Final Paper

Kenneth (Hsuan) Chen

6/11/2021

Feature	Response
Name	Kenneth (Hsuan An) Chen
SID	005544529
Kagge Nickname	Kenneth (Hsuan An) Chen
Kaggle Rank	7
Kaggle R^2	0.63156
Total Number of Predictors	21
Total Number of βs	22
BIC	12523.91
Complexity Grade	108

Abstract

This project aimed to create a multiple linear regression model that could successfully predict a professional basketball player's salary given various statistics that measure the value of a player. Before building the model, we first examined the data and transformed various predictors as well as created new predictors in order to provide better information for the model. We then took a look at the predictors that were most correlated with the response variable and proceeded to create interactions that increased the accuracy of the model. Finally, we chose the best predictors for the model via exhaustive selection over 50 different variables and optimizing with adjusted R^2.

After some fine-tuning and minor adjustments, we were able to produce a model with $R^2 = 0.7987$ with the training dataset and $R^2 = .63156$ with the test dataset. This model was then able to place 7^{th} out of 52 on Kaggle (Kaggle name: Kenneth (Hsuan An) Chen) with a total of 21 predictors. Overall, although there were some issues regarding the validity of the model and the dataset as a whole, we believe that, in terms of prediction, the model was quite accurate at predicting salary from player statistics given the constraints of this project.

Introduction

The National Basketball Association is a professional basketball league comprising of 30 North American teams (29 in the United States and 1 in Canada). For the 2020-2021 NBA season, Stephen Curry was the highest paid player, making over $43 million that season (ESPN). Thus, it is quite clear that NBA players are paid a hefty sum each year as part of their NBA salaries. However, with the large amount of money involved, are players paid appropriately for the amount that they contribute on the court? In other words, how strong of a relationship does the performance of a player have with the amount that they are paid at the end of the season (Almohalwas 2021, 2)? This is the question that we would like to answer in this project.

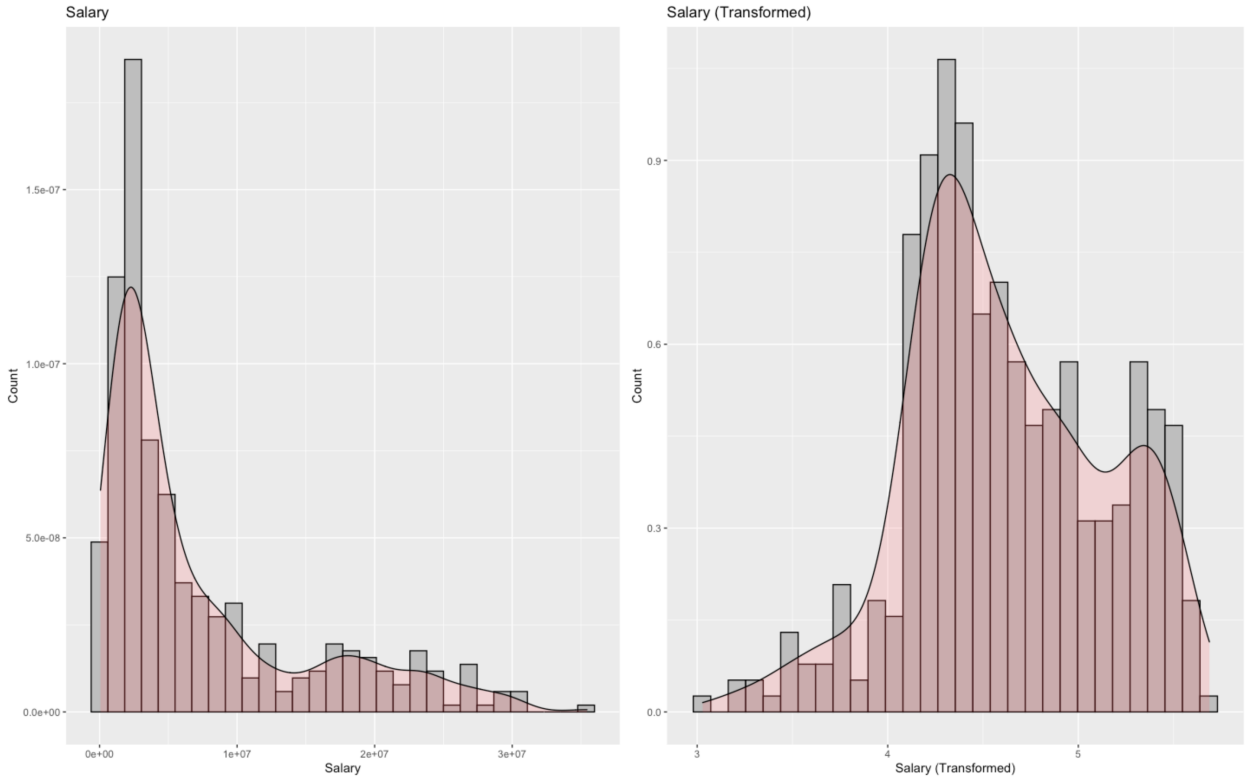
The training data set in question consists of 420 observations and 68 columns containing various advanced game statistics that help describe a player's value and the test data set consists of 180 observations and 67 columns (no response variable). Specifically, there are 60 numeric predictors, 6 categorical predictors, the response variable Salary, and one column Obs that simply indicates the row number (Kaggle 2021). Furthermore, "this dataset is oriented from Kaggle," and it has been cleaned and imputed of its missing values by the Professor; the players that were traded to other teams during the season had their performance data combined from all the teams that they played for (Almohalwas 2021, 2-3).

Methodology

Response Variable Salary

Upon inspection of the response variable, we see that like most salary/income data, it is heavily skewed. Thus we try to transform the variable to have a normal distribution using the Box-Cox transform below:

	Est Power	Rounded Pwr	Wald Lwr Bnd	Wald Upr Bnd
Y1	0.1062188	0.11	0.0342216	0.178216



After transforming Salary with $\lambda = 0.1$, we see that although the response variable is no longer as skewed as before, it does not necessarily resemble a normal distribution. Thus, the decision was made to leave the response variable as it is now, and see if a transformation is necessary after we fit a model.

Summary statistics for Salary:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
65019	1982576	3771013	7227809	9532681	35439394

Feature Engineering

Before we can dive further into the predictors, we must first transform a few variables. Since the demography of NBA players is mostly American while the rest of the countries have relatively few observations each, we transform NBA_Country into a variable that simply indicates whether a player is American or not:

NBA_Country	Freq
No	76
Yes	344

Further, we see a similar situation with the variable Pos where the hybrid positions PG-SG and SF-SG only have one and four observations, respectively. Thus we transform Pos by reclassifying PG-SG as SG and SF-SG as SF:

Pos	Freq
C	75
PF	77
PG	100
SF	75
SG	93

The remaining variables that were transformed/created are, for the most part, either additional statistics that were not included in the original dataset (but could be calculated using the original predictors) or just spontaneous ideas that popped up given our background knowledge of the NBA and basketball. These variables are summarized in the table below:

Variable	Description
MPG	Minutes Per Game
GS.	% of games that the player started
Superstar	"Yes" if player was in top 20% of player ranking
Aggressive	"Yes" if player averaged more turnovers than mean turnovers per game
PTS.Total	Total number of points a player scored in a season
Winning	"Yes" if player's team had a winning season
Traded	"Yes" if player was traded
Center	"Yes" if player plays center
Playoffs	"Yes" if player's team made the playoffs
AST.TOV	Assist to Turnover ratio
is.PG	"Yes" if player plays point guard
eFG.	Effective Field Goal %
four.factors	Dean Oliver's "Four Factors of Basketball Success"
GmSc	Game Score
TSA	True Shooting Attempts
ASTr	Assist % (I did not realize this was already in the data set)
tT.Div	Grouped Division into either Pacific, Southwest, or Other
tPos	Grouped Position into either Frontcourt or Backcourt

Variable	Description
AvgTMSalary	Average salary of player's team
Role	Grouped players into either Benchwarmer, Role Player, or Starter given their MPG value

Notes:

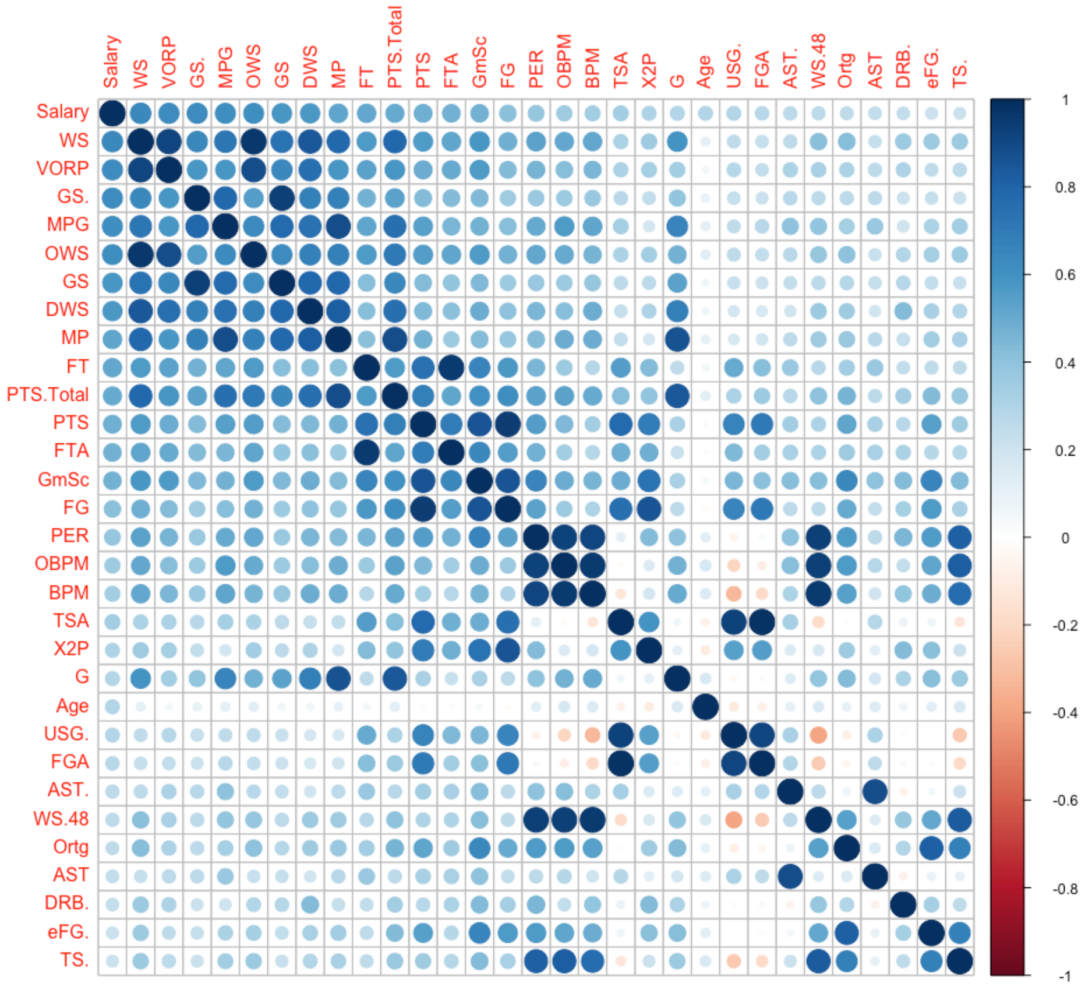
1. More information about the advanced statistics can be found on basketball-reference.com (Basketball Reference).
2. If a variable has "Yes" for one of its values then the only other value for that variable would be "No."
3. AvgTMSalary was not created until the last minute so most of the methodology will not involve this variable.

Numeric Predictors

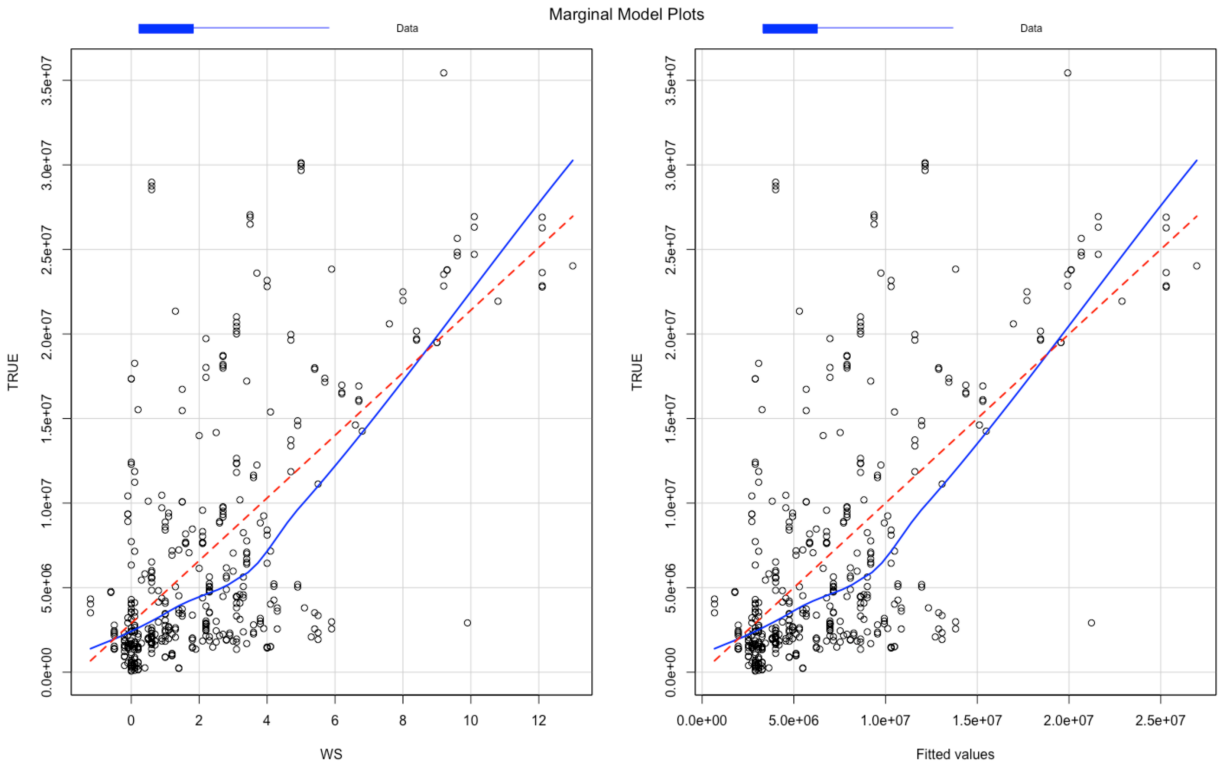
We now examine the numeric predictors. Below we have the 10 numeric predictors that are most correlated with Salary:

Variable	Correlation
WS	0.6464196
VORP	0.6262785
GS.	0.6218517
MPG	0.6158582
OWS	0.6118452
GS	0.5827094
DWS	0.5708446
MP	0.5234995
FT	0.5189536
PTS.Total	0.5022498

and a correlation plot of the 30 numeric predictors that are most correlated with Salary:



From the correlation chart, we see that the numeric predictors are, in general, not strongly (linearly) related to Salary. For instance, if we look at the MMPS plot of a model that uses WS to predict Salary:



we see that WS, the predictor most correlated to Salary, does not seem to have a linear relationship with the response variable. Lastly, not only are the numeric predictors not very linearly related to Salary, we see from the correlation plot that the predictors that are most correlated to the response variables are also quite correlated with each other. This may cause problems related to multicollinearity somewhere down the line.

Categorical Predictors

In total, we have 17 categorical predictors. One of the biggest issues that we had with the categorical predictors was that some of the predictors had too many levels. For instance, Team and TM had 26 distinct levels each. We addressed this problem in two different ways. The first method was to simply transform the variable to have a far fewer number of levels (like with NBA_Country). The second method was to "encode" the variable into a discrete, numeric predictor (like with AvgTMSalary). Both methods were able to preserve most of the information as well as reduce the number of betas produced by a single categorical predictor.

Interaction Terms

In order to compensate for the relatively "weak" predictors in the original dataset, our model had to rely heavily on interaction terms. Furthermore, in order to be have more of a "creative" advantage, we decided to not only utilize numeric-categorical interations, but numeric-numeric interactions as well. For instance, the interaction MPG:Age was consistently one of the most significant predictors that we had in the various models that we tried.

For the selection of numeric-numeric interaction pairs (let's say a pair consists of predictors A and B), we would select predictor A to be one that was among the most correlated with Salary and we would select predictor B to be not as correlated with Salary but also not correlated with A (thus A and B would be independent). This was done by referring to the correlation plot above with the 30 predictors most

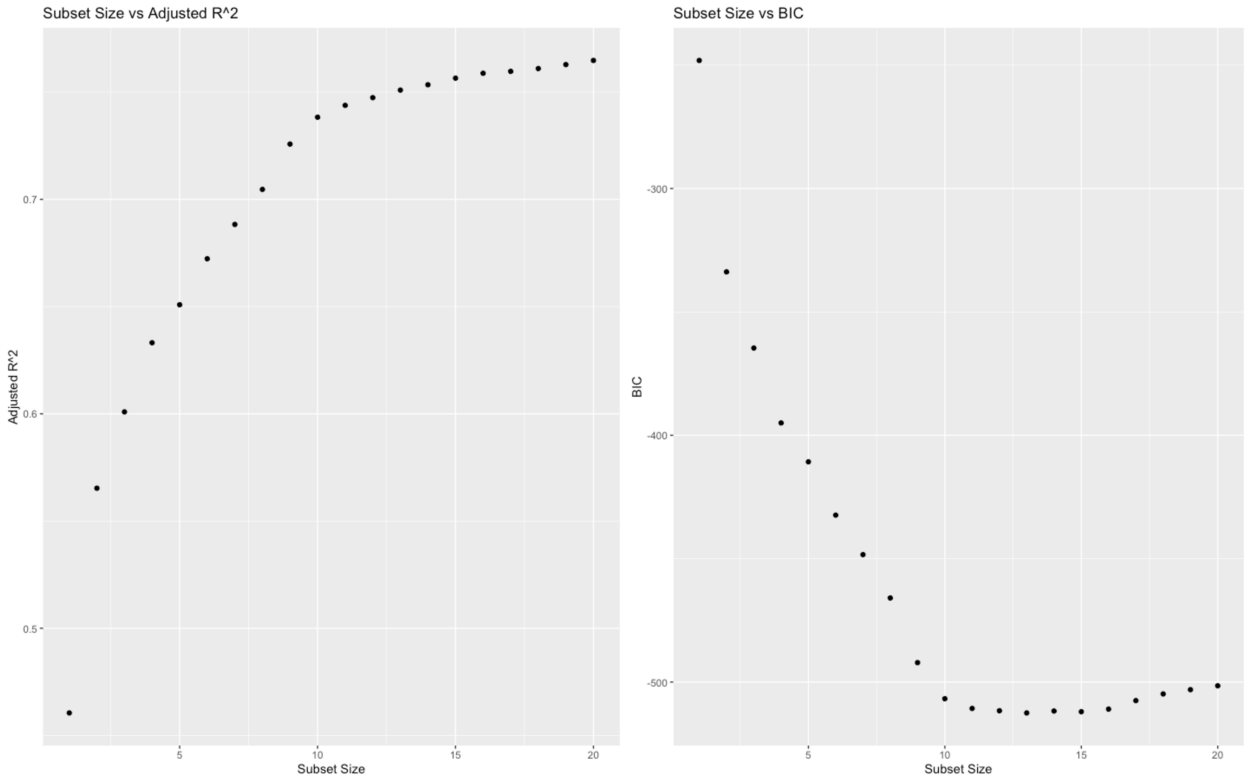
correlated with Salary. We then kept interaction pairs that had the most significant p-values when we conducted partial F-tests. Specifically, in the partial F-test, the reduced model could be the model we are testing at the time (with a given selection of predictors) and the full model would simply be the reduced model with the additional interaction. Among the most consistently significant interactions were MPG:Age and VORP:WS.48.

For the selection of numeric-categorical interaction pairs, we paired each of the categorical predictors with the numeric predictors that were in the correlation plot above and kept the pairs that displayed a significant effect when plotted and/or had a significant p-value when we conducted partial F-tests. Two examples are shown below:

From the scatter plots above we see that the interactions are quite significant so we put these pairs into consideration when selecting our predictors.

Feature Selection

For feature selection, the most effective method that we found was exhaustive selection. Specifically, we ran exhaustive selection from a choice of 50 total predictors that included various interactions, numeric predictors, and categorical predictors; and we chose 20 predictors to be the largest subset size chosen. Below we plotted subset size against adjusted R^2 and BIC, respectively:



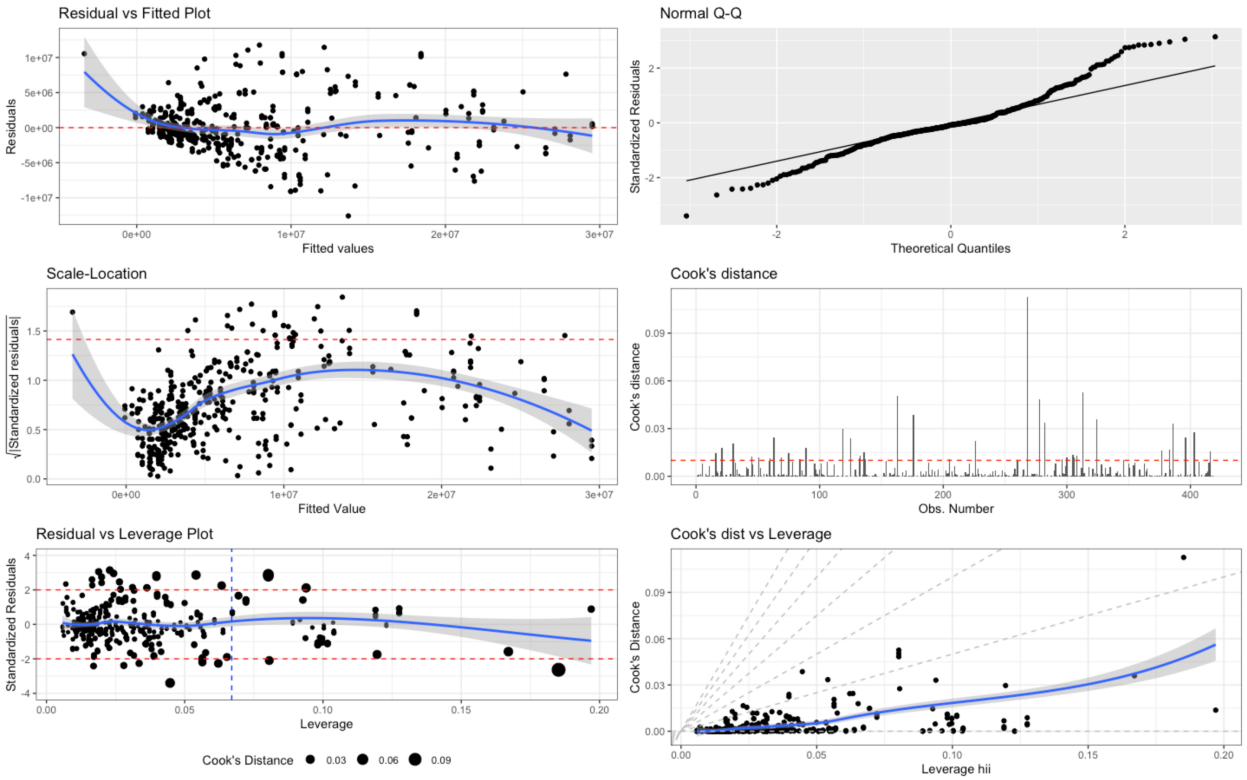
From the graphs, we see that a subset size of 13 predictors yields the lowest BIC and a subset size of 20 predictors yields the highest adjusted R^2.

Building the Model

Model 1

We first build a model based on the predictors selected from exhaustive selection by optimizing BIC (14 βs total).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1960007.281	564200.5975	3.473955	0.0005680
MPG	-1854072.220	191011.2839	-9.706611	0.0000000
OWS	-1133228.704	283332.6757	-3.999640	0.0000754
G	436984.430	67204.4212	6.502317	0.0000000
Age:MPG	76249.730	6901.3969	11.048449	0.0000000
Age:G	-18080.151	2488.4594	-7.265600	0.0000000
MPG:FT	20629.077	5720.0131	3.606474	0.0003490
MPG:GS	4253.941	569.7383	7.466482	0.0000000
T.W.L.PERC:WS	1021375.896	393043.2759	2.598635	0.0097003
GS:AST.TOV	-41107.037	7684.8887	-5.349074	0.0000001
tT.DivSW:DWS	-1548544.280	327227.8750	-4.732312	0.0000031
TradedYes:OWS	1774992.014	299997.0863	5.916697	0.0000000
VORP:WS.48	10386430.391	2276974.0093	4.561506	0.0000067
AST.:tPosFrontcourt	182421.354	33691.5900	5.414448	0.0000001



Note: The code for the diagnostics plot was taken from the Professor's Chapter 5 notes (Almohalwas 2020, 2)

We inspect the bad leverage points from this model:

Obs	Age	G	MPG	FG	FGA	AST	PTS	TRB	STL	BLK	TOV	PF	Salary
163	30	12	31.08333	8.7	22.8	6.6	27.8	3.7	1.6	0.4	2.4	3.3	28750441
268	23	71	27.53521	10.7	16.4	1.6	24.8	19.4	1.4	3.3	2.5	4.5	2917360
278	30	12	31.08333	8.7	22.8	6.6	27.8	3.7	1.6	0.4	2.4	3.3	28533996
313	30	12	31.08333	8.7	22.8	6.6	27.8	3.7	1.6	0.4	2.4	3.3	28974644
386	29	51	31.98039	12.6	25.5	9.2	39.8	7.7	2.4	0.2	4.5	3.4	35439394
403	32	78	33.50000	7.8	13.5	1.8	18.3	10.7	1.1	1.1	1.6	4.0	14252702

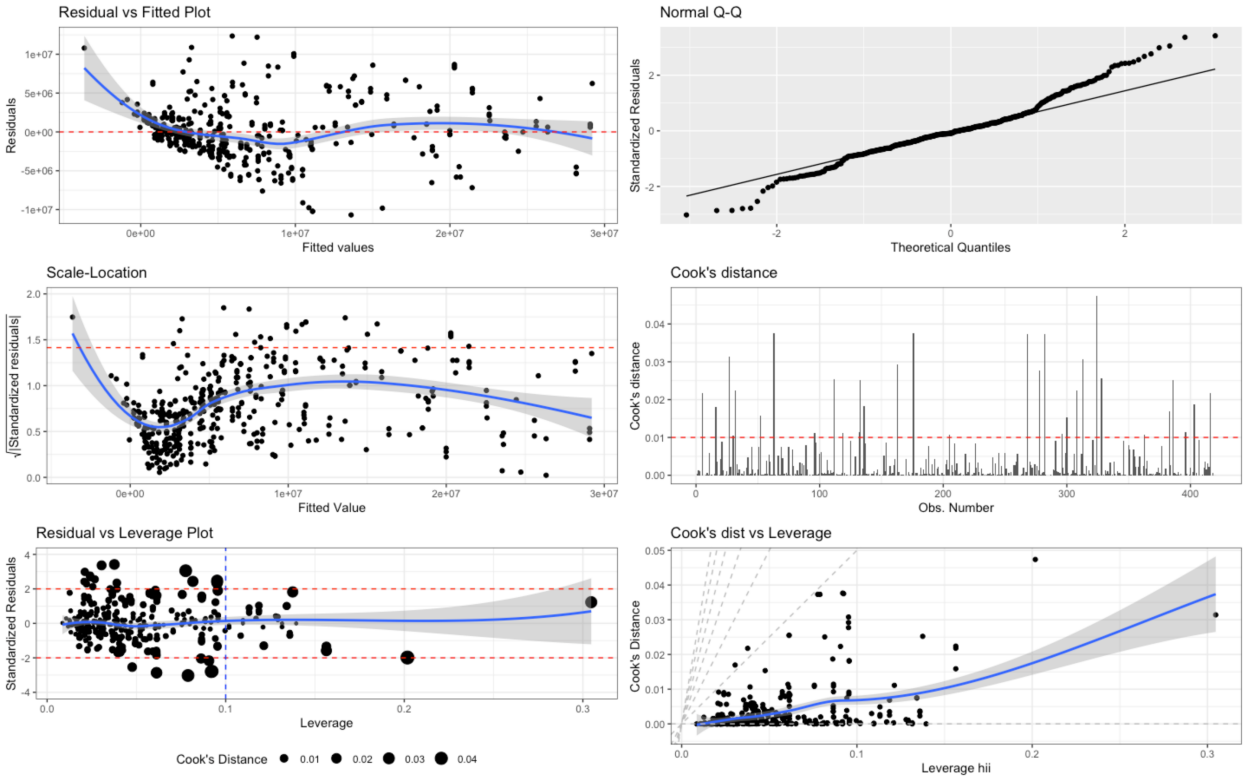
Aside from observation 403, the bad leverage points are extremely suspicious. In particular, observations 163, 287, and 313 seem to be the duplicate observations but each with slightly different salaries for some reason. Further, observation 278 averaged 19.4 rebounds over 71 games and observation 386 averaged 39.8 points over 51 games. Both of these accomplishments have never been done by anyone in the NBA in the last 30 years (I checked) and so we can probably remove those points from the model.

Observations	Training R^2	Training Adjusted R^2	Test R^2 (Kaggle)
415	0.7625	0.7548	0.55533

Model 2

For our second model, we built it based on the predictors selected from exhaustive selection by optimizing R^2_{adj} (21 βs total).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2988817.925	1061697.2704	2.815132	0.0051174
GS.	3314819.226	1389485.0543	2.385646	0.0175149
MPG	-1945561.665	191336.6871	-10.168263	0.0000000
FG	1077676.573	273598.0318	3.938905	0.0000966
G	452788.762	66570.0658	6.801687	0.0000000
TS.	-3623144.342	1750290.3578	-2.070025	0.0390929
X3P	-670930.223	238068.3180	-2.818226	0.0050693
TRB.	-191685.085	59535.2390	-3.219691	0.0013883
Age:MPG	76596.555	7046.2799	10.870496	0.0000000
Age:G	-18129.485	2510.0128	-7.222866	0.0000000
MPG:FT	27819.280	6445.5555	4.316041	0.0000201
MPG:GS	3948.863	701.9621	5.625464	0.0000000
PTS:X2P	-38987.801	8836.1368	-4.412313	0.0000132
FG:TradedYes	-961208.506	278395.8317	-3.452669	0.0006145
X2P:TradedYes	1344368.089	332231.8581	4.046475	0.0000624
GS:AST.TOV	-66381.755	8189.9601	-8.105260	0.0000000
T.ConfW:VORP	-1286395.460	320027.6931	-4.019638	0.0000697
GS:tT.DivP	53130.664	14156.8631	3.752997	0.0002007
TradedYes:OWS	1910860.731	336140.0979	5.684715	0.0000000
VORP:WS.48	15716609.549	1701993.2008	9.234238	0.0000000
AST.:tPosFrontcourt	215853.551	41856.3671	5.157006	0.0000004

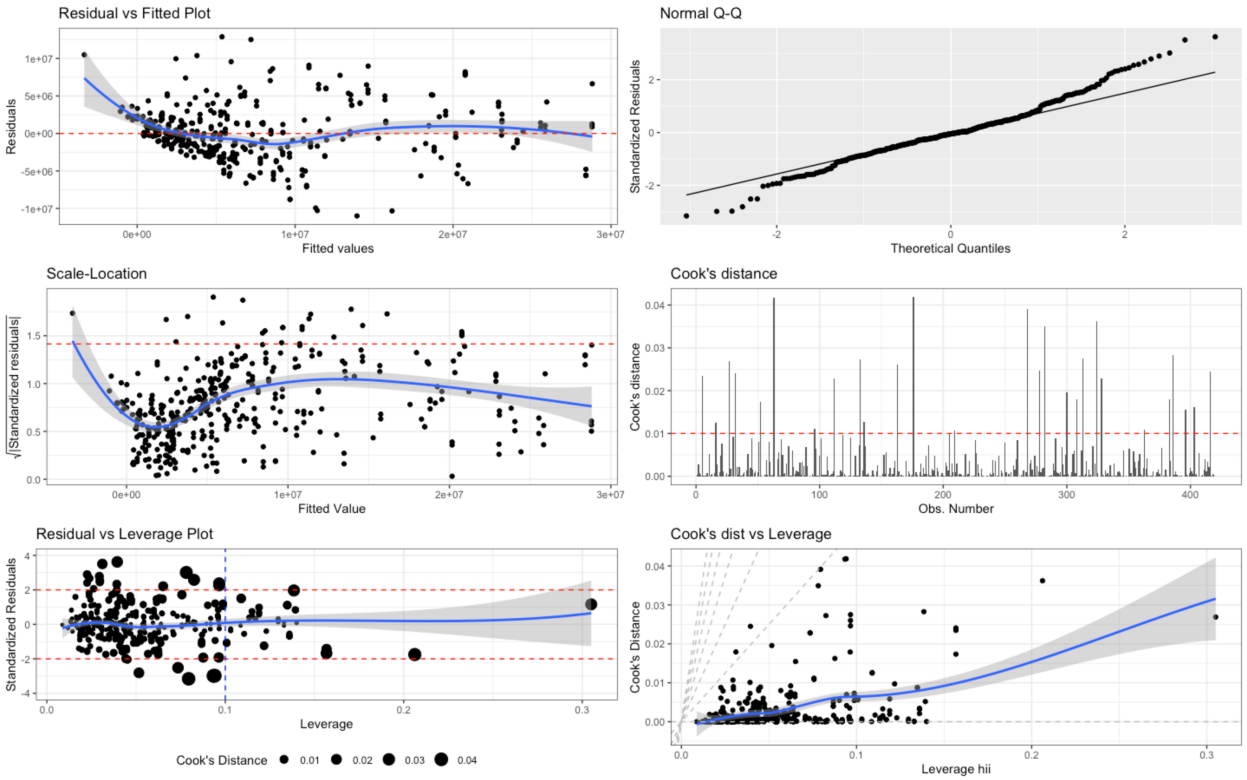


Observations	Training R^2	Training Adjusted R^2	Test R^2 (Kaggle)
420	0.776	0.7648	0.59895

Model 3

For our third model, we decided (at the last minute) to create a variable that incorporated the information that Team provided but without the large number of βs that came with adding Team to our model. Thus, we decided to create the AvgTMSalary (Average Salary per Team) variable that both distinguishes observations on different teams and relates that difference to the salary a player earns. Since model 2 had the higher score on Kaggle, we added AvgTMSalary to model 2 (thus 22 βs total).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.397992e+05	1.220401e+06	0.6061938	0.5447317
GS.	3.232997e+06	1.369544e+06	2.3606374	0.0187250
MPG	-1.725822e+06	1.983016e+05	-8.7030126	0.0000000
FG	1.021931e+06	2.700831e+05	3.7837659	0.0001782
G	3.845525e+05	6.831792e+04	5.6288664	0.0000000
TS.	-3.916948e+06	1.726882e+06	-2.2682201	0.0238509
X3P	-5.790163e+05	2.360194e+05	-2.4532569	0.0145849
TRB.	-1.789112e+05	5.878103e+04	-3.0436899	0.0024921
Age:MPG	6.887202e+04	7.271640e+03	9.4713190	0.0000000
Age:G	-1.557906e+04	2.574176e+03	-6.0520548	0.0000000
MPG:FT	2.839515e+04	6.354204e+03	4.4687184	0.0000103
MPG:GS	3.600604e+03	6.985968e+02	5.1540518	0.0000004
PTS:X2P	-3.855360e+04	8.708958e+03	-4.4268903	0.0000124
FG:TradedYes	-9.669350e+05	2.743669e+05	-3.5242404	0.0004740
X2P:TradedYes	1.357354e+06	3.274384e+05	4.1453715	0.0000415
GS:AST.TOV	-6.206818e+04	8.160727e+03	-7.6057171	0.0000000
T.ConfW:VORP	-1.254024e+06	3.155205e+05	-3.9744618	0.0000838
GS:tT.DivP	4.358171e+04	1.420440e+04	3.0681851	0.0023009
TradedYes:OWS	1.868181e+06	3.314843e+05	5.6358042	0.0000000
VORP:WS.48	1.530649e+07	1.681241e+06	9.1042799	0.0000000
AST.:tPosFrontcourt	2.183441e+05	4.125579e+04	5.2924474	0.0000002
AvgTMSalary	3.199281e-01	8.935920e-02	3.5802461	0.0003857



Observations	Training R^2	Training Adjusted R^2	Test R^2 (Kaggle)
420	0.783	0.7715	0.62339

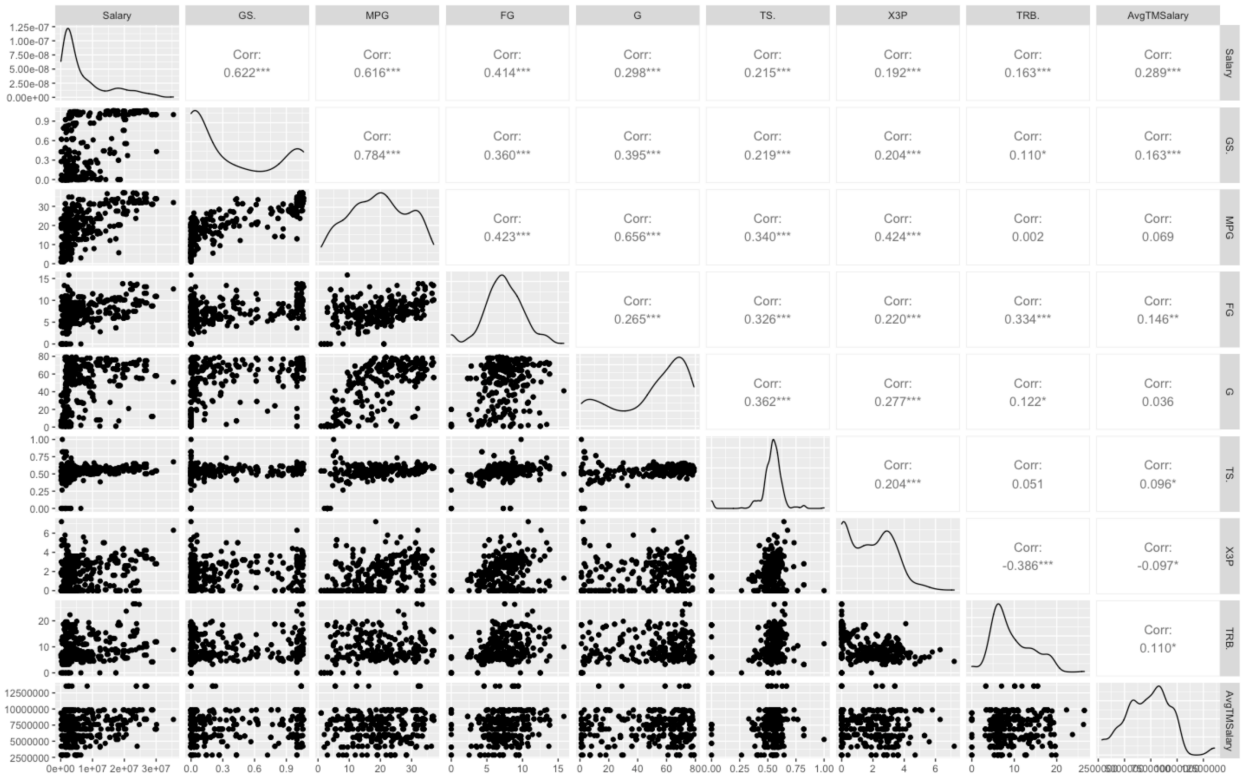
Since model 3 produced the highest score on Kaggle, we decided to focus our attention here.

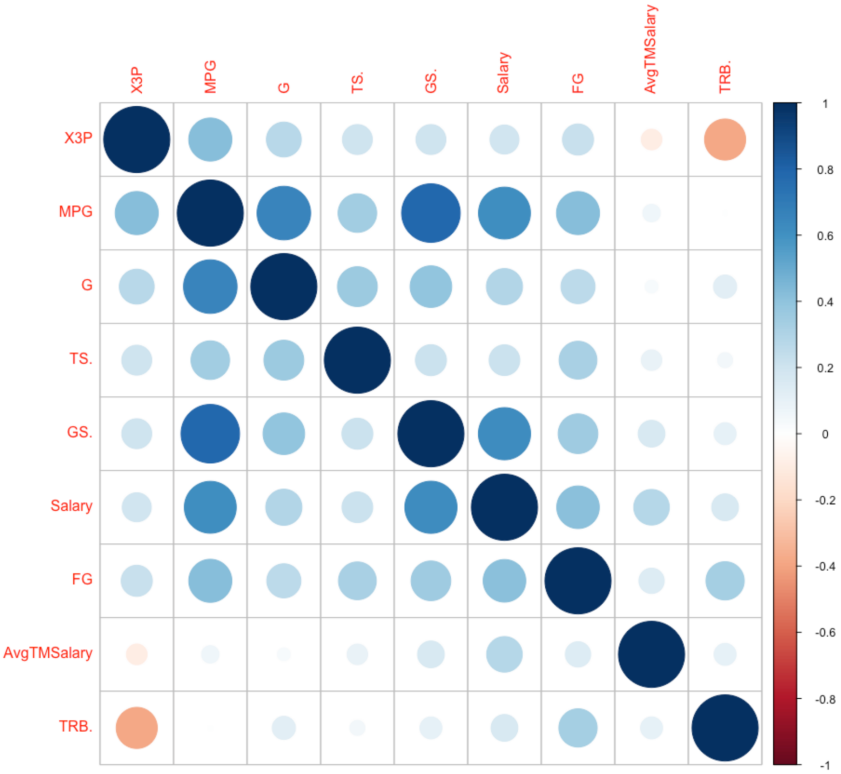
Checking Validity

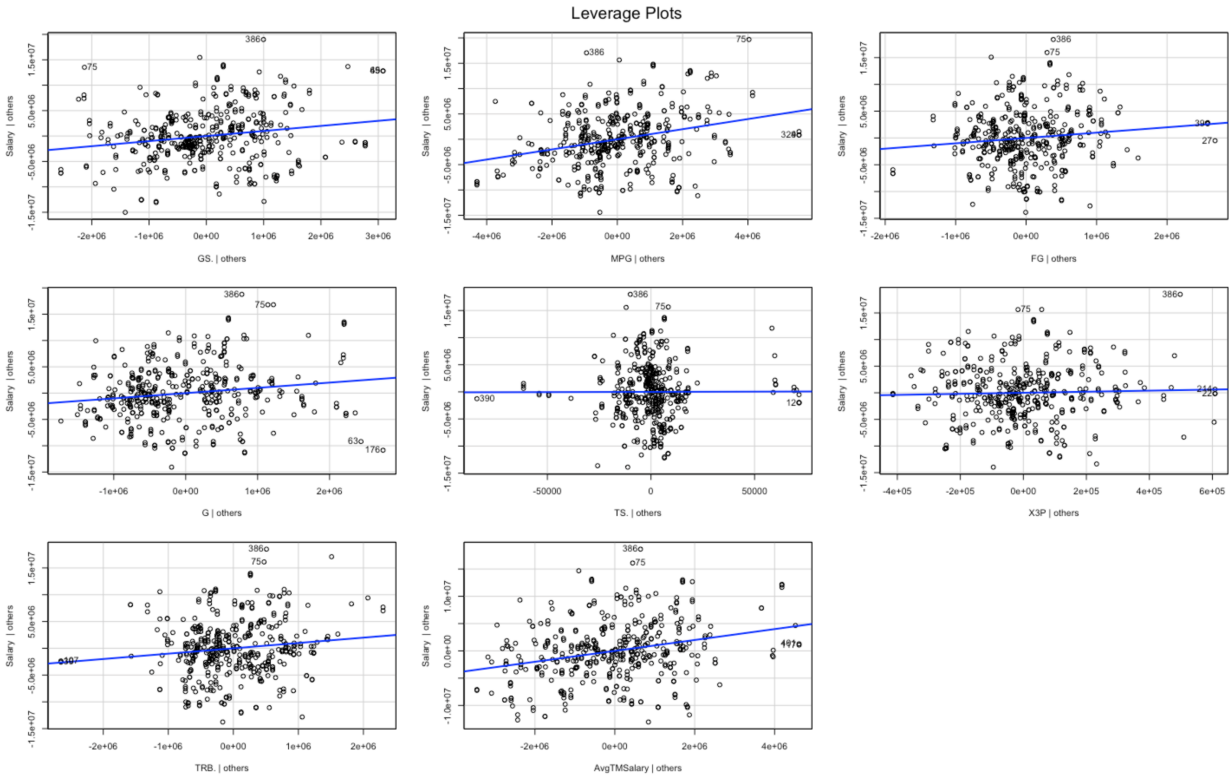
Multicollinearity

We first check for multicollinearity by examining a matrix plot, a correlation plot, a leverage plot, and a VIF chart of our numeric predictors from model 3:

Variable	VIF
GS.	3.072998
MPG	5.151768
FG	1.608806
G	2.066431
TS.	1.250938
X3P	1.663405
TRB.	1.594652
AvgTMSalary	1.072422


Salary
GS.
MPG
FG
G
TS.
X3P
TRB.
AvgTMSalary
Corr: 0.622***
Corr: 0.616***
Corr: 0.414***
Corr: 0.298***
Corr: 0.215***
Corr: 0.192***
Corr: 0.163***
Corr: 0.289***
Corr: 0.784***
Corr: 0.360***
Corr: 0.395***
Corr: 0.219***
Corr: 0.204***
Corr: 0.110*
Corr: 0.163***
Corr: 0.423***
Corr: 0.656***
Corr: 0.340***
Corr: 0.424***
Corr: 0.002
Corr: 0.069
Corr: 0.265***
Corr: 0.326***
Corr: 0.220***
Corr: 0.334***
Corr: 0.146**
Corr: 0.362***
Corr: 0.277***
Corr: 0.122*
Corr: 0.036
Corr: 0.204***
Corr: 0.051
Corr: 0.096*
Corr: -0.386***
Corr: -0.097*
Corr: 0.110*


X3P
MPG
G
TS.
GS.
Salary
FG
AvgTMSalary
TRB.
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1



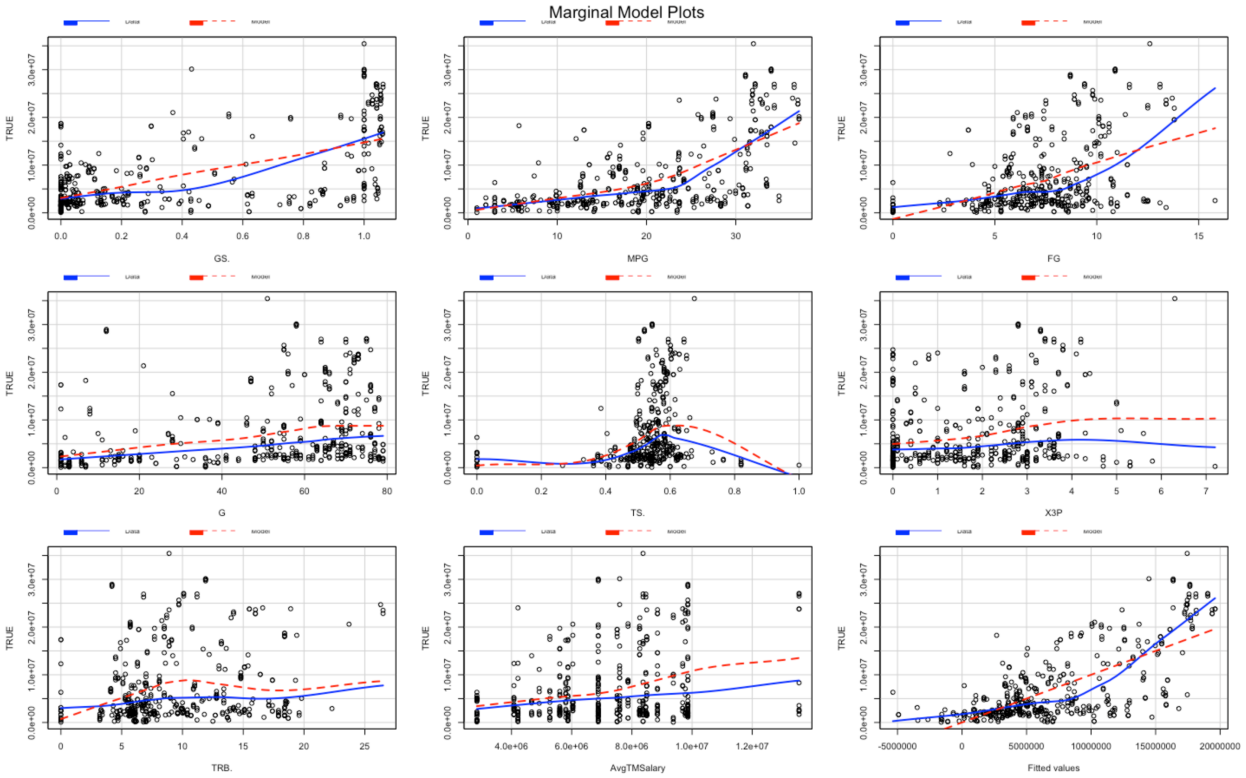
From the VIF chart, we see that there may be an indication of multicollinearity with the MPG variable since $VIF_{MPG} > 5$. Upon inspection of the matrix plot and the correlation plot, we see that MPG is not highly correlated with any of the remaining numeric predictors except GS% (r = 0.784). However, if we run a partial F-test by removing MPG from model 3, we see that there is an extremely significant difference between the reduced model and the full model:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	399	6.2428e+15					
2	398	5.2447e+15	1	9.9811e+14	75.742	< 2.2e-16	***

Furthermore, we do not see any strong patterns in the Leverage plots that would suggest to us that there is high multicollinearity. Lastly, since VIF_{MPG} is only slightly above 5, and 5 is not exactly a strict cutoff (some may even recommend a VIF cutoff of 10), we decide to keep the variable MPG instead of removing it.

Linearity

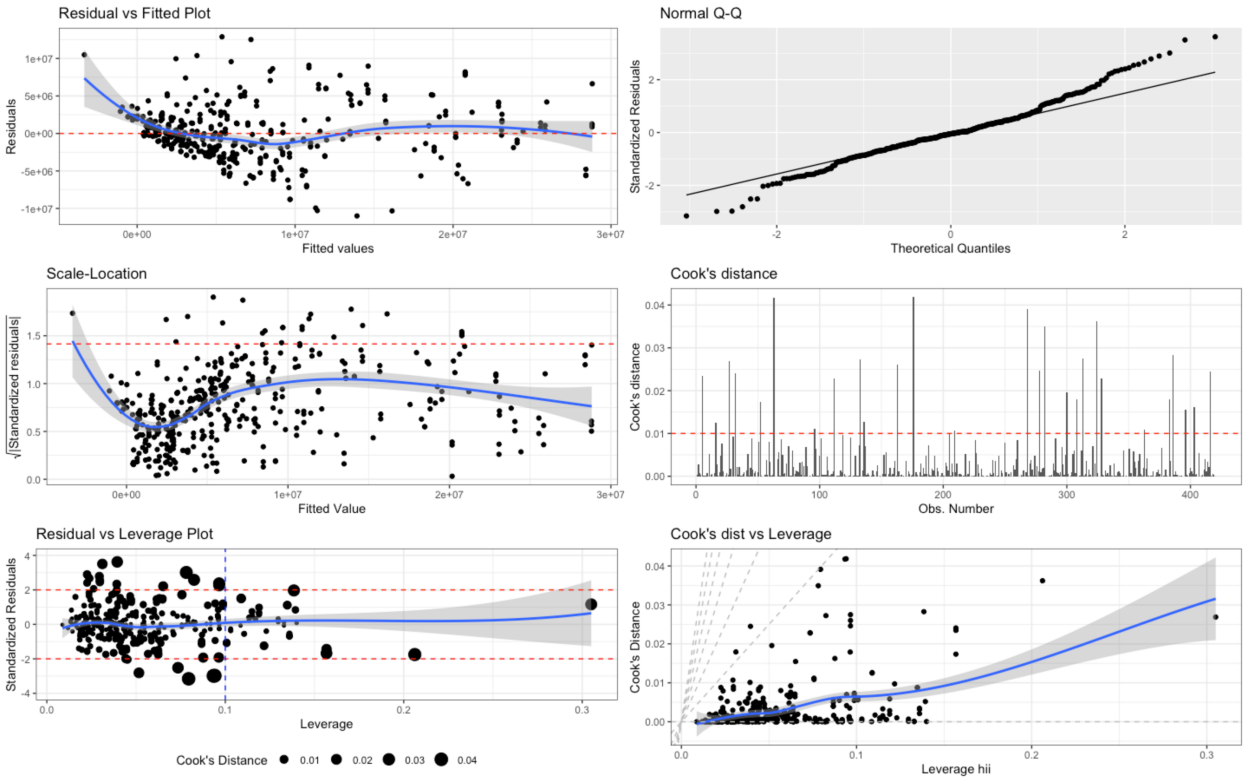
We inspect the MMPS plot of model 3:



As stated previously, it was difficult to find predictors that had a strong linear relationship with the response variable, so we decided to compensate for this by relying heavily on interactions.

Residual Analysis

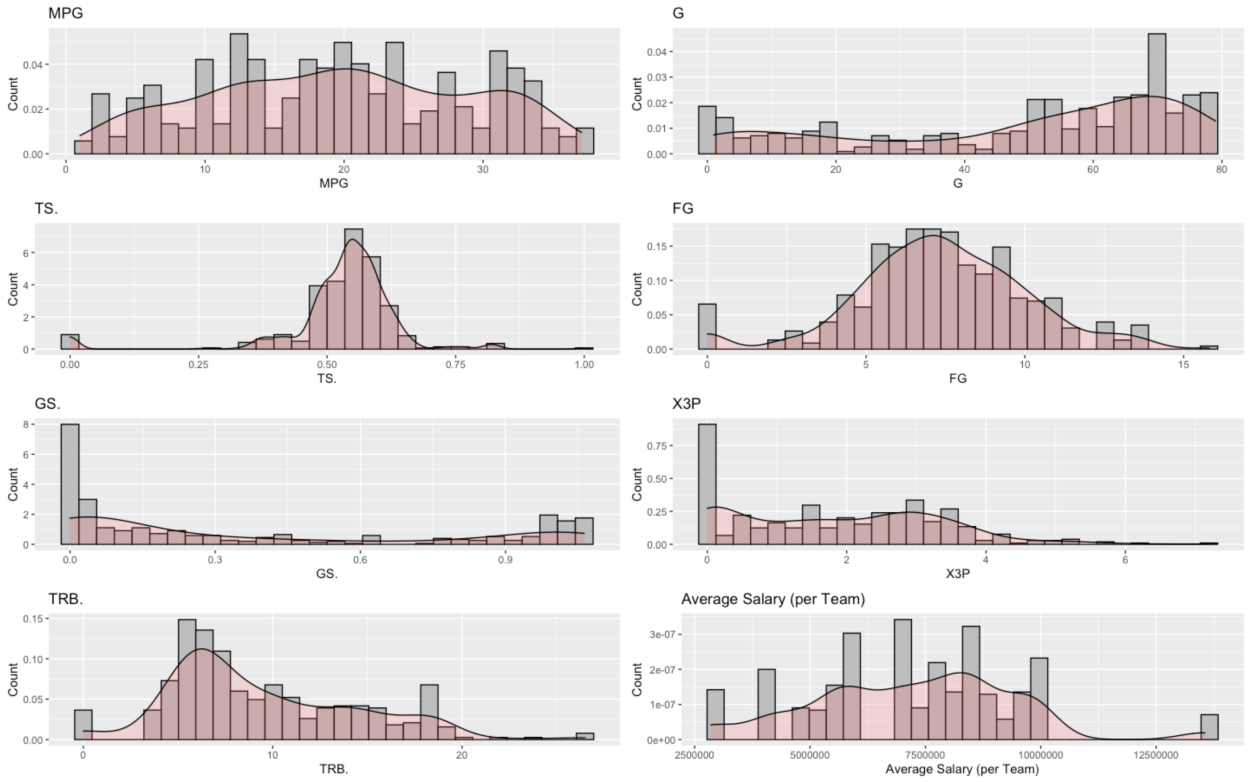
We examine the validity plots of model 3 again:



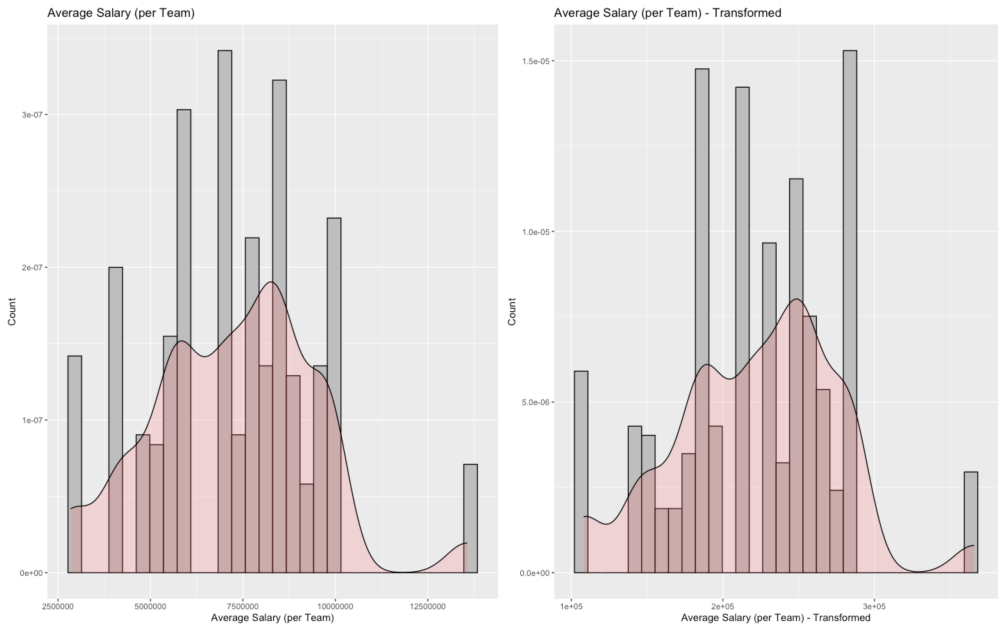
While there are no bad leverage points in the model, we see several violations of the assumptions for multiple linear regression. In particular, we see violations in the random residual assumption, the constant variance assumption, and the normality assumption in the diagnostics. To ameliorate this, we perform a Box-Cox transformation on the reponse variable along with its numeric predictors:

Variable	Est Power	Rounded Pwr	Wald Lwr Bnd	Wald Upr Bnd
Salary	0.1557777	0.16	0.0864434	0.2251119
GS. + 1	-1.7923597	-2.00	-2.2464505	-1.3382689
MPG	0.9122409	1.00	0.7889902	1.0354915
FG + 1	0.9684239	1.00	0.8142625	1.1225853
G	0.9768199	1.00	0.8502935	1.1033462
TS. + 1	3.6317678	3.63	3.0738145	4.1897210
X3P + 1	0.2722915	0.33	0.0935937	0.4509893
TRB. + 1	0.7258341	0.73	0.5956181	0.8560501
AvgTMSalary	0.7823881	1.00	0.5467028	1.0180735

We examine the densities of the numeric predictors before we transform:

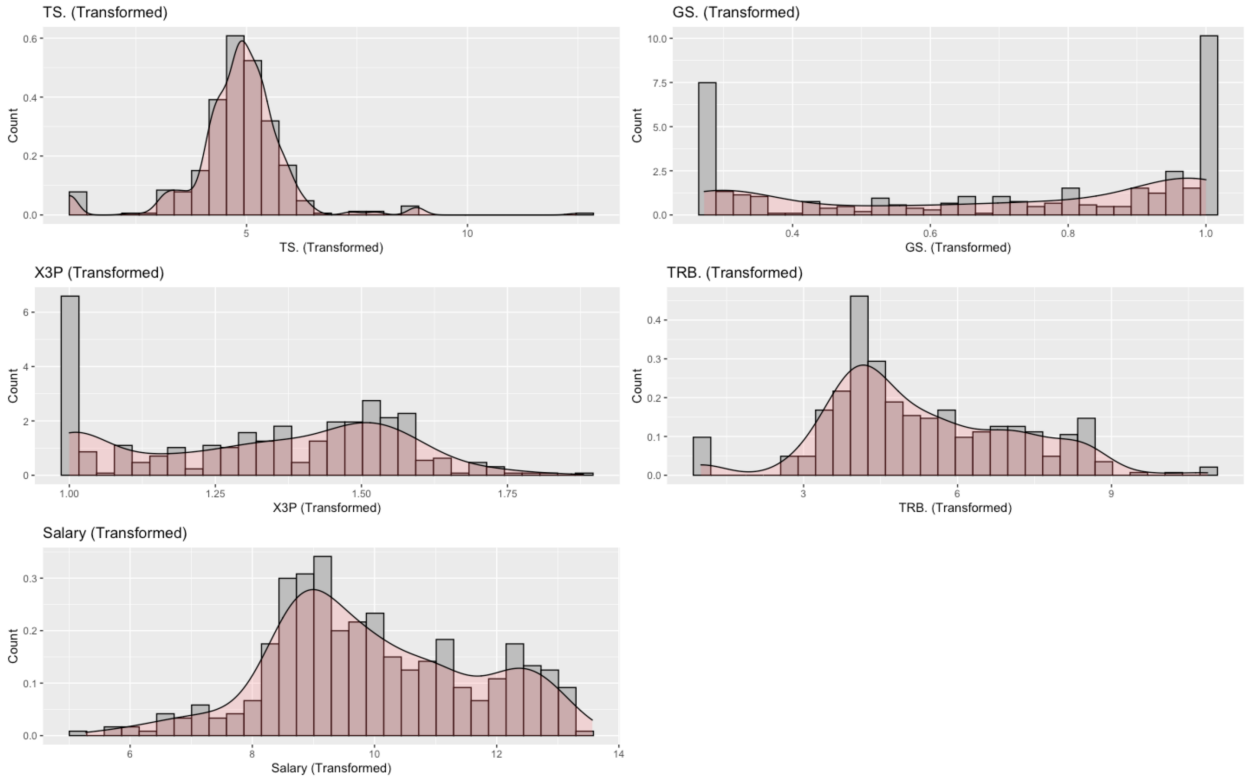


Note: the distribution of AvgTMSalary does not seem to change much with $\lambda = 0.78$ so we will not transform the variable:

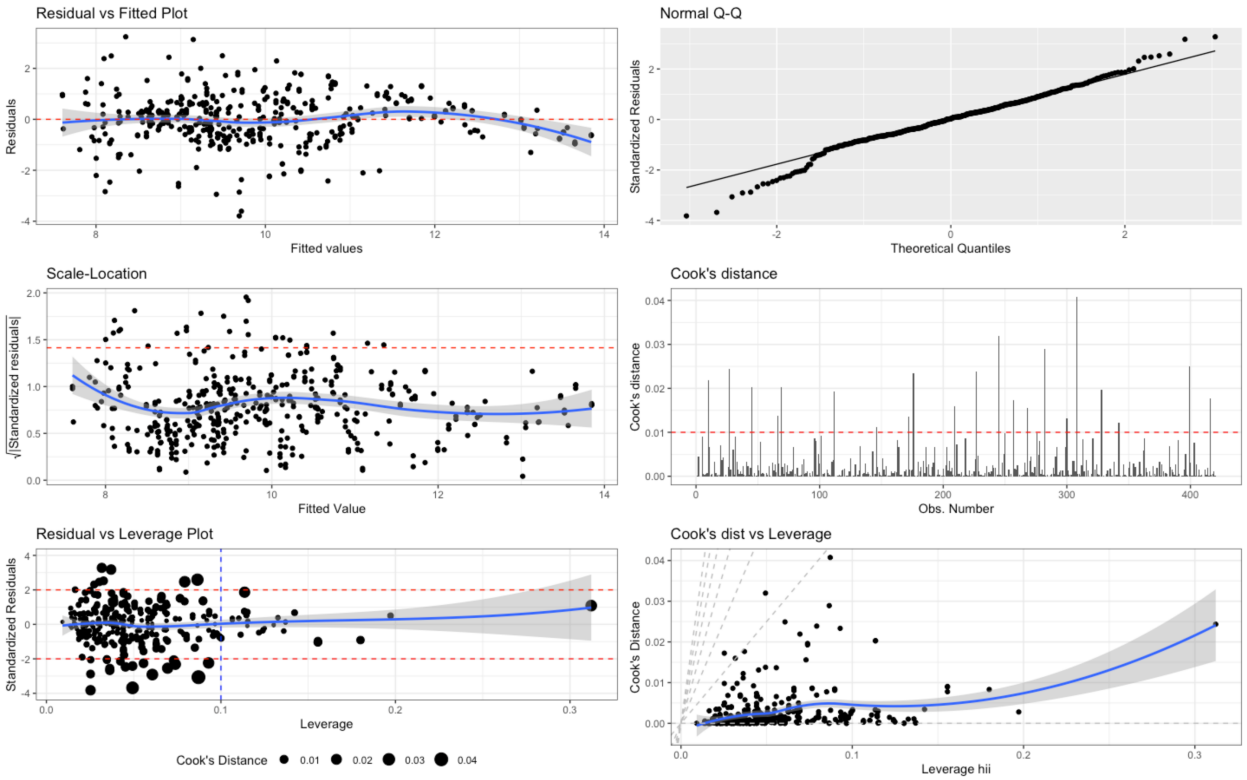


We then perform a Box-Cox transformation on the response variable and its predictors with the following λs:

Variable	Salary	GS.	MPG	FG	G	TS.	X3P	TRB.	AvgTMSalary
λ	0.15	-1.8	1	1	1	3.64	0.3	0.72	1



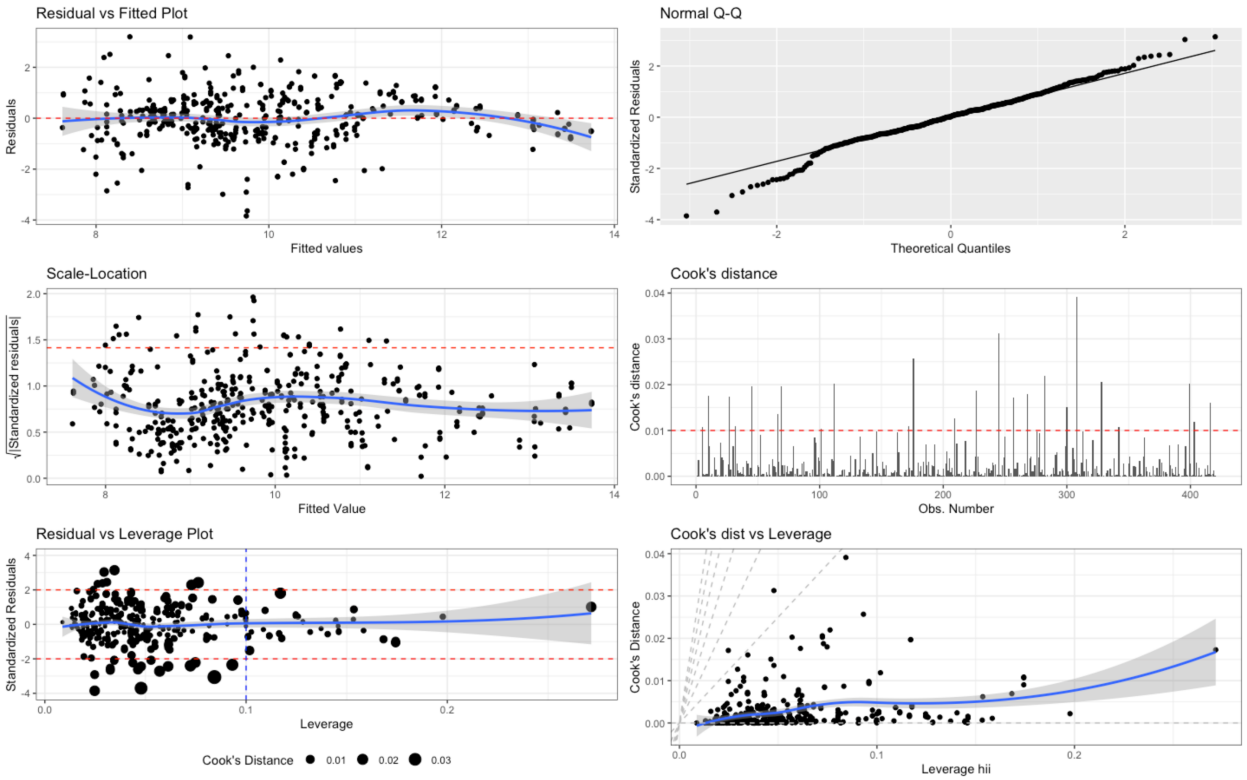
Finally, we fit a model around the transformed predictors and response variable, and examine the diagnostics:



Unfortunately, we see a funnel pattern in the Residuals vs Fitted plot and a bit of a curve in the Scale-Location plot which indicates nonconstant variance from the residuals. We try to ameliorate this problem by using Weighted Least Squares.

Weighted Least Squares

The diagnostics and performance of the Weighted Least Squares model is shown below:



Observations	Training R^2	Training Adjusted R^2	Test R^2 (Kaggle)
420	0.6822	0.6654	0.55498

Note: the code for performing weighted least squares can be found on the website for Penn State's Department of Statistics (Penn State).

After transformations and weighted least squares, we still see violations in the regression assumptions. Furthermore, the transformed model with WLS performs much worse than the original model. Thus, it does not appear that we can create a valid model from model 3.

Final Adjustment

Outliers

Over the course of the project, we experimented with over 100 different models before reaching the finalized product that is shown in this report. Moreover, as we experimented with different ideas and variations, we also kept track of the bad leverage points that consistently appeared and improved our models when removed. A total of eight observations were collected and are shown below:

Obs	Age	G	MPG	FG	FGA	AST	PTS	TRB	STL	BLK	TOV	PF	Salary
63	29	11	30.36364	7.0	17.3	4.7	19.5	6.6	2.8	0.7	3.6	4.9	5815206
154	39	77	24.67532	9.1	20.0	3.2	24.5	11.6	1.1	1.2	1.3	3.9	5185818
163	30	12	31.08333	8.7	22.8	6.6	27.8	3.7	1.6	0.4	2.4	3.3	28750441
311	26	79	21.82278	6.7	14.3	2.2	17.7	5.6	0.5	0.5	1.6	3.4	3549110
343	22	55	33.41818	9.6	22.2	3.5	26.2	11.7	1.5	1.2	2.7	2.8	5972652

Obs	Age	G	MPG	FG	FGA	AST	PTS	TRB	STL	BLK	TOV	PF	Salary
176	29	3	25.33333	13.4	26.7	6.5	36.9	5.5	2.9	0.8	4.2	1.6	1115910
268	23	71	27.53521	10.7	16.4	1.6	24.8	19.4	1.4	3.3	2.5	4.5	2917360
324	29	1	25.00000	9.7	23.3	7.8	35.0	0.0	0.0	0.0	5.8	5.8	12298793

Upon inspection, we see that these observations are all quite unusual and should be removed from the model. In particular, the most blatant of them of all is observation 324 (let's call him Bill). Bill here played 25 minutes in 1 game, shot 23 times in those 25 minutes, but still had the generosity (and the time) to dish out 7.8 assists. It would be quite the understatement to say that Bill was doing the impossible.

Thus, for one last adjustment, we remove the observations above from model 3, and we see a slight improvement from model 3 in terms of prediction.

Results

Final Model

Our final model is given by:

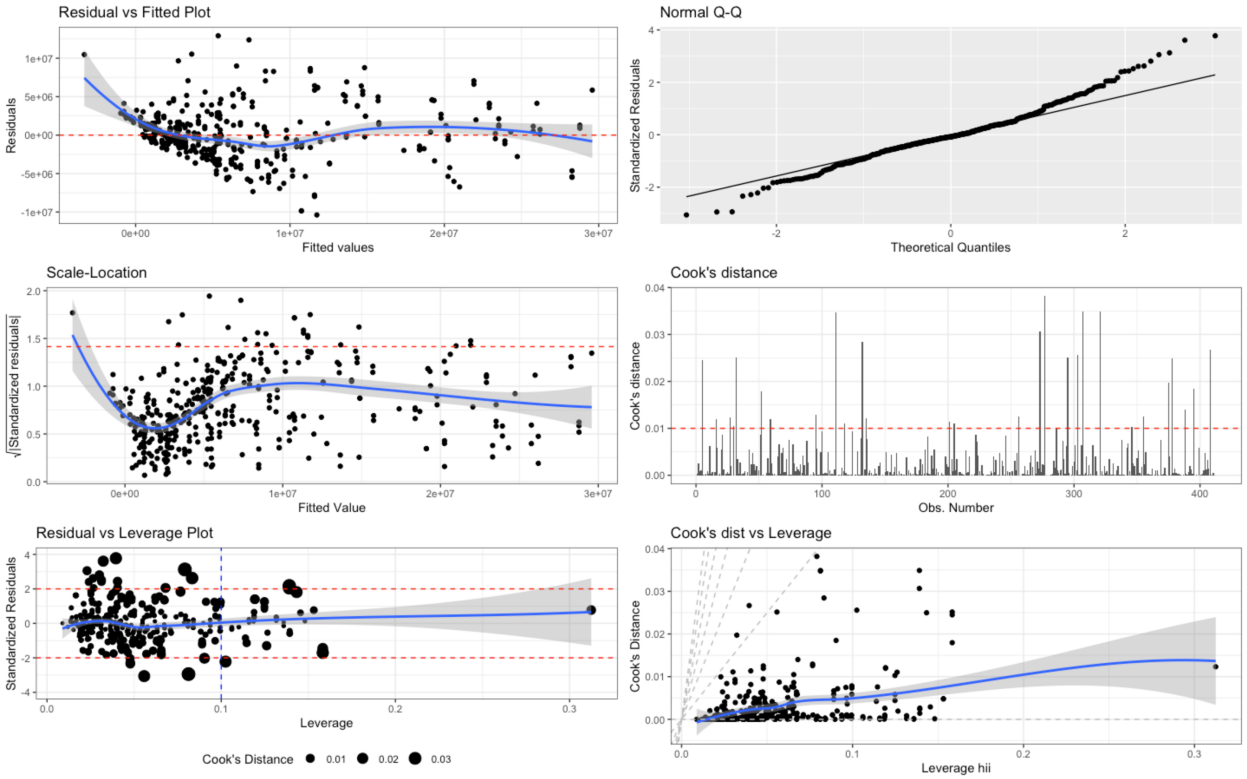
$$\hat{Salary} = 306953.9 + 4596537(GS.) - 1721041(MPG) + 869643.1(FG) + 388481.6(G) - 3279435(TS.) - 428147.8(X3P)$$
$$- 154671.2(TRB.) + 69663.95(Age)(MPG) - 15941.98(Age)(G) + 23953.95(MPG)(FT) + 3107.404(MPG)(GS)$$
$$- 31734.86(PTS)(X2P) - 940794.7(FG)(TradedYes) + 1300400(X2P)(TradedYes) - 65262.32(GS)(AST.TOV)$$
$$- 1137243(T.ConfW)(VORP) + 39564.5(GS)(tT.DivP) + 1939565(TradedYes)(OWS) + 15147610(VORP)(WS.48)$$
$$+ 206262(AST.)(tPosFrontcourt) + 0.3268309(AvgTMSalary)$$

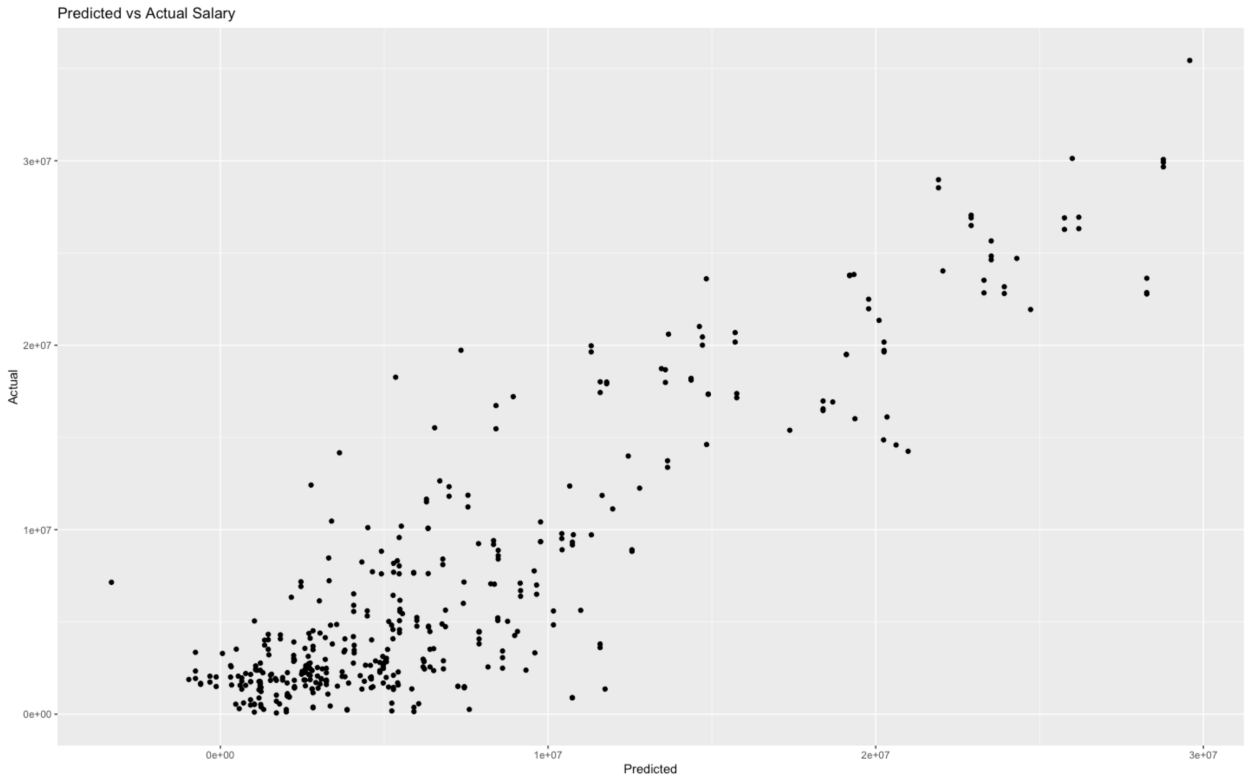
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.069539e+05	1.185517e+06	0.2589199	0.7958337
GS.	4.596537e+06	1.381653e+06	3.3268399	0.0009618
MPG	-1.721041e+06	1.943089e+05	-8.8572414	0.0000000
FG	8.696431e+05	2.614831e+05	3.3258105	0.0009652
G	3.884816e+05	6.684845e+04	5.8113779	0.0000000
TS.	-3.279435e+06	1.668150e+06	-1.9659109	0.0500179
X3P	-4.281478e+05	2.304427e+05	-1.8579357	0.0639315
TRB.	-1.546712e+05	5.704274e+04	-2.7114972	0.0069942
Age:MPG	6.966395e+04	7.149129e+03	9.7443970	0.0000000
Age:G	-1.594198e+04	2.527671e+03	-6.3069847	0.0000000
MPG:FT	2.395395e+04	6.567028e+03	3.6476086	0.0003006
MPG:GS	3.107404e+03	6.950859e+02	4.4705318	0.0000102
PTS:X2P	-3.173486e+04	8.493075e+03	-3.7365575	0.0002144
FG:TradedYes	-9.407947e+05	2.647528e+05	-3.5534836	0.0004268
X2P:TradedYes	1.300400e+06	3.158830e+05	4.1167153	0.0000469
GS:AST.TOV	-6.526232e+04	8.049727e+03	-8.1073961	0.0000000
T.ConfW:VORP	-1.137243e+06	3.053196e+05	-3.7247613	0.0002243
GS:tT.DivP	3.956450e+04	1.380566e+04	2.8658176	0.0043846
TradedYes:OWS	1.939565e+06	3.200150e+05	6.0608567	0.0000000
VORP:WS.48	1.514761e+07	1.646719e+06	9.1986615	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
AST.:tPosFrontcourt	2.062620e+05	3.990750e+04	5.1685022	0.0000004
AvgTMSalary	3.268309e-01	8.664110e-02	3.7722385	0.0001869

Observations	Training R^2	Training Adjusted R^2	Test R^2 (Kaggle)
412	0.7987	0.7878	0.63156

Final Diagnostics





Note: these are predicted and actual values from the training data

Summary statistics for fitted values:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3324692	2595396	5136129	7208919	9203021	29583855

Summary statistics for predicted values:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2226091	2879500	5392609	7112014	9809197	29583855

Discussion

Leverages and Outliers

Leverage	Outlier	
	No	Yes
No	355	22
Yes	33	2

The final model contained two bad leverage points, observations 278 and 313:

	Age	G	MPG	FG	FGA	AST	PTS	TRB	STL	BLK	TOV	PF	Salary
278	30	12	31.08333	8.7	22.8	6.6	27.8	3.7	1.6	0.4	2.4	3.3	28533996
313	30	12	31.08333	8.7	22.8	6.6	27.8	3.7	1.6	0.4	2.4	3.3	28974644

As discussed back in model 1, these two observations seem to be duplicates with different response values. However, since removing these two observations actually lowered our score on Kaggle, we decided to leave them in the model.

Inverse Response Plot

We tried transforming Salary with $\lambda = 1.3$ to see if we could improve our model. However, while it did improve our training R^2, it lowered our test R^2 on Kaggle. Thus we decided against using IRP in our final model.

Lambda	RSS
1.307504	3.627160e+15
-1.000000	1.777143e+16
0.000000	8.822852e+15
1.000000	3.795268e+15

ANOVA

We create an ANOVA table for our final model to identify the most important predictors in our final model. The predictors are sorted in descending sum of squares in the table below:

Variable	Df	Sum Sq	Mean Sq	F value	Pr(>F)
GS.	1	9.420769e+15	9.420769e+15	773.1623130	0.0000000
Age:MPG	1	2.347940e+15	2.347940e+15	192.6954210	0.0000000
VORP:WS.48	1	1.118919e+15	1.118919e+15	91.8296426	0.0000000
MPG	1	1.002266e+15	1.002266e+15	82.2559690	0.0000000
MPG:FT	1	9.632742e+14	9.632742e+14	79.0558928	0.0000000
FG	1	6.826374e+14	6.826374e+14	56.0240398	0.0000000
TradedYes:OWS	1	6.168714e+14	6.168714e+14	50.6266209	0.0000000
Age:G	1	5.495988e+14	5.495988e+14	45.1055588	0.0000000
GS:AST.TOV	1	4.021161e+14	4.021161e+14	33.0016557	0.0000000
AST.:tPosFrontcourt	1	3.213124e+14	3.213124e+14	26.3700976	0.0000004
FG:TradedYes	1	3.073762e+14	3.073762e+14	25.2263570	0.0000008
MPG:GS	1	2.764950e+14	2.764950e+14	22.6919418	0.0000027
TRB.	1	2.303837e+14	2.303837e+14	18.9075836	0.0000175
G	1	2.012388e+14	2.012388e+14	16.5156620	0.0000584
AvgTMSalary	1	1.733860e+14	1.733860e+14	14.2297834	0.0001869
GS:tT.DivP	1	1.369114e+14	1.369114e+14	11.2363191	0.0008805
PTS:X2P	1	3.315018e+13	3.315018e+13	2.7206349	0.0998638
X3P	1	3.310643e+13	3.310643e+13	2.7170438	0.1000870
T.ConfW:VORP	1	1.349668e+13	1.349668e+13	1.1076723	0.2932401
X2P:TradedYes	1	1.309230e+13	1.309230e+13	1.0744851	0.3005768
TS.	1	5.744352e+12	5.744352e+12	0.4714389	0.4927340

From the ANOVA table, we see that the numeric predictor GS. and the interaction Age:MPG were the most significant predictors in our final model by a large margin.

Limitations

Validity of the Model

The most glaring problem with this project has to be the fact that we were unable to produce a valid linear regression model. Specifically, even after transformations and weighted least squares, we still experienced violations in the various assumptions of a multiple linear regression model. The invalidity of the model would make it extremely difficult to make statistical inference on the model.

Linearity

From the various MMPS plots above, it was quite clear that the numeric predictors in the dataset did not have strong linear relationships with the response variable Salary. This limitation made it quite difficult to produce a model that could accurately predict the salary of a given NBA player.

Quality of Data

Another glaring issue with the project was the quality of data that was provided. For starters, as seen in many of the bad leverage points produced, there were several cases of observations with identical predictor values but differing salary values in the dataset. In fact, if we exclude Salary and only included basic statistics like PTS, TRB, AST, STL, etc., only 255 observations out of the original 420 in the training data would be unique.

The issues do not stop there. The NBA consists of 30 teams, but both the training and testing dataset only contains 26. That is incredibly strange considering that there are 600 total observations (the number of active players in a season usually numbers around 400-500). Furthermore, if a player commits more than six fouls in an NBA game, that player is then ejected and no longer allowed to continue playing in that game. However, there are over 53 observations in the dataset that average over 6 fouls a game which is literally impossible. Lastly, we came across a large number of strange observations such as Bill (observation 324) from the **Final Adjustments** section that did things on the court that were basically impossible for a human being.

Below are examples of players that averaged over *ten* fouls a game:

Obs	Age	G	MPG	FG	FGA	AST	PTS	TRB	STL	BLK	TOV	PF	Salary
36	22	28	12.60714	7.7	17.3	1.9	19.2	17.3	0.0	3.8	5.8	13.5	5324699
39	20	1	4.00000	12.5	25.0	0.0	25.0	0.0	12.5	0.0	0.0	12.5	2429826
107	20	1	4.00000	12.5	25.0	0.0	25.0	0.0	12.5	0.0	0.0	12.5	2614357
174	22	28	12.60714	7.7	17.3	1.9	19.2	17.3	0.0	3.8	5.8	13.5	5599522
208	20	1	4.00000	12.5	25.0	0.0	25.0	0.0	12.5	0.0	0.0	12.5	2403689
227	30	2	3.00000	0.0	8.4	0.0	0.0	0.0	0.0	0.0	0.0	16.7	6334706

Conclusion

Although we did not have strong linearity between the predictors and the response, it was still possible to produce accurate models. This was evidenced by the fact that some of the best models in the Kaggle competition produced R^2 scores above 0.75 with the test set. Furthermore, this also demonstrated the fact that there were probably still plenty of ways to improve our model given the fact that our R^2 was much lower than 0.75 with the test set.

Moreover, since we did not have a strong linear relationship between the predictors and the response, perhaps there are other models out there that are more appropriate for this dataset than a multiple linear regression model.

Finally, with problems involving model validity and data quality, it is quite difficult to make solid conclusions with the model that was produced from this project. In particular, with an invalid model, we are unable to make statistical inferences on the model and with data quality issues, it became hard to trust that the data represented actual NBA players. Thus, even if we were to find a strong relationship between player performance and their salaries with our regression model, our findings woud not be valid because our model was not valid and the data quality was not ideal.

All that being said, despite certain shortcomings, I am proud of the model that I was able to produce in this project. Out of 52 students, I was able to place 7^{th} with an $R^2 = 0.63156$ on Kaggle. Within the given time frame, I believe that I gave every effort that I could to make my model the best that I can, and I truly had fun doing so.

References

Almohalwas, Akram. 2020. *Chapter 5 Updated Winter 2020.*

Almohalwas, Akram. 2021. *STAT 101 A Spring 2021 Kaggle Competition: Predicting NBA Players' Salary.*

Basketball Reference. "Glossary." Accessed June 13, 2021. https://www.basketball-reference.com/about/glossary.html

ESPN. Accessed June 13, 2021. http://www.espn.com/nba/salaries.

Kaggle. 2021. "NBATrain.csv." https://www.kaggle.com/c/nba-players-salaries/data.

PennState: Statistics Online Courses. "R Help 13: Weighted Least Squares: STAT 501." Accessed June 13, 2021. https://online.stat.psu.edu/stat501/lesson/r-help-13-weighted-least-squares.

Sheather, Simon J. 2009. *A Modern Approach to Regression with R.* New York: Springer.