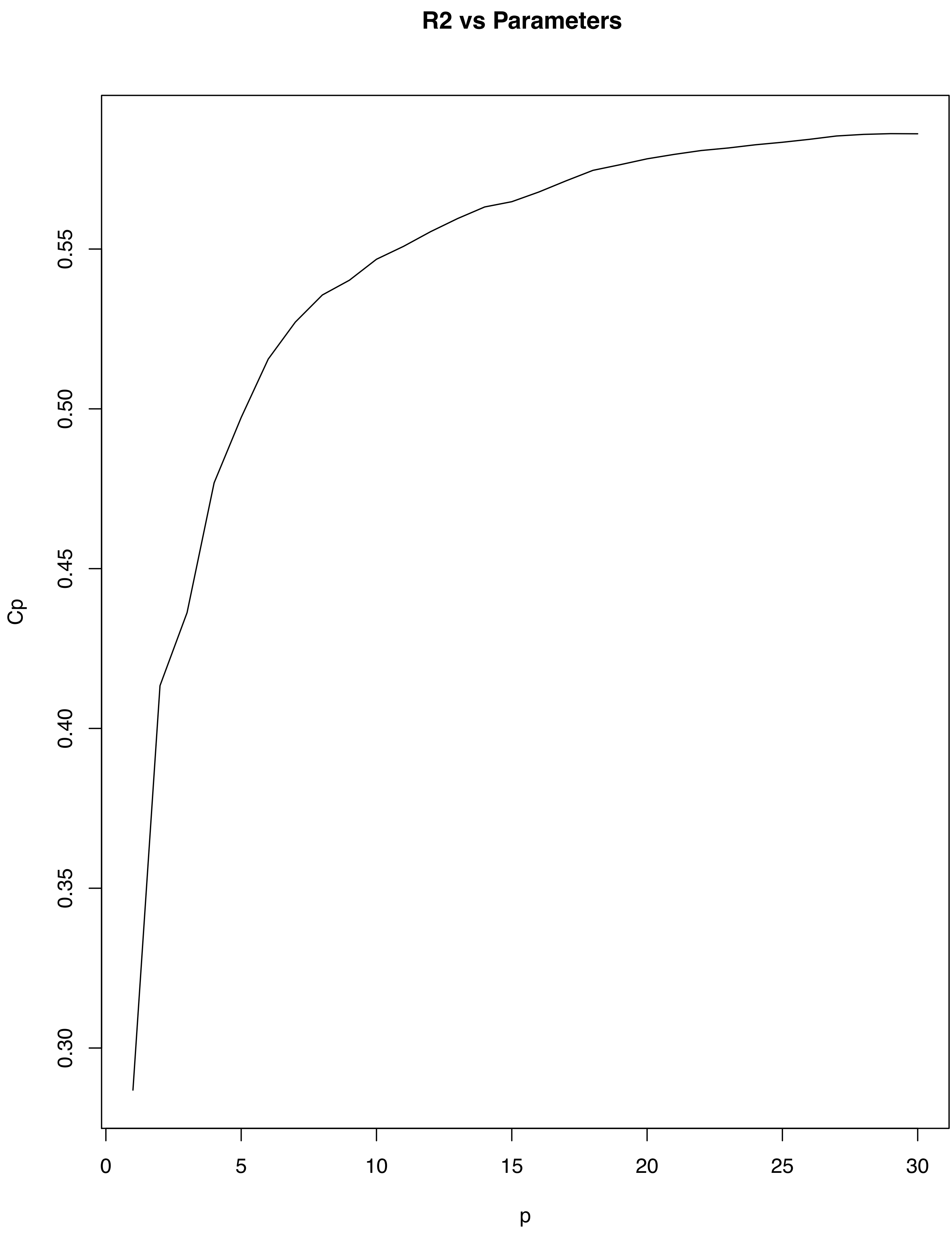
order to display such behavior. I think it would also be interesting to look at how playoff performance impacts voting. Although the MVP award is a regular season award, it does tend to retrospectively favor high performers in the previous year's playoffs. Lastly, I was surprised that my model chose Westbrook at the 2016-2017 MVP, I personally expected Harden to be chosen.

| Term | Definition |
|---|---|
| **MVP** | Most Valuable Player. Awarded to the “Best Player” in the league every year. |
| **Votes** | A panel of about 125 sportswriters every year is awarded ballots by the league. For each ballot the first vote is worth 10 points, second place is worth 7, third is worth 5, fourth is worth 3, and fifth is worth 1. “Votes” as used in this paper refers to total vote points a player has for a given season. |
| **FGM** | Field Goals Made |
| **FG_PCT** | Field Goal Percentage |
| **FT** | Free Throws |
| **FT_PCT** | Free Throw Percentage |
| **W_PCT** | Win Percentage for that player’s team |
| **FG3** | 3-PT Field Goals Made |
| **FG3_PCT** | 3-PT Percentage |
| **PTS** | Points |
| **BLKS** | Blocks |
| **REB** | Total Rebounds |
| **AST** | Assists |
| **STL** | Steals |
| **TOV** | Turnovers |
| **Plus Minus** | The point differential for a given player’s time on court. |
| **Double Double (DD2)** | The amount of games in which a player records double digit totals in two of five categories (Assists, Blocks, Points, Rebounds, Steals) |
| **Triple Double (TD2)** | The amount of games in which a player records double digit totals in three of five categories. Considerably rarer and more prestigious than a Double Double. |

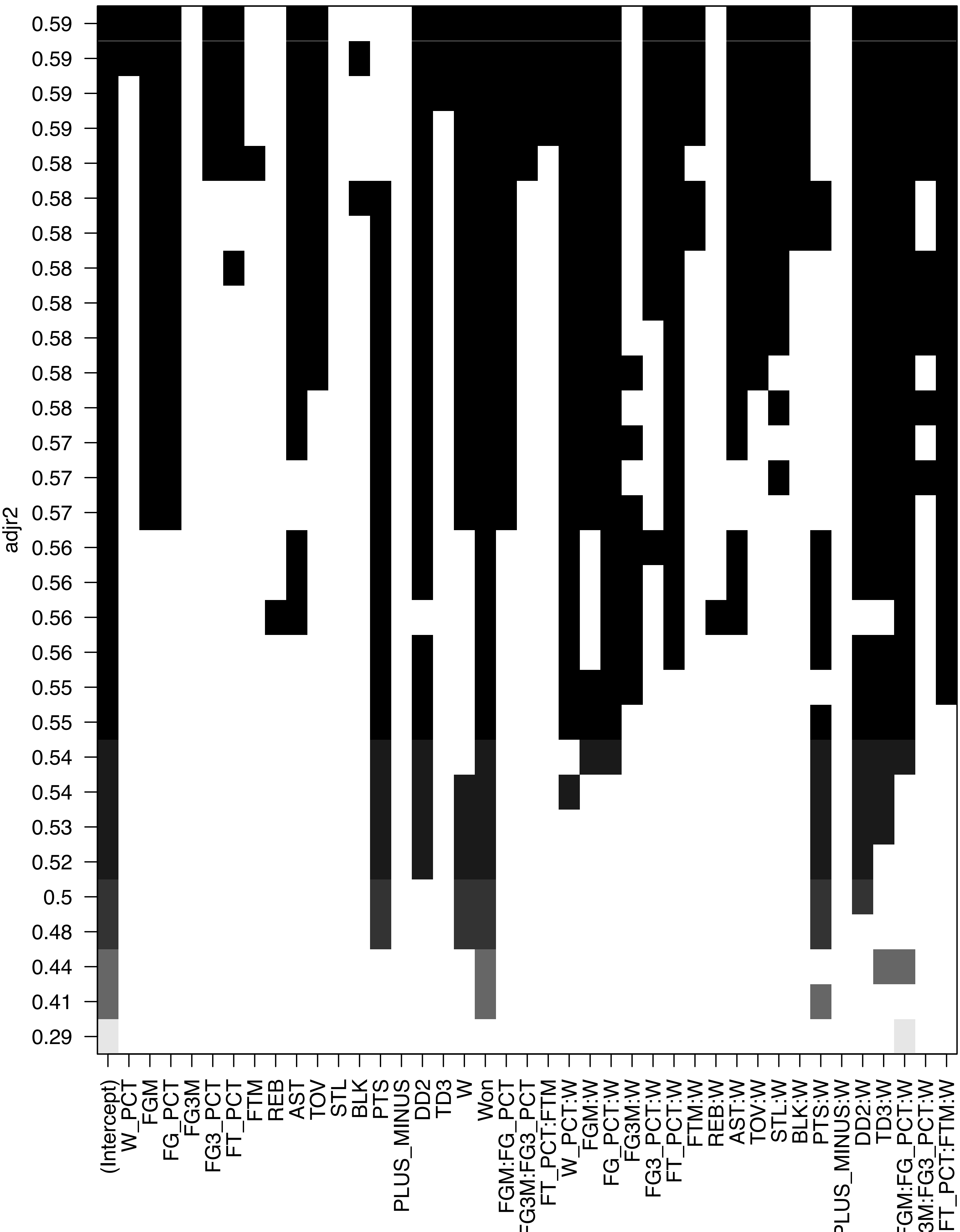

Residuals vs Fitted
1050
1358
415
1000
500
0
Residuals
-200
0
200
400
Fitted values
lm(Votes ~ .)


Residuals vs Fitted
Residuals
Fitted values
lm(Votes ~ .)
1050
1358
242
1000
500
0
−300
−200
−100
0
100
200


Cp vs Parameters
Cp
p
0
200
400
600
800
1000
0
5
10
15
20
25
30
Cp Minimizing Point
R2 Maximizing Point



R2 vs Parameters
Cp
p
0.30
0.35
0.40
0.45
0.50
0.55
0
5
10
15
20
25
30



adjr2
0.59
0.59
0.59
0.59
0.58
0.58
0.58
0.58
0.58
0.58
0.58
0.58
0.57
0.57
0.57
0.56
0.56
0.56
0.56
0.55
0.55
0.54
0.54
0.53
0.52
0.5
0.48
0.44
0.41
0.29
(Intercept)
W_PCT
FGM
FG_PCT
FG3M
FG3_PCT
FT_PCT
FTM
REB
AST
TOV
STL
BLK
PTS
PLUS_MINUS
DD2
TD3
W
Won
FGM:FG_PCT
G3M:FG3_PCT
FT_PCT:FTM
W_PCT:W
FGM:W
FG_PCT:W
FG3M:W
FG3_PCT:W
FT_PCT:W
FTM:W
REB:W
AST:W
TOV:W
STL:W
BLK:W
PTS:W
PLUS_MINUS:W
DD2:W
TD3:W
FGM:FG_PCT:W
3M:FG3_PCT:W
FT_PCT:FTM:W