Titolo del progetto

Domenico Plantamura, Eduardo David Lotto, Manuel D'Alterio Grazioli, Gabriele Fugagnoli

Contents

Introduction

Objective of the project

Our goal is to investigate whether the salaries earned by the NBA players during the 2023-2024 season are fair in proportion to their performance during the current year's Regular season. To analyse performance, we selected several statistics: from the most common such as points, rebounds, assists to advanced metrics like Usage, Player Impact Estimated and Winning Shares. The idea is to deep dive into the relationship between salaries and performance through different models in order to understand what kind of relationship there is and which model best fits the data. Finally, we will compare actual salaries with those predicted by our models to find out which players (according to the models) are the most overpaid or underpaid.

Steps followed

To perform our analysis we followed these steps:

1. Data collection;
2. Data exploration;
3. Analysis;
4. Interpretation.

We now explain in depth each step.

Data collection

We performed a web scraping operation from the Official NBA Stats website, from which we collected most of the stats. Additionally, we downloaded data about the salaries from Hoopshype and other stats of interest from Basketball reference. All data concerns the 2023-2024 NBA Regular Season.

Why consider only Regular Season data?

Considering only data about Regular Season without considering players performance during playoffs limits a bit the potential of our analysis. On one hand, it's reasonable to infer that player performance during playoffs should have an important weight in determining his salary. On the other hand, considering playoffs in the analysis carries different issues.

There are teams (and consequently players) that go further than others: 14 out of 30 teams can't qualify for the playoffs. For the teams which qualify, playoff stats are calculated on a number of games that could differ greatly between different teams (e.g. if a team loses in the first round, it plays from 4 to 7 games. If a team reaches the finals, it plays from 16 to 28 games). During Regular Season every team plays a fixed number of games, 82.

Additionally, coaches usually rotate players at their disposal in a different way during playoffs: for instance, during regular season approximately 10-12 players for each team take part in the game; during playoffs it is not uncommon to observe only 7-8 players that come into play for each team. Furthermore, usually in a playoff game the best players are more involved compared to Regular season games. It means that, first of all, they play several more minutes. Moreover, they have the ball in their hands for a lot of time and consequently their stats grow a lot; hence, it could happen that few players record a large part of the entire team's statistics. Considering this, including playoffs data in the analysis could lead to an overestimation of performance of 2-3 players and to an underestimation of the performance of the rest of the team.

All in all, it is undeniable that playoffs are a fundamental part of the season. It is also obvious that if a player has more responsibilities in that phase he probably deserves a higher salary. But we think that for the purposes of our analysis, the addition of statistics collected on a small sample of matches, different for practically every team, with highly polarized data between the various players may lead to biases if not handled properly.

We think that considering only the regular season, although leading to a limited analysis, may be sufficient to grasp the main relationships between salaries and performance.

Glossary

- **PLAYER NAME**: name of a player;
- **SALARY**: salary earned by a player for 2023-2024 season (collected from Hoopshype);
- **AGE**: age of a player;
- **POS**: "Position", the playing position of a player.

Traditional stats (collected from the NBA website)

- **GP**: "Games played", the number of games played by a player during the 2023-2024 regular season;
- **FG_PCT**: "Field Goal Percentage", the percentage of field goal attempts that a player makes. Formula: (FGM)/(FGA);
- **FG3_PCT**: "3 Points "Field Goal Percentage", the percentage of 3pt field goal attempts that a player makes;
- **FT_PCT**: "Free throws Percentage", the percentage of free throws attempts that a player makes;
- **OREB**: "Offensive Rebounds", the number of rebounds a player or team has collected while they were on offense;

- **DREB**: “Defensive Rebounds”, the number of rebounds a player or team has collected while they were on defense;
- **REB**: “Rebounds”, a rebound occurs when a player recovers the ball after a missed shot. This statistic is the number of total rebounds a player has collected on either offense or defense;
- **AST**: “Assists”, the number of assists (passes that lead directly to a made basket) by a player;
- **TOV**: “Turnovers”, a turnover occurs when a player on offense loses the ball to the defense;
- **STL**: “Steals”, number of times a defensive player takes the ball from a player on offense, causing a turnover;
- **BLK**: “Blocks”, a block occurs when an offensive player attempts a shot, and the defense player tips the ball, blocking their chance to score;
- **BLKA**: “Blocks Against”, The number of shots attempted by a player or team that are blocked by a defender
- **PF**: “Personal fouls”, the number of personal fouls a player or team committed;
- **PFD**: “Personal fouls drawn”, the number of personal fouls that are drawn by a player or team;
- **PTS**: “Points”, the number of points scored by a player;
- **MIN**: “Minutes played”, number of minutes played by a player during the 2023-2024 Regular season;
- **MIN_G**: “Minutes played per game”.

Advanced stats (collected from the NBA website)

- **OFF_RATING**: “Offensive Rating”, measures a team’s points points scored per 100 possessions while a player is on the court. Formula: 100*((Points)/(POSS);
- **DEF_RATING**: “Defensive Rating”, the number of points per 100 possessions that the team allows while a player is on the court. Formula: 100*((Opp Points)/(Opp POSS));
- **NET_RATING**: “Net Rating”, Measures a team’s point differential per 100 possessions while a player is on the court. Formula: OFFRTG - DEFRTG;
- **AST_TO**: “Assist to Turnover Ratio”, the number of assists for a player compared to the number of turnovers committed;
- **TS_PCT**: “True Shooting Percentage”, a shooting percentage that factors in the value of three-point field goals and free throws in addition to conventional two-point field goals. Formula: Points/ [2*(Field Goals Attempted+0.44*Free Throws Attempted)];
- **USG_PCT**: “Usage Percentage”, the percentage of team plays used by a player when they are on the floor. Formula: (FGA + Possession Ending FTA + TO) / POSS;
- **PIE**: “Player Impact Estimate”, measures a player’s overall statistical contribution against the total statistics in games they play in. PIE yields results which are comparable to other advanced statistics (e.g. PER) using a simple formula. Formula: (PTS + FGM + FTM - FGA - FTA + DREB + (.5 * OREB) + AST + STL + (.5 * BLK) - PF - TO) / (GmPTS + GmFGM + GmFTM - GmFGA - GmFTA + GmDREB + (.5 * GmOREB) + GmAST + GmSTL + (.5 * GmBLK) - GmPF - GmTO).

The stats below are collected from Basketball Reference:

- **WS**: “Win Shares”, attempts to divvy up credit for team success to the individuals on the team. It is calculated using player, team and league-wide statistics and the sum of player win shares on a given team will be roughly equal to that team’s win total for the season (more details on the Basketball Reference page);
- **BPM**: “Box Plus/Minus”, a box score estimate of the points per 100 possessions that a player contributed above a league-average player, translated to an average team;
- **VORP**: “Value Over Replacement Player”, a box score estimate of the points per 100 TEAM possessions that a player contributed above a replacement-level (-2.0) player, translated to an average team and prorated to an 82-game season. Multiply by 2.70 to convert to wins over replacement.

BPM and VORP are calculated per 100 possessions; MIN and WS are calculated over the whole regular season, MIN_G is calculated per game. The other stats are considered per 48 minutes.

Why statistics per 48 minutes?

Considering most statistics projected over 48 minutes avoids overestimating performance for players who play, on average, more minutes in a game. In this way we think that the contribution of each player is fairly evaluated and not distorted by the minutes played.

Data integration and cleaning

Once we had obtained the tables of interest, we selected from each table the statistics useful for analysis (those given in the glossary) and then merged the slices of the various datasets, removing all the players who played less than 480 minutes during the entire regular season.

```r
data_trad_tot <- data_traditional_tot[data_traditional_tot$MIN > 480, ]

data_st <- merge(data_salary, data_traditional_per48, by = "PLAYER_NAME", all = TRUE)
data_ast <- merge(data_st, data_advanced, by = "PLAYER_NAME", all = TRUE)
data_mast <- merge(data_ast, data_miscellaneous, by = "PLAYER_NAME", all = TRUE)
data_mastt <- merge(data_mast, data_trad_tot, by = "PLAYER_NAME", all = TRUE)
final_dataset <- merge(data_mastt, data_vorp, by = "PLAYER_NAME", all = TRUE)
```

The reason why we selected players with at least 480 minutes played is that we wanted to avoid considering stats taken on a too small amount of minutes. After these operation, the final dataset consists of 360 rows and 31 columns.

At this stage, we cleaned the data following these other steps:

- NA removal;
- Matching players' names;
- Transforming the Salary column into a numeric one;
- Putting the players' name as row names for the dataset and thus removing the PLAYER_NAME column.

Data exploration

Before studying the data with formal models, we got an overview through an exploratory data analysis. For the first part of our analysis we used only numeric variables, so the categorical parameter 'Pos', which you can see on the table below, was removed from the dataset at this stage.

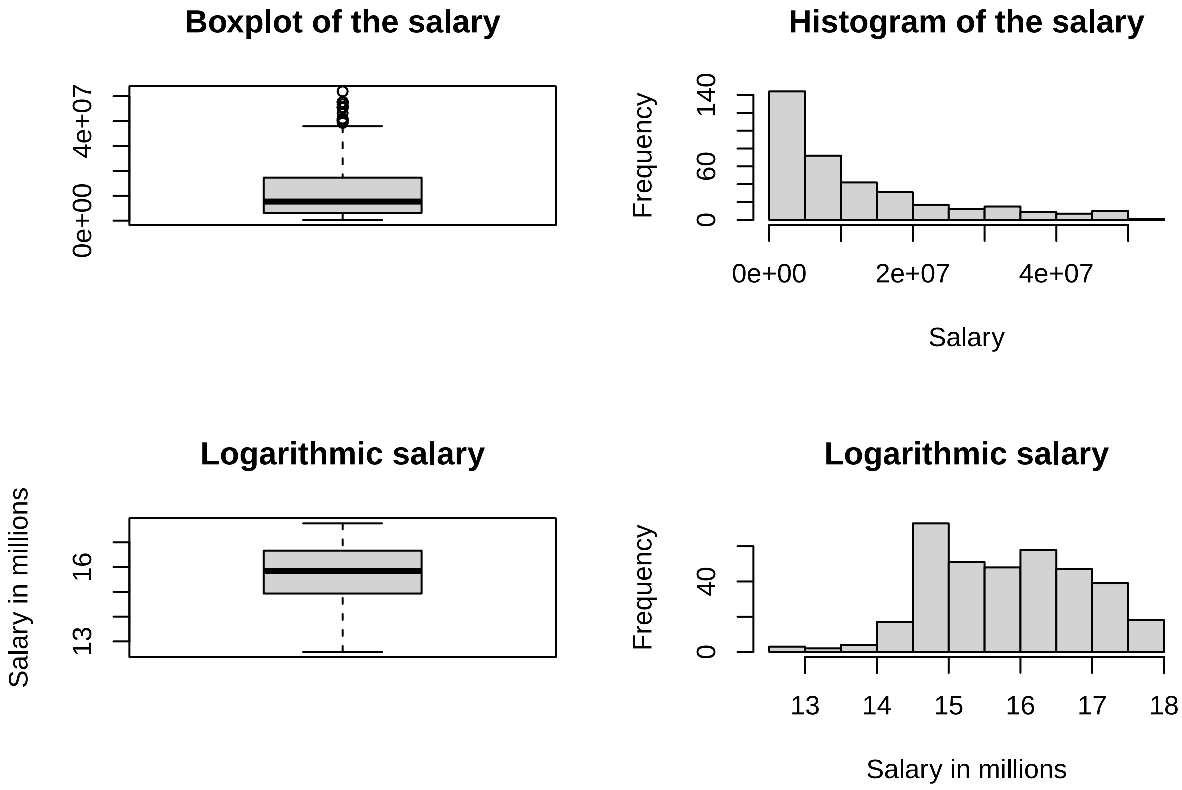
	Salary	AGE	GP	FG_PCT	FG3_PCT	FT_PCT	OREB	DREB	REB	AST
Aaron Gordon	22266182	28	73	0.556	0.290	0.658	3.6	6.2	9.8	5.4
Aaron Holiday	2346614	27	78	0.446	0.387	0.921	0.9	3.8	4.7	5.3
Aaron Nesmith	5634257	24	72	0.496	0.419	0.781	1.5	5.1	6.6	2.6
Aaron Wiggins	1836096	25	78	0.562	0.492	0.789	2.3	4.9	7.3	3.4
Al Horford	10000000	37	65	0.511	0.419	0.867	2.3	9.1	11.4	4.6

	TOV	STL	BLK	BLKA	PF	PTS	OFF_RATING	DEF_RATING	NET_RATING	AST_
Aaron Gordon	2.2	1.2	0.9	1.2	3.0	21.2	119.8	111.1	8.7	2.4
Aaron Holiday	2.0	1.6	0.2	0.8	4.7	19.4	110.5	107.6	2.9	2.6
Aaron Nesmith	1.5	1.6	1.2	1.2	5.8	21.1	119.3	115.0	4.3	1.6

	TOV	STL	BLK	BLKA	PF	PTS	OFF_RATING	DEF_RATING	NET_RATING	AST_
Aaron Wiggins	2.2	2.2	0.7	1.3	3.6	21.2	115.6	110.0	5.7	1.5
Al Horford	1.3	1.0	1.7	0.3	2.6	15.5	120.9	109.5	11.4	3.5

	TS_PCT	USG_PCT	PIE	PFD	MIN	MIN_G	Pos	WS	BPM	VORP
Aaron Gordon	0.607	0.174	0.103	4.7	2296.810	31.46315	PF	7.1	1.3	1.9
Aaron Holiday	0.578	0.158	0.078	2.5	1269.297	16.27303	PG	2.5	-1.5	0.2
Aaron Nesmith	0.631	0.158	0.071	3.5	1994.655	27.70354	SF	4.1	-0.5	0.8
Aaron Wiggins	0.664	0.163	0.096	2.3	1227.938	15.74280	SG	3.7	0.7	0.8
Al Horford	0.650	0.119	0.105	0.8	1739.797	26.76610	C	6.2	3.6	2.5

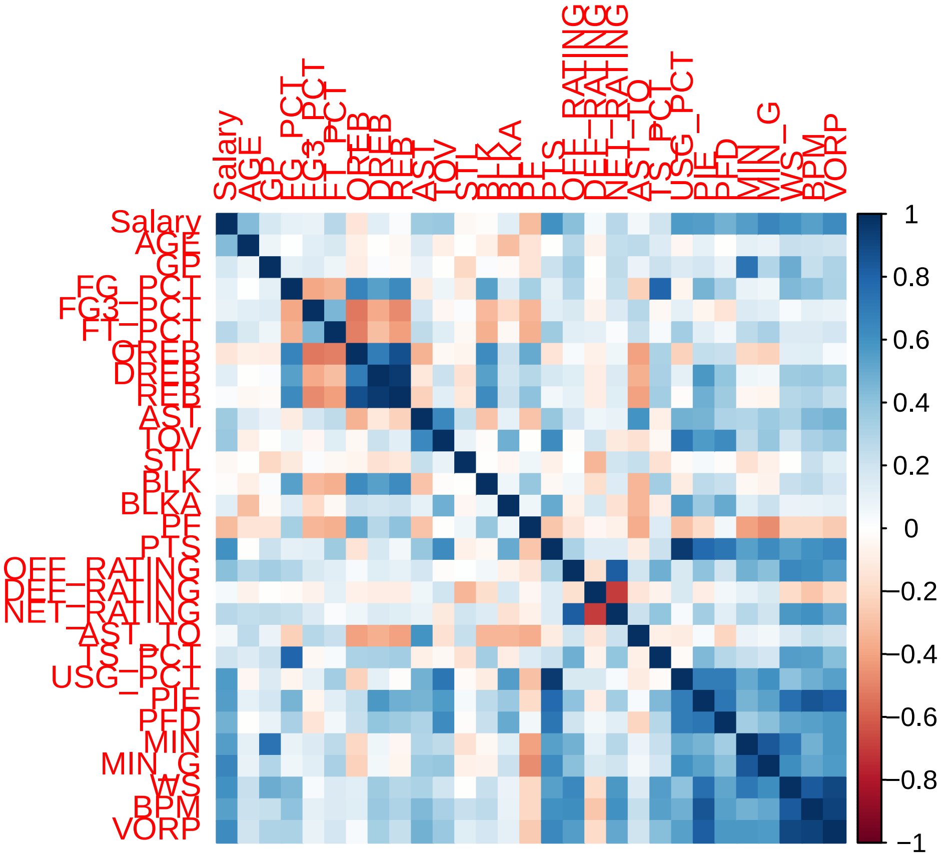
Firstly, an analysis of the variable Salary that will be the dependent variable in the models.



The boxplot shows that the salary distribution is right skewed, with some outliers in the right side. We expected this kind of distribution, the outliers are the players earning the highest salaries. The histogram also highlights the right skewed distribution. It can be seen that Salary's log transformation reduces the skewness and makes the distribution of the variable closer to normal.

In order to study correlations between the predictors of the model, we used the corrplot function.

```r
library(corrplot)
corrplot(cor(fd_numeric), method = 'color')
```



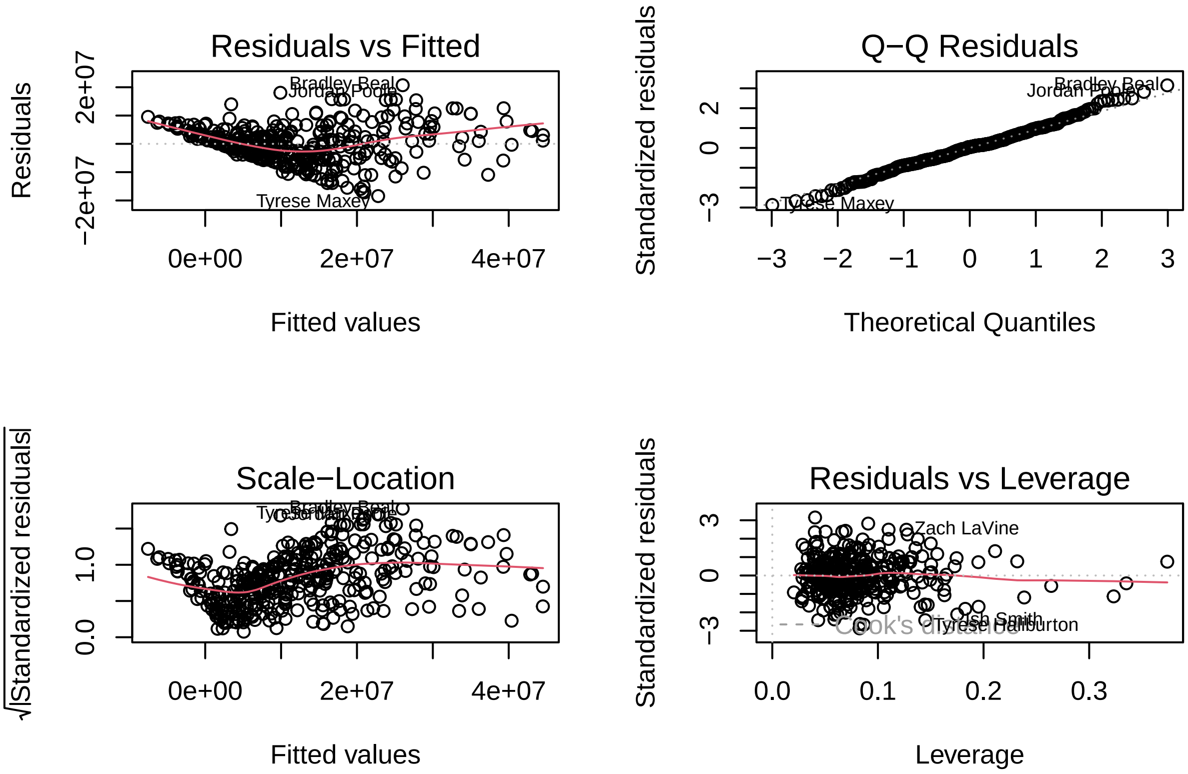
Different correlations between the variables emerge from the corrplot. With regard to the variable Salary, it is interesting to notice that Salary is positively correlated with PTS and advanced stats like USG_PCT, BPM and VORP: all of these variables are related to players' shots and point contribution. For what concerns the other variables, there are some obvious correlations: for instance, between variables MIN (total minutes played during the regular season) and MIN_G (minutes played per game) and between variables REB, OREB and DREB (all related to rebounds, with the relation REB = OREB + DREB). Additionally, we expected the positive correlation between BPM and VORP because are both related to players point estimation. A strong positive correlation emerges between PTS and USG_PCT. The usage percentage is "The percentage of team plays used by a player when they are on the floor. Formula: (FGA + Possession Ending FTA + TO) / POSS". Thus, players with a high USG_PCT often make the last play in an offensive possession (a shot, a free throw or a turnover): it is straightforward that if a player often ends the offensive possession of his team, he has more opportunities to score points. For what concerns the negative correlations, the most interesting are the ones between rebounds variables (OREB, DREB, REB), FT_PCT and FG3_PCT. Players that grab a lot of rebounds are usually the tallest ones and these players are not great free throws shooters or 3 point shooters (on average).

Models

We started creating a linear regression model in order to predict salaries.

```
##
## Call:
## lm(formula = Salary ~ +., data = fd_numeric)
##
## Residuals:
```

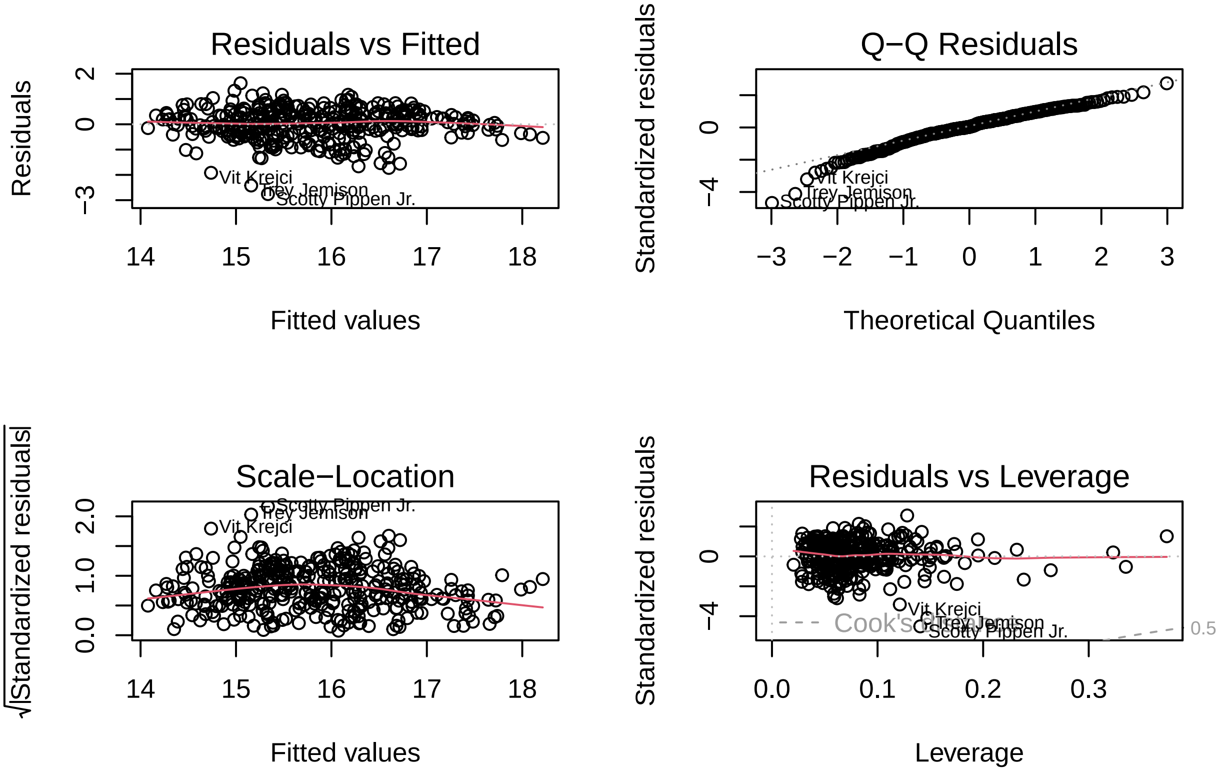
```
##       Min        1Q    Median        3Q       Max
## -18450260  -4028989    276645   4003025  20712902
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -12550803   28966256  -0.433   0.6651
## AGE            1057586      93293  11.336   <2e-16 ***
## GP              -21024     118116  -0.178   0.8588
## FG_PCT        35836089   19086013   1.878   0.0613 .
## FG3_PCT         -56045    4195524  -0.013   0.9893
## FT_PCT          975535    6135667   0.159   0.8738
## OREB           4054377    6865112   0.591   0.5552
## DREB           4473997    6846450   0.653   0.5139
## REB           -4315225    6838752  -0.631   0.5285
## AST             -98527     667680  -0.148   0.8828
## TOV            2003183    1516145   1.321   0.1873
## STL             -69046     985541  -0.070   0.9442
## BLK             601287     664109   0.905   0.3659
## BLKA          -2253383    1230646  -1.831   0.0680 .
## PF             -616626     640191  -0.963   0.3362
## PTS            1117890     623630   1.793   0.0740 .
## OFF_RATING    16646653    6955245   2.393   0.0172 *
## DEF_RATING   -16681974    6953008  -2.399   0.0170 *
## NET_RATING   -16610236    6957987  -2.387   0.0175 *
## AST_TO         -252115     978605  -0.258   0.7969
## TS_PCT       -63710004   30332753  -2.100   0.0365 *
## USG_PCT      -40539821   73391644  -0.552   0.5811
## PIE         -131534170  115933751  -1.135   0.2574
## PFD             101829     397847   0.256   0.7981
## MIN              -4774       4689  -1.018   0.3094
## MIN_G           696311     299872   2.322   0.0208 *
## WS             1845668     740418   2.493   0.0132 *
## BPM            -391702     764769  -0.512   0.6089
## VORP            663105    1436430   0.462   0.6446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6712000 on 331 degrees of freedom
## Multiple R-squared:  0.7081, Adjusted R-squared:  0.6834
## F-statistic: 28.67 on 28 and 331 DF,  p-value: < 2.2e-16
```



```
## [1] 4.142347e+13
```

The complete model has a good R-squared of 0.68 and a MSE of 4.14e+13. It emerges that many variables are not significative in determining the response. Through the residual analysis it is noticeable that the relationship between fitted values and residuals is not exactly linear (1st graph). Additionally, in the third graph the points are not are included in a band of constant amplitude parallel to the x-axis, hence the omoschedasticity assumption can be doubted.

```
##
## Call:
## lm(formula = log(Salary) ~ +., data = fd_numeric)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75872 -0.32207  0.02064  0.42623  1.62364
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.258e+01  2.744e+00   4.583 6.49e-06 ***
## AGE          9.669e-02  8.838e-03  10.940  < 2e-16 ***
## GP          -5.087e-04  1.119e-02  -0.045   0.9638
## FG_PCT       3.669e+00  1.808e+00   2.029   0.0432 *
## FG3_PCT     -7.625e-02  3.975e-01  -0.192   0.8480
## FT_PCT      -1.574e-01  5.813e-01  -0.271   0.7867
## OREB         3.180e-01  6.504e-01   0.489   0.6252
## DREB         4.105e-01  6.486e-01   0.633   0.5273
## REB         -3.435e-01  6.479e-01  -0.530   0.5963
## AST          3.954e-02  6.325e-02   0.625   0.5323
## TOV          1.536e-03  1.436e-01   0.011   0.9915
## STL          5.077e-02  9.337e-02   0.544   0.5870
## BLK          7.242e-02  6.291e-02   1.151   0.2505
## BLKA        -1.802e-01  1.166e-01  -1.545   0.1232
## PF          -5.294e-02  6.065e-02  -0.873   0.3834
## PTS          9.972e-02  5.908e-02   1.688   0.0924 .
## OFF_RATING   1.508e+00  6.589e-01   2.289   0.0227 *
## DEF_RATING  -1.499e+00  6.587e-01  -2.276   0.0235 *
## NET_RATING  -1.504e+00  6.592e-01  -2.282   0.0231 *
## AST_TO      -4.552e-02  9.271e-02  -0.491   0.6237
## TS_PCT      -6.498e+00  2.874e+00  -2.261   0.0244 *
## USG_PCT     -2.881e+00  6.953e+00  -0.414   0.6789
## PIE         -1.715e+01  1.098e+01  -1.562   0.1193
## PFD          3.123e-02  3.769e-02   0.829   0.4079
## MIN         -5.447e-05  4.442e-04  -0.123   0.9025
## MIN_G        5.720e-02  2.841e-02   2.014   0.0449 *
## WS           1.246e-01  7.014e-02   1.777   0.0765 .
## BPM          5.637e-02  7.245e-02   0.778   0.4371
## VORP        -1.639e-01  1.361e-01  -1.204   0.2293
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6359 on 331 degrees of freedom
## Multiple R-squared:  0.6516, Adjusted R-squared:  0.6222
## F-statistic: 22.11 on 28 and 331 DF,  p-value: < 2.2e-16
```



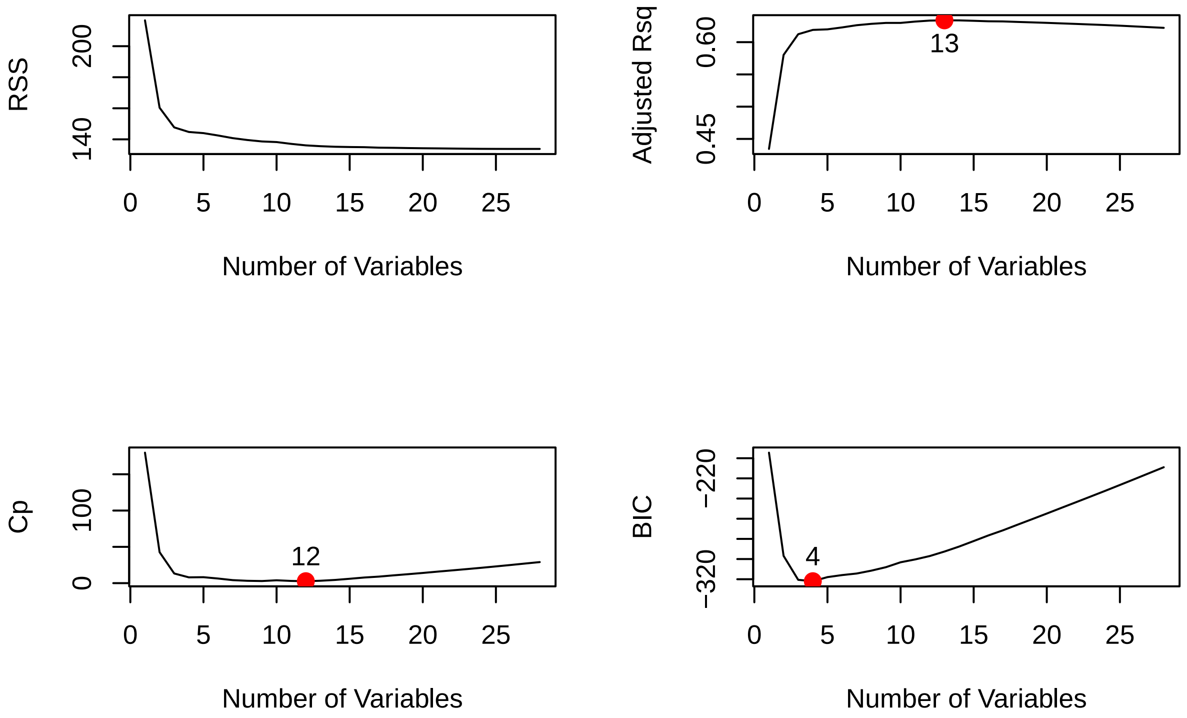
```
## [1] 4.703165e+13
```

With a logarithmic transformation of the dependent variable, the model shows a slightly lower R-squared (0.62 vs the previous 0.68) and a slightly higher MSE (4.70e+13 vs the previous 4.14e+13). Applying a logarithmic transformation to the dependent variable, the first graph shows a more linear relationship and the third graph allows to infer a more constant variance in the error terms. In both models many variables are not significative in determining the response: for this reason, to avoid a model that is unnecessary complex, we performed a variable selection. A logarithmic transformation of the dependent variable Salary will be applied because, although it slightly worsens the performance of the model, it makes the salaries distribution closer to normal, it improves the linearity of the model and it reduces residuals eteroschedasticity.

Variable selection

We selected a subset of relevant features starting from the predictors used in the complete model in order to have a simpler model that is easier to interpret, without redundant variables and less prone to overfitting. To do so, we used The regsubsets function which performs best subset selection by identifying the best model that contains a given number of predictors, where best is quantified using RSS. We set the function to return results up to the best 28-variables model.

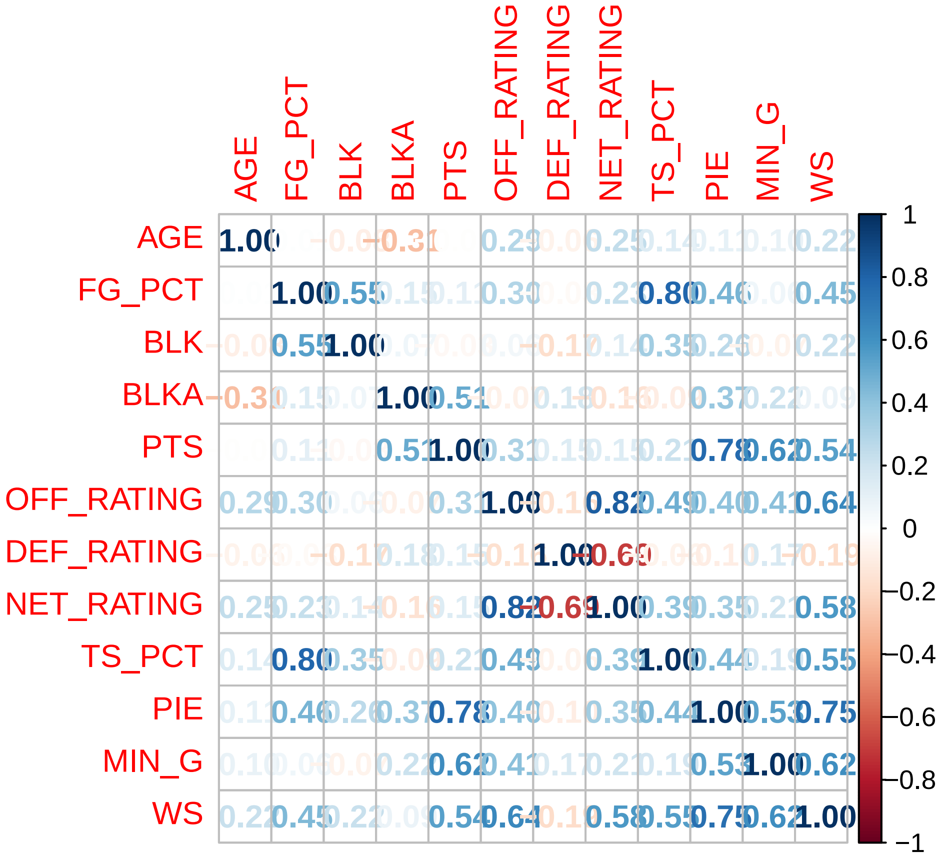
To find the best balance between model simplicity and precision, we evaluated the number of parameters to be included in the model through Mallow's Cp, BIC and Adjusted R-squared.



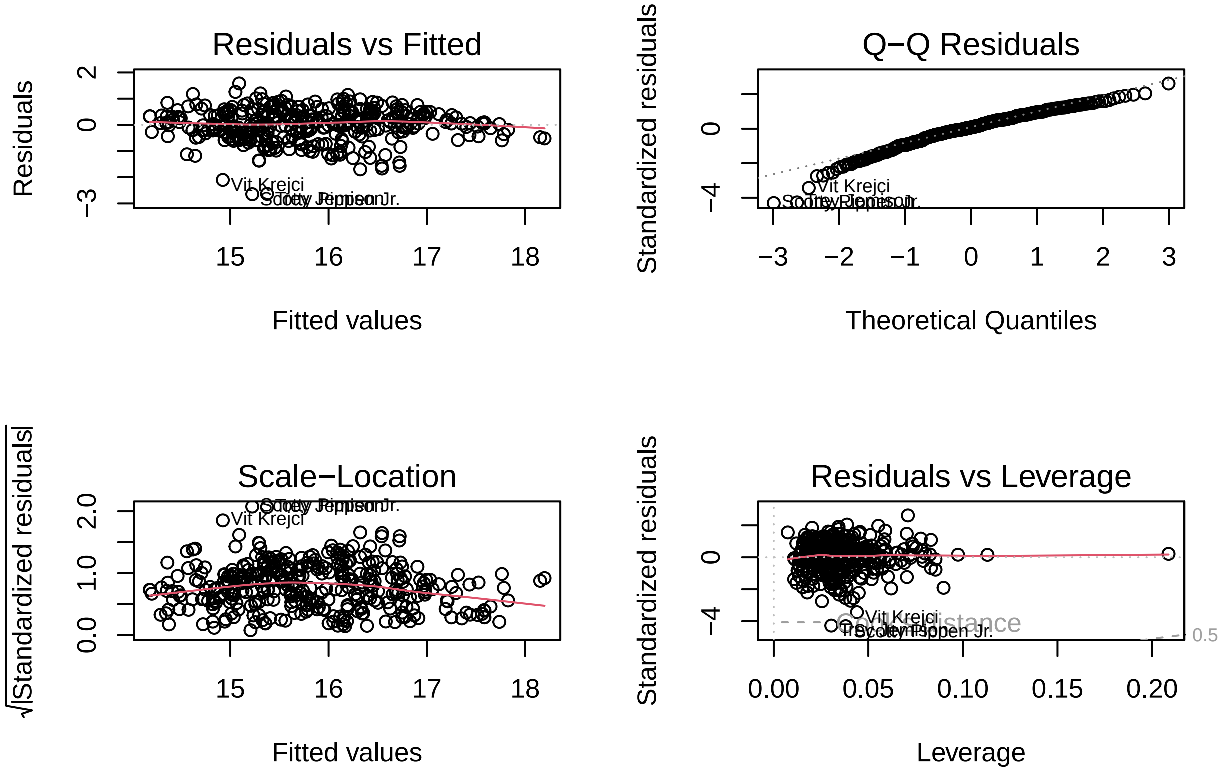
Considering Mallow's Cp, the best number of parameters for our model is 12. We obtained the list of parameters from the regsubset function to get the best model with 12 parameters.

```
##
## Call:
## lm(formula = selected.formula, data = fd_numeric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6476 -0.3193  0.0482  0.4315  1.5814
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.852025   1.626347   6.673 9.95e-11 ***
## AGE          0.095227   0.008503  11.199  < 2e-16 ***
## FG_PCT       3.445692   1.155704   2.981  0.00307 **
## BLK          0.083159   0.049934   1.665  0.09674 .
## BLKA        -0.167470   0.106759  -1.569  0.11764
## PTS          0.058816   0.011460   5.132 4.78e-07 ***
## OFF_RATING   1.531106   0.629095   2.434  0.01544 *
## DEF_RATING  -1.513509   0.629188  -2.405  0.01667 *
## NET_RATING  -1.520988   0.629121  -2.418  0.01614 *
## TS_PCT      -5.839664   1.438599  -4.059 6.09e-05 ***
## PIE         -6.258145   2.750650  -2.275  0.02351 *
## MIN_G        0.059152   0.006890   8.585 3.10e-16 ***
## WS           0.053790   0.026074   2.063  0.03986 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6263 on 347 degrees of freedom
## Multiple R-squared:  0.6457, Adjusted R-squared:  0.6334
## F-statistic:  52.7 on 12 and 347 DF,  p-value: < 2.2e-16
```

INTERPRETARE MODELLO



INTERPRETARE E USARE DOPO PER SPIEGARE PERCHE' RIDGE E LASSO



INTERPRETARE

```
## [1] 4.775116e+13
```

INTERPRETARE