Relation between NBA performance and Salary

22/12/2020

Abstract

Three models are created to predict the value of an NBA player's salary. The models are based on a per game statistics, more advanced statistics like Win Shares, and Box Plus/Minus and a model using both of those stats. The model using residuals compares how players are getting paid and what is predicted. This helps determine how much a player should be getting paid, and concludes if a player is being under or over paid based on their performance. Further it can help predict what the player should then be paid the next contracting year.

The code and dataset for this analysis can be found here: https://github.com/drTricked/NBA-Analysis

Keywords

Keywords: Observational Study, Multiple linear Regression, NBA, Basketball, Salary

Introduction

In the NBA, a player's performance is often associated with the amount of money they are being paid. Player's with bigger and more impactive roles are often the recipients of these large paychecks. These players can have large stat totals which introduces the question, what values correlate with a players salary. Are players really overpaid or underpaid relative to their salaries?

What will show an impactive role? Potentially some variables that could be most important are the offensive statistics that aid in a team's win. This includes points and assists (a pass to a teammate that leads to a score) to name a few. Salary is important value to investigate, because a team's decision is often influenced by the salary of their current player and the players that they aim to sign.

Often when looking at relation between variables a type of regression analysis is used. This statistical method can be used to help determine how strong a connection between one response variables and other explanatory variables. In this report, a multilinear regression analysis will be used to investigate the relationship between different variables mainly from a set of per game statistics and more advanced basketball statistics. These include points per game or win shares (an estimate of how much a player contributes to a win) for example.

We will be combining datasets to create one single dataset holding all the information about the players. In the Methodology section (Section 2), details about the dataset and the model that we create will be shown. Then following, the Results section (Section 3) will show the results from the model. Then our analysis will end with a discussion and conclusion found in the Conclusion section (Section 4).

Methodology

Data:

Three datasets will be used for the analysis. All the data comes from the website Basketball Reference, where they were converted into csv files. The first dataset is for the salary amount of the contract of all the players in the 2019-2020 season. The second dataset holds the per game statistics of each player, and the third dataset holds some more advanced statistics of the player.

For the first dataset, it can be seen that there are two columns one for the player name, and one for the respective salary. One issue with the dataset is that some players will show up more than once, this is due to other contract issues that often arises from trades or waivers that allow a player to be paid from a previous team then their current team. An example of this is like the player Alfonzo McKinnie.

Table 1: Salary Issue

	Player	Salary
436	Alfonzo McKinnie	1361046
476	Alfonzo McKinnie	708871
528	Alfonzo McKinnie	183114

He had received a contract from the Cleveland Cavaliers, only to be waived. McKinnie was removed from the roster, he later was signed with the same team to a 10-day contract and then resigned with another 10-day contract. For these observations we combine them into one data value.

The dataset also contains the values of the salary originally as strings, so the data is cleaned to produce salary as numeric values.

Looking at the data, salary appears similar to a log normal distribution, then a logarithmic transformation is suitable.

Table 2: Salary Dataset Head

Player	Salary
A.J. Hammons	350087
Aaron Gordon	19863636
Aaron Holiday	2239200
Abdel Nader	1618520
Admiral Schofield	898310

The second dataset holds the values of a player's per game statistic. One problem with this data is the existence of NA values. This occurs due to one of the values being a percentage, the ratio between shots made and shots attempted. Clearly the percentage cannot exist if the player has 0 attempted shots. So those observations are removed from the dataset. The players with these NA values fortunately are players that often have very small amounts of games played. Thus they should not be representative of the population that is the players of the 2019-2020 season.

Another issue is that for a player that existed on more than one team, they will have multiple observations, including a TOT (two or more teams) which combines the statistics for each of the observations. In cleaning this dataset, only the variables with TOT will remain over the other individual observations for a player.

Table 3: TOT example

	Player	Pos	Age	Tm
19	Trevor Ariza	SF	34	TOT
20	Trevor Ariza	SF	34	SAC
21	Trevor Ariza	SF	34	POR

Table 4: Per Game Dataset Head

	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG.
2	Bam Adebayo	PF	22	MIA	72	72	33.6	6.1	11.0	0.557
3	LaMarcus Aldridge	C	34	SAS	53	53	33.1	7.4	15.0	0.493
4	Nickeil Alexander-Walker	SG	21	NOP	47	1	12.6	2.1	5.7	0.368

	Player	X3P	X3PA	X3P.	X2P	X2PA	X2P.	eFG.	FT	FTA	FT.
2	Bam Adebayo	0.0	0.2	0.143	6.1	10.8	0.564	0.558	3.7	5.3	0.691
3	LaMarcus Aldridge	1.2	3.0	0.389	6.2	12.0	0.519	0.532	3.0	3.6	0.827
4	Nickeil Alexander-Walker	1.0	2.8	0.346	1.1	2.8	0.391	0.455	0.5	0.8	0.676

	Player	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
2	Bam Adebayo	2.4	7.8	10.2	5.1	1.1	1.3	2.8	2.5	15.9
3	LaMarcus Aldridge	1.9	5.5	7.4	2.4	0.7	1.6	1.4	2.4	18.9
4	Nickeil Alexander-Walker	0.2	1.6	1.8	1.9	0.4	0.2	1.1	1.2	5.7

Table 5: Per Game Statitics Meaning

Variables	Meanings
Player	Player Name
Pos	Position
Age	Age
Tm	Team Initials
G	Games Played
GS	Games Started
MP	Minutes Per Game Played
FG	Field Goals Per Game
FGA	Field Goals Attempted Per Game
FG.	Field Goals Percentage
X3P	3 Point Field Goals Per Game
X3PA	3 Point Field Goals Attempted Per Game
X3P.	3 Point Field Goal Percentage
X2P	2 Point Field Goals Per Game
X2PA	2 Point Field Goals Attempted Per Game
X2P.	2 Point Field Goal Percentage
eFG.	Effective Field Goal Percentage
FT	Free Throws Per Game
FTA	Free Throws Attempted Per Game
FT.	Free Throw Percentage
ORB	Offensive Rebounds Per Game
DRB	Defensive Rebounds Per Game
TRB	Total Rebounds Per Game
AST	Assists Per Game
STL	Steals Per Game
BLK	Blocks Per Game
TOV	Turnovers Per Game
PF	Personal Fouls

Variables	Meanings
PTS	Points Per Game

Some further explanation to the variables: A field goal is a basket scored that does not come from a free throw. Free throws are shots taken behind the free throw line that are awarded from a foul. A rebound is when a player retrieves the ball after a missed field goal or free throw. An assist is given for the player who's pass to their teammate lead to a score by field goal from that teammate. Blocks occur when a defensive player legally deflects an opposing players shot. A steal is when a defensive player leggally causes a turnover. Turnovers are when a player loses possession of the ball to the opposing team by the players actions.

The third dataset holds more advanced statistics and follows similar issues as the previous dataset. Thus similar data cleaning is used.

Table 6: Per Game Dataset Head

	Player	Pos	Age	Tm	G	MP	PER	TS.	X3PAr	FTr
2	Bam Adebayo	PF	22	MIA	72	2417	20.3	0.598	0.018	0.484
3	LaMarcus Aldridge	C	34	SAS	53	1754	19.7	0.571	0.198	0.241
4	Kyle Alexander	C	23	MIA	2	13	4.7	0.500	0.000	0.000

	Player	ORB.	DRB.	TRB.	AST.	STL.	BLK.	TOV.	USG.	OWS	DWS
2	Bam Adebayo	8.5	24.9	17.0	24.2	1.7	3.8	17.6	21.2	4.6	3.9
3	LaMarcus Aldridge	6.3	17.8	12.0	11.4	1.0	4.4	7.8	23.4	3.0	1.4
4	Kyle Alexander	17.9	8.3	12.9	0.0	0.0	0.0	33.3	10.2	0.0	0.0

	Player	WS	WS.48	OBPM	DBPM	BPM	VORP
2	Bam Adebayo	8.5	0.168	1.4	2.0	3.4	3.3
3	LaMarcus Aldridge	4.5	0.122	1.8	-0.5	1.4	1.5
4	Kyle Alexander	0.0	-0.003	-6.1	-3.5	-9.6	0.0

Table 7: Advanced Statitics Meaning

Variables	Meanings
Player	Player Name
Pos	Position
Age	Age
Tm	Team Initials
G	Games Played
MP	Minutes Played
PER	Player Efficiency Rating
TS.	True Shooting Percentage
X3PAr	3 Point Attempt Rate
FTr	Free Throw Attempt Rate
ORB.	Offensive Rebound Percentage
DRB.	Defensive Rebound Percentage
TRB.	Total Rebound Percentage
AST.	Assist Percentage

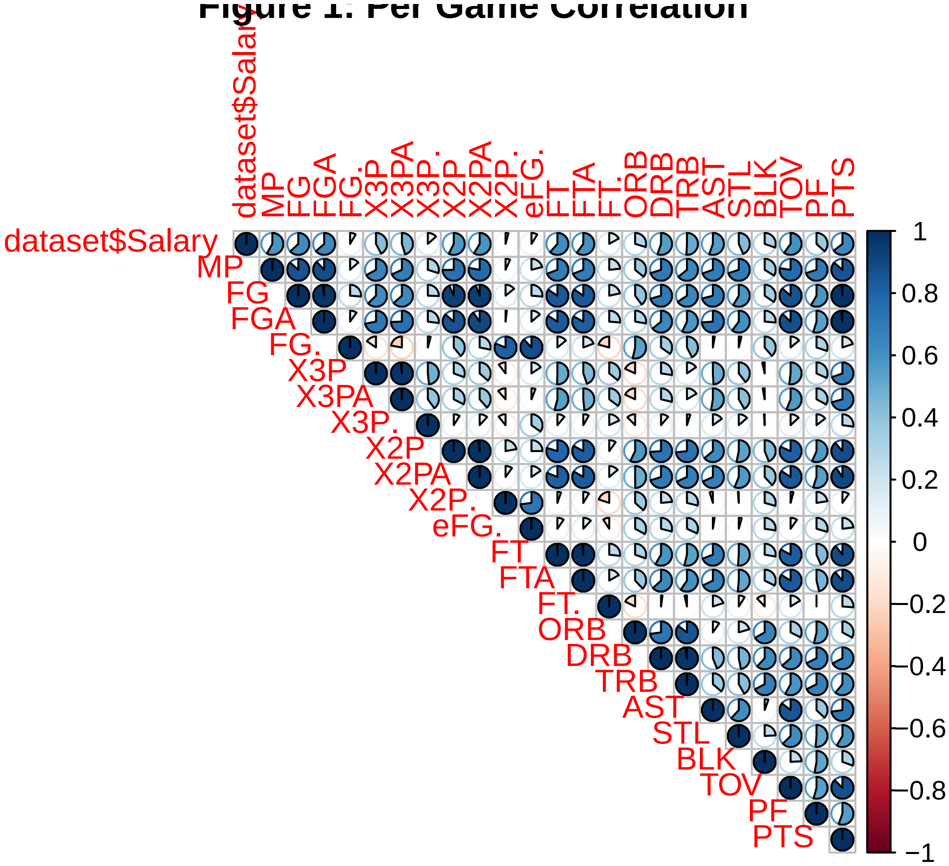
Variables	Meanings
STL.	Steal Percentage
BLK.	Block Percentage
TOV.	Turnover Percentage
USG.	Usage Percentage
OWS	Offensive Win Shares
DWS	Defensive Win Shares
WS	Win Shares
WS.48	Win Shares / 48 Minutes
OBPM	Offensive Box Plus/Minus
DBPM	Defensive Box Plus/Minus
BPM	Box Plus/Minus
VORP	Value over Replacement Player

Some further explanation to the variables: Many of the percentages are calculated for when the player is on the court. Win shares is an estimate of the number of wins a player contributes to. Box Plus/Minus is an estimate of the points per 100 possessions a player contributes relative to the league average player.

Each of these datasets come from observational data. It is collected as public information from the National Basketball Association (NBA). The population would be all the players in the NBA over its 74 years, while this frame is based on the players of the 2019-2020 season. Both of the two datasets, for statistics of per game and for more advanced statistics they include statistics like the Win Shares or the Box Plus/Minus.

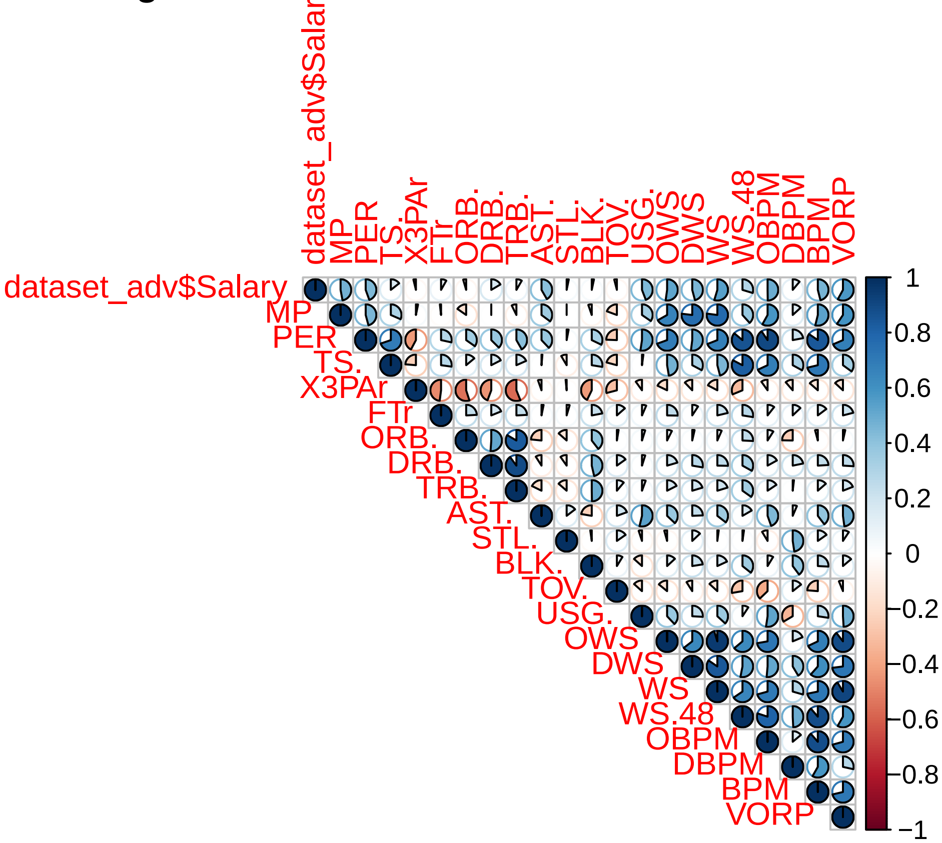
For the first model it is only based on the combination of salary and the points per game statistics, the second model builds on the combination of salary and the advanced statistics, while the third model will be based on both points per game statistics and the advanced statistics. When looking at the values for the first model, there are some strongly correlated values. This comes from the fact that the values used to calculate the percentage statistics are included, thus only the percentage variables are kept. Further we remove the predictor of Turnovers per game since it has high collinearity with many other variables like assists or points. Since the variables are quite similar its fine to keep only the percentage variables instead.

Figure 1: Per Game Correlation



The second model looks at more advanced statistics, like Box Plus/Minus, and Win Share. Looking at the correlation coefficients there are overlapping variables so we remove them. Similarly with this reasoning, the variables that offer both offensive and defensive specifics are dropped for their normal total counterparts. For instance the variables of offensive win shares and defensive win shares are dropped where win shares is kept.

Figure 2: Advanced Statistics Correlation



The third model uses all three of the datasets, and similarly the same reasonings are applied on what variables are dropped. Looking at the correlation coefficients, although the dataset has both a per game statistic and a similar percentage statistic, only if there is high correlation will the variable be dropped. For instance Assist percentage is dropped.

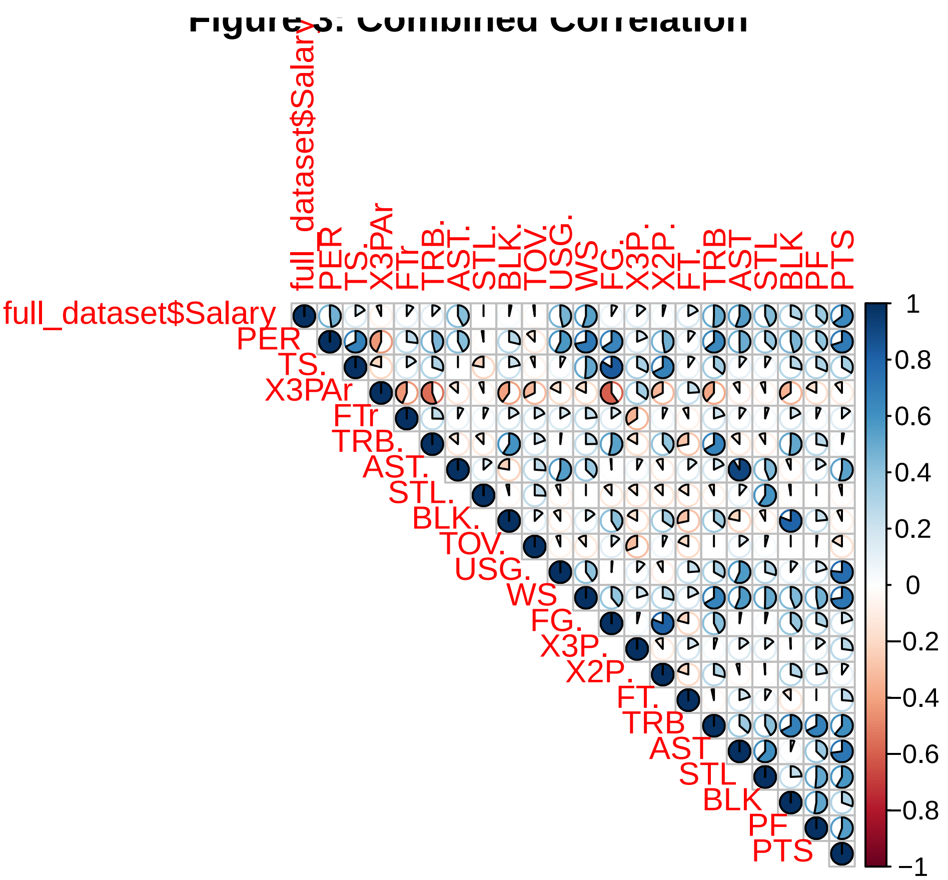


Figure 3: Combined Correlation

Model:

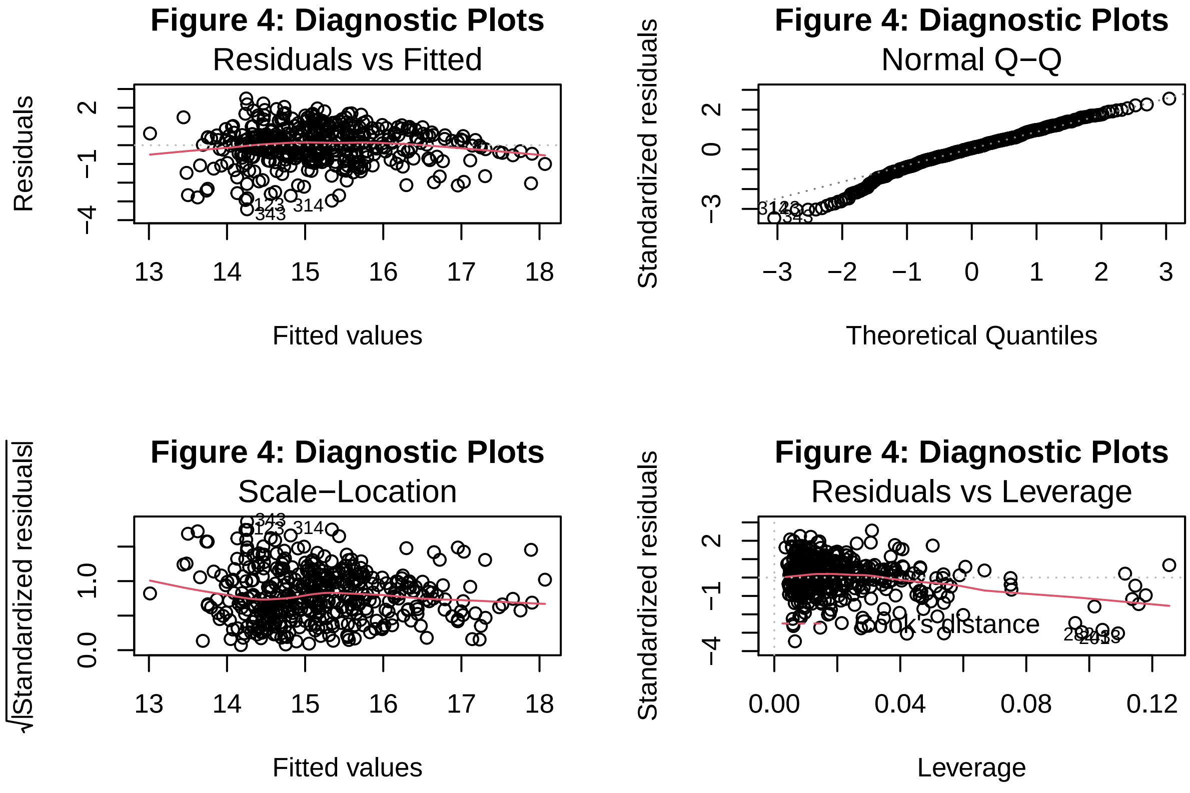
Given our data we choose to do a multiple linear regression on our variable of interest Salary. As mentioned in the data section, Salary shows a similar pattern to an exponential distribution so we take a log transformation on the value of salary. We hope to predict the value of a players salary so multiple regression is a good choice. Our dataset offers a lot of options of predictor values which we can use to help predict the response variable of log(salary).

We end up with three models. The first model is based upon looking only at simple per game statistics. The second looks at using more advanced statistics, and the third model combines both of them. The first model is based on the stats: Field Goal Percentage, 3 Point Percentage, 2 Point Percentage, Free Throw Percentage, Total Rebounds, Assists, Blocks, and Points per game. Where assists are the most significant variable, while based on correlation coefficients, Points is first with Total Rebounds following it as the highest correlation coefficients. Viewing the diagnostic plots the model satisfies assumptions.

Table 8: Per Game Model

term	estimate	std.error	statistic	p.value
(Intercept)	12.7808214	0.4562154	28.014885	0.0000000
FG.	-2.2397044	1.0670663	-2.098937	0.0364105
X3P.	1.1327285	0.4480188	2.528306	0.0118217
X2P.	2.1879591	0.8517554	2.568765	0.0105458
FT.	0.6294972	0.3996407	1.575158	0.1159623
TRB	0.1343258	0.0344639	3.897575	0.0001128
AST	0.1628662	0.0397010	4.102317	0.0000490
BLK	0.3415833	0.1701735	2.007265	0.0453527

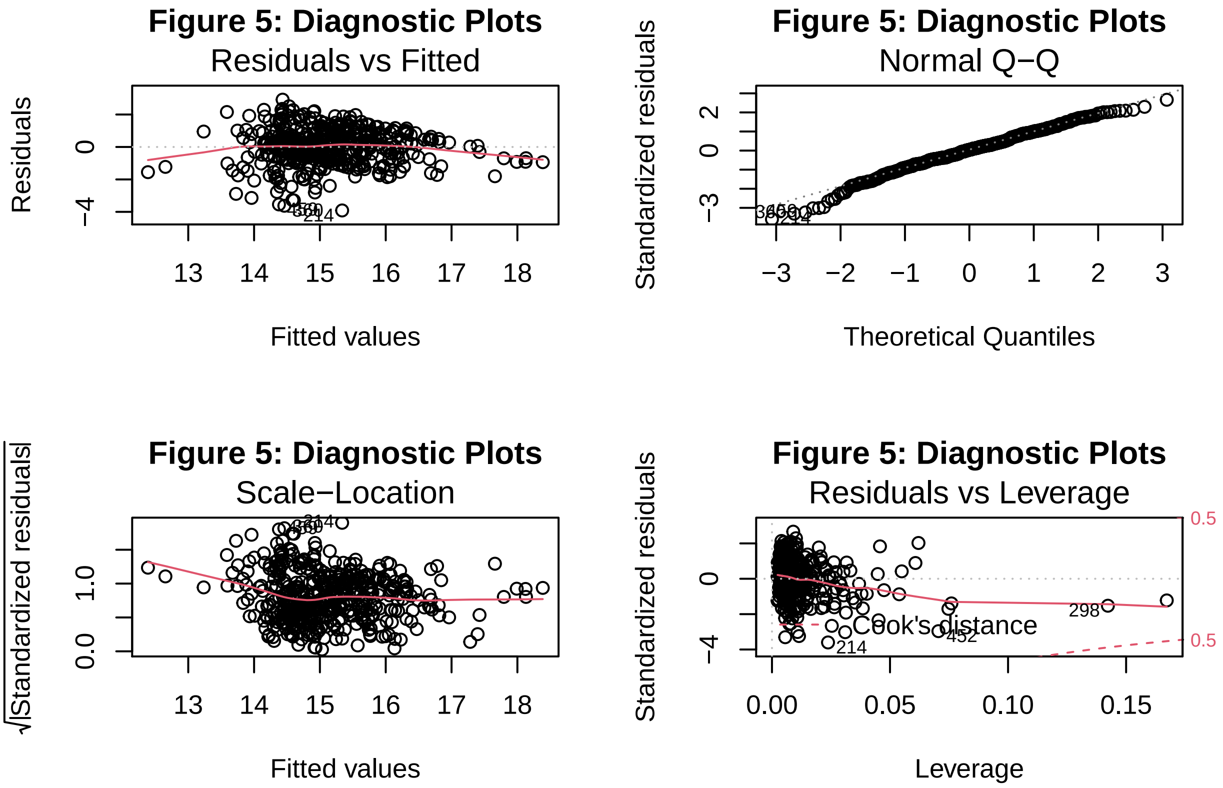
term	estimate	std.error	statistic	p.value
PTS	0.0448864	0.0139848	3.209664	0.0014294



The second model is based on the stats: Player Efficiency Rating, True Shooting Percentage, 3 Point Attempt Rate, Free Throw Attempt Rate, Assist Percentage and Win Shares. Where Win Shares are the most significant, while based on correlation coefficients, Win Shares has the highest correlation coefficient and Player Efficiency Rating second. Viewing the diagnostic plots the model satisfies assumptions.

Table 9: Per Game Model

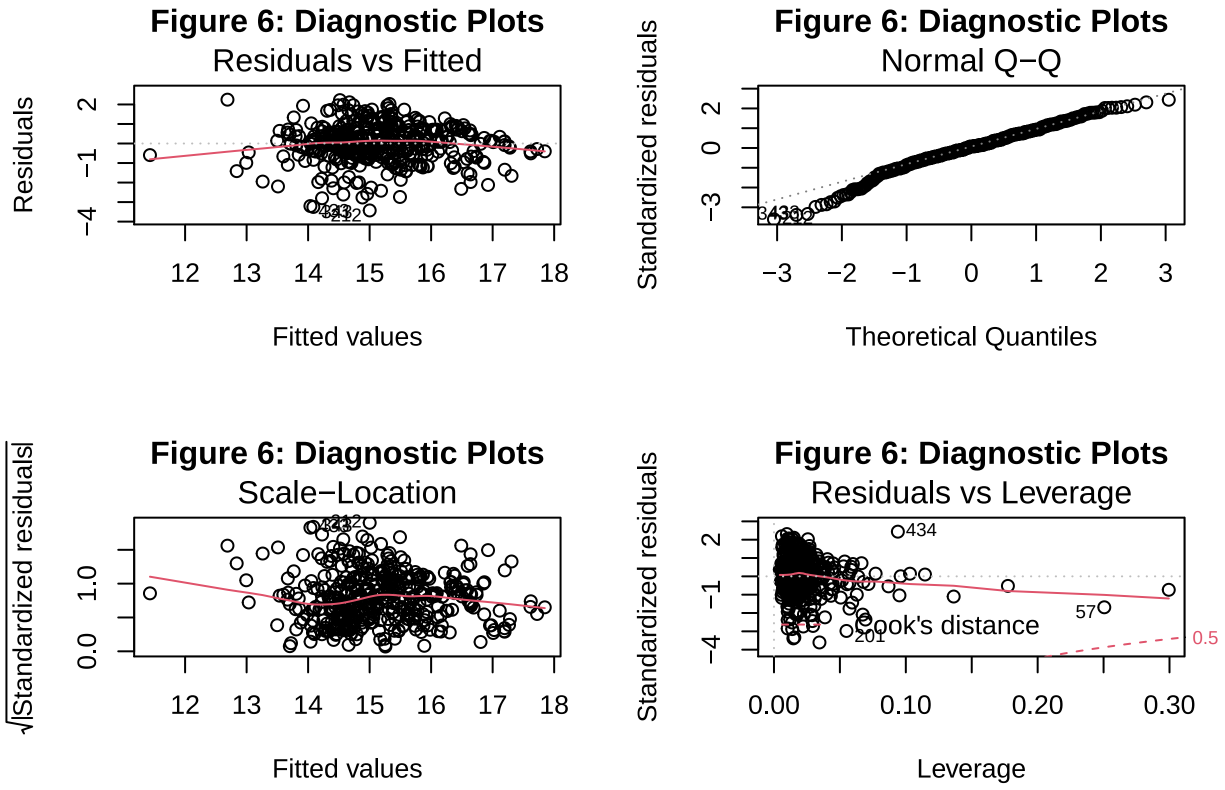
term	estimate	std.error	statistic	p.value
(Intercept)	14.1896750	0.1573244	90.193761	0.0000000
PER	0.0332279	0.0130926	2.537911	0.0114842
FTr	-1.4229820	0.3379744	-4.210325	0.0000307
AST.	0.0239028	0.0064415	3.710744	0.0002322
WS	0.2383535	0.0304772	7.820706	0.0000000



The third model is based on the stats: Free Throw Rate, Total Rebound Percentage, Steals Percentage, Usage Percentage, Assists, Steals, Blocks, Personal Fouls, and Points. Points is the most significant value, with Points and Win Shares having the highest correlation coefficient. Viewing the diagnostic plots the model satisfies assumptions.

Table 10: Per Game Model

term	estimate	std.error	statistic	p.value
(Intercept)	14.8356841	0.2842226	52.197419	0.0000000
FTr	-1.2287144	0.3494552	-3.516085	0.0004851
TRB.	0.0418581	0.0129583	3.230216	0.0013331
STL.	-0.2225049	0.1045334	-2.128553	0.0338658
USG.	-0.0581108	0.0151563	-3.834109	0.0001451
AST	0.1295977	0.0422912	3.064409	0.0023199
STL	0.6254740	0.2800453	2.233475	0.0260374
BLK	0.4098176	0.1616846	2.534672	0.0116131
PF	-0.1766121	0.0979700	-1.802716	0.0721411
PTS	0.1167484	0.0191630	6.092399	0.0000000



To produce the model we use stepwise model selection with both AIC and BIC, where we end up with the same model using both methods. Stepwise is a combination of both forward and backward selection. This is one method of variable selection, this choice allows us to lower the prediction error of the model.

Results

Looking at all three models, and the coefficient values. There seems to be values that have a larger effect by the values.

In the first model, variable selection chooses some percentage values as well as some per game statistics. Notably, the assists per game value is the most significant. This is not as surprising as assists correlate with getting points. A player with more assists can increase the offensive abilities of their whole team. This is where the importance of point guards, who direct the offense and often possess the ball the most, have an impact on the flow of the offense. Assists can also imply that the team's offense involves passing the ball around a lot, ball movement is often important in many strong offensive systems. Assists also could imply how a player is making the players around them better. Recall that based on correlation coefficients, points and total rebounds had been the two highest values. This is not surprising either as points are required to win the game. Further, many of the players with high salaries are the superstar caliber players who average high amounts of points every game. These are the star players who are at the top of their team. Similarly, rebounds is another important stat that can often have high values.

In the second model, variable selection chooses a mix of different advanced statistics. Recall win shares had the most significance, which by definition would be an important stat to base a salary on. Win shares is an estimate of the number of wins contributed by a player. Therefore it makes sense that a player who contributes to more wins would be worth more. Free throw attempt rate was second significant which also carries more meaning. A player who shoots more freethrows will be fouled more, this allows easier points and can cause foul trouble for opposing player.

In the third model, variable selection chooses a mix of variables including variables that were not chosen in the individual models above. The model still chooses points as the most significant value with win shares as the second significance similar to the previous models. When looking at the coefficients, some of them include negative correlations which can be surprising as most values should contribute to a positive increase in salary. This is one thing that could be further investigated.

Table 11.1: Lower Residuals

	Player	Residuals
197	Jeff Green	-0.00881685492946624
44	Bryn Forbes	-0.0157640751756584
125	Dwight Howard	-0.0208717731667279
80	Damian Lillard	-0.0230474703509008
16	Andre Drummond	-0.0242927342552393
54	Carsen Edwards	-0.0254587556752157

Table 11.2: Upper Game Model

	Player	Residuals
39	Brandon Knight	1.9316814627641
332	Otto Porter	1.96423622360949
415	Tyler Johnson	2.05007305780408
304	Michael Kidd-Gilchrist	2.18200363667793
13	Allen Crabbe	2.23069235626048
138	Evan Turner	2.49324382659728

Table 12.1: Lower Game Model

	Player	Residuals
157	Harrison Barnes	-0.0021441819313289
385	T.J. McConnell	-0.00393220945201203
295	Matisse Thybulle	-0.0256684810678366
407	Trae Young	-0.0323825997837884
53	Carmelo Anthony	-0.0351503897357301
47	Caleb Martin	-0.0509861597043264

Table 12.2: Upper Game Model

	Player	Residuals
443	Alec Burks	2.2682952657104
13	Allen Crabbe	2.28041479219898
320	Nerlens Noel	2.29135962893012
260	Kyle Lowry	2.34677517094556
60	Chris Boucher	2.50729154821752
32	Bogdan BogdanoviÄ‡	2.91506491135146

Table 13.1: Lower Game Model

	Player	Residuals
79	Damian Jones	-0.00867661079159424
262	KZ Okpala	-0.0134550913470961
432	Zach LaVine	-0.0180988416812163
54	Carsen Edwards	-0.0194974483849172
50	Cameron Johnson	-0.0330764597466685
132	Enes Kanter	-0.0331414513028195

Table 13.2: Upper Game Model

	Player	Residuals
13	Allen Crabbe	1.96093077672119
39	Brandon Knight	1.98937534775467
29	Blake Griffin	2.02508038111754
415	Tyler Johnson	2.1110984251087
138	Evan Turner	2.22132603945138
434	Zhaire Smith	2.24475090088665

Looking at the residual values, a player who is underpaid based on the model will have a larger negative value, and a player who is overpaid will have a large positive value. Comparing models, it can be seen that players show up in the top 5 overpaid or underpaid for more than one model. For instance players like Allen Crabbe and Brandon Knight are deemed underpaid in both models.

Discussion

Summary:

After cleaning and removing variables with collinearity, three models were created to predict the value of the log transformed salary values. The models showed similarity to each other, and had similar predictions.

Conclusion:

The main variables in the prediction of models were Points and Win Shares, a player's value in their salary is mainly concerned with these two values. Simplistically, a player is paid based on their contribution to winning. A player who scores a large amount of points should be getting paid a large amount for those abilities. Win shares captures the value of more variables that can help sum up their importance into one single statistic. Thus a player contributing all over through other statistics like blocks or assists, can change their win share values. The models show that even with more advanced statistics like how much a player attempts three point shots, or their Usage percentage the best model comes from its simplicity. At the end of the day, the best team is the team with more points. The player that helps achieve those more points, often themselves individually scoring, will receive the most money.

Weakness and next step:

Some of the weakness the model suffers come from the choice in data cleaning. Players with NA values were omitted and justified due to having low amount of games played, this could be looked at again, to see if there was any influence. Similarly, the values in the pergame statistic are calculated based on games played. This could potentially lead to values being higher in certain contexts. Say a player who only plays a couple games, but in those couple games their stats are quite high. Would this player still be able to maintain those stats given more games? Similarly with percentages, a player may have a high percentage due to small amount of the stat taken (ex: 1 three point made, 1 three point attempted) which could influence the model.

Another weakness is that most NBA players will sign a 3-5 year contract, the model does not take this into any consideration. This can skew the model because a player could be on decline from their initial season of the contract and progressively gets worse even if the money stayed the same or increased per year on the contract. This could lead to a player being deemed overpaid and underperforming. Salary also has more depth, as salary can be quite relative to the team. One team could offer more to a player based on their current salary limits (salary cap).

The next steps would include looking at models that can explain more with other information, than the statistics. More statistics could be contributed, for instance a newer model can look at how many years a players has been in the league or what position they are. Looking at how many seasons the player plays, intuitively is a good choice. Since many rookie contracts are quite low, and veteran players will take a smaller wage to play on championship caliber teams. Another model which takes into consideration the position of the player could also be insightful, depending on the position certain stats are more expected. For instance, a power forward or center is expected to get more rebounds, thus their average is higher which could be influencial. Similarly, with point guards often getting more assists.

Furthermore, this analysis only looked at the 2019-2020 season, potentially looking at other seasons as a whole or individually could show different results.

References

All of the information relative to the basketball statistics and data used come from the website basketball-reference.com, specifically these three websites:
http://web.archive.org/web/20201113205930/https://www.basketball-reference.com/contracts/players.html https://www.basketball-reference.com/leagues/NBA_2020_per_game.html https://www.basketball-reference.com/leagues/NBA_2020_advanced.html

This is analysis was written using r, and other r libraries. These include stringr, dplyr, corrplot, knitr, and broom.