USF MSDS 601 Project: NBA Player of the Week

USF MSDS 601: Linear Regression Analysis

Project: NBA Player of the Week

Team Members

- Shirley Li (@Shirleyiscool)
- Charles Siu (@chunheisiu)

Description of Dataset

Our dataset is a combination of the following datasets with regards to NBA:

- NBA Player of the Week (1985 - 2019)
  https://www.kaggle.com/jacobbaruch/nba-player-of-the-week
  1,187 Rows, 14 Columns
- NBA Player Salary from basketball-reference.com (1991 - 2017)
  https://www.kaggle.com/whitefero/nba-player-salary-19902017
  11,837 Rows, 7 Columns
- NBA Player Salary from basketball-reference.com (2018 - 2019)
  https://web.archive.org/web/20181002194236/www.basketball-reference.com/contracts/players.html
  578 Rows, 11 Columns
- NBA Player Statistics (1985 - 2019)
  https://www.basketball-reference.com/leagues/
  18,480 Rows, 30 Columns
- NBA Yearly Summary (1985 - 2019)
  https://www.basketball-reference.com/leagues/
  35 Rows, 8 Columns

Combining the aforementioned datasets, we created a dataset in which, each row is an NBA player per season, and each column is a statistic of the player. We filtered the rows so that only the players who have both statistics and salary data for that particular season are included.

There are 9,003 Rows and 38 Columns in the dataset.

Index of the Dataset

Variable	Definition	Type
`Year`	(e.g. 1991 means the NBA 1990 - 1991 Season)	Numerical
`Player`	Player name	Categorical

Variables of the Dataset

Variable	Definition	Type
`Pos`	Player Position	Categorical
`Age`	Age of Player at the start of February 1st of that season	Numerical
`Tm`	Team of Player	Categorical
`G`	Number of games played	Numerical
`GS`	Number of games played when the game started	Numerical
`MP`	Minutes played per game	Numerical
`FG`	Field Goals per game	Numerical

Variable	Definition	Type
FGA	Field Goal attempts per game	Numerical
FG_Prct	Field Goal percentage	Numerical
Three_P	3-Point Field Goals per game	Numerical
Three_PA	3-Point Field Goal attempts per game	Numerical
Three_P_Prct	3-Point Field Goal percentage	Numerical
Two_P	2-Point Field Goals per game	Numerical
Two_PA	2-Point Field Goal attempts per game	Numerical
Two_P_Prct	2-Point Field Goal percentage	Numerical
ePF_Prct	Effective Field Goal percentage	Numerical
FT	Free Throws per game	Numerical
FTA	Free Throw attempts per game	Numerical
FTA_Prct	Free Throw percentage	Numerical
ORB	Offensive Rebounds per game	Numerical
DRB	Defensive Rebounds per game	Numerical
TRB	Total Rebounds per game	Numerical
AST	Assists per game	Numerical
STL	Steals per game	Numerical
BLK	Blocks per game	Numerical
TOV	Turnovers per game	Numerical
PF	Personal Fouls per game	Numerical
PTS	Points per game	Numerical
Potw	Was the player named *Player of the Week* during the season?	Binary
APG_Leader	Was the player named *Assists Per Game Leader* during the season?	Binary
MVP	Was the player named *Most Valuable Player* during the season?	Binary
PPG_Leader	Was the player named *Points Per Game Leader* during the season?	Binary
RPG_Leader	Was the player named *Rebounds Per Game Leader* during the season?	Binary
Rookie	Was the player named *Rookie of the Year* during the season?	Binary
WS Leader	Was the player named *Win Shares Leader* during the season?	Binary
Salary	Player Salary	Numerical

Statement of Research Problems and Methods

Using the dataset, we stemmed two main research problems:

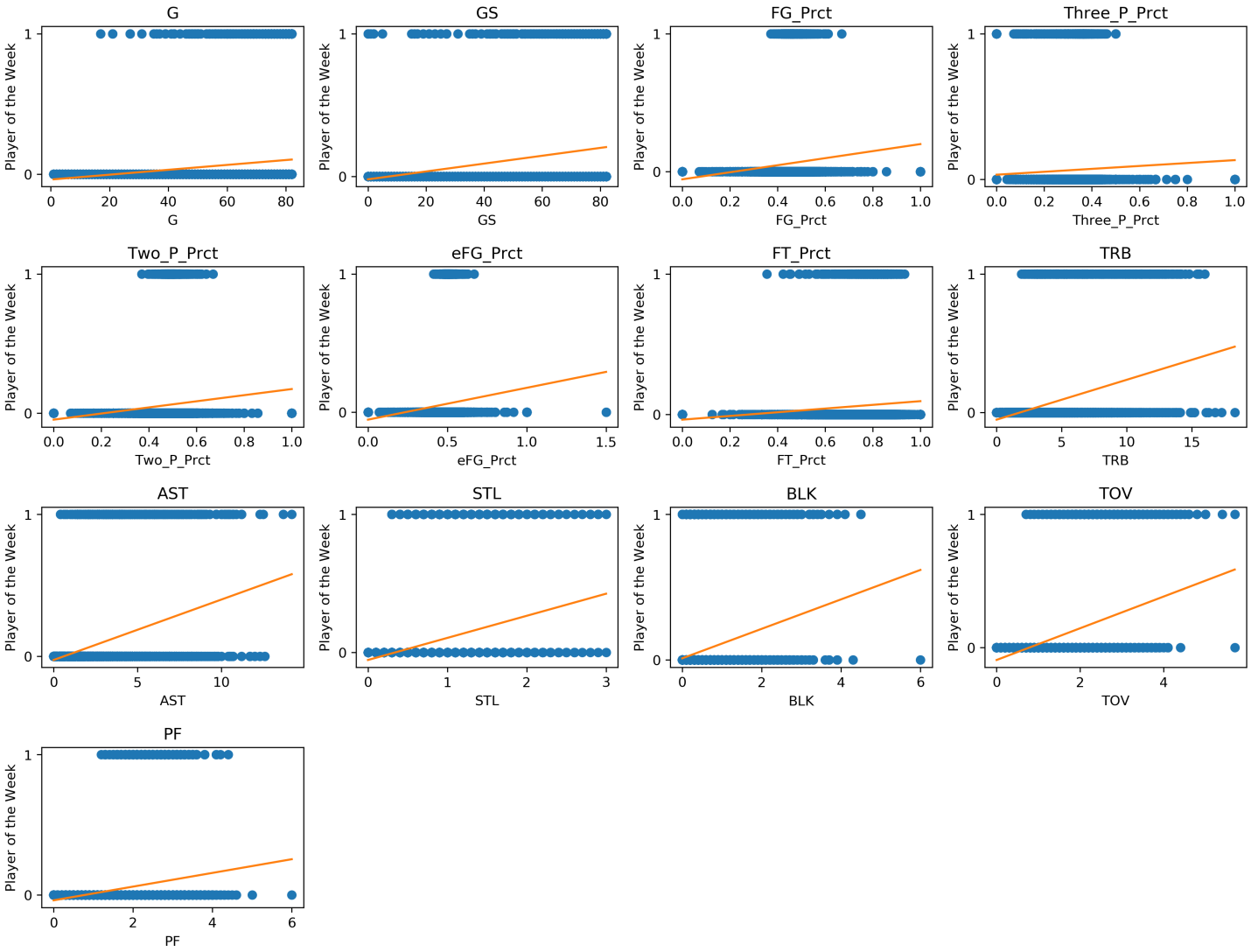
- **What player statistic contributes the most to the event that the player is named Player of the Week?**
  Since whether a player is named Player of the Week is a binary variable, we decided to approach this problem using the logistic regression model.
- **What NBA title, including Player of the Week, has the most weight on the salary of the player?**
  Since the salary of a player is a numerical variable, we decided to approach this problem using the multiple linear regression model.

For both problems, model selection was performed to find the optimal model, and model diagnosis was performed to mitigate the possible issues of heteroscedasticity, multicollinearity and autocorrelation.

Problem 1: Relationship between Player Statistics and Player of the Week

Explanatory Analysis

After extracting the relevant player statistics and Player of the Week from the dataset, we plotted the relationship between the statistics and Player of the Week using a scatter plot.



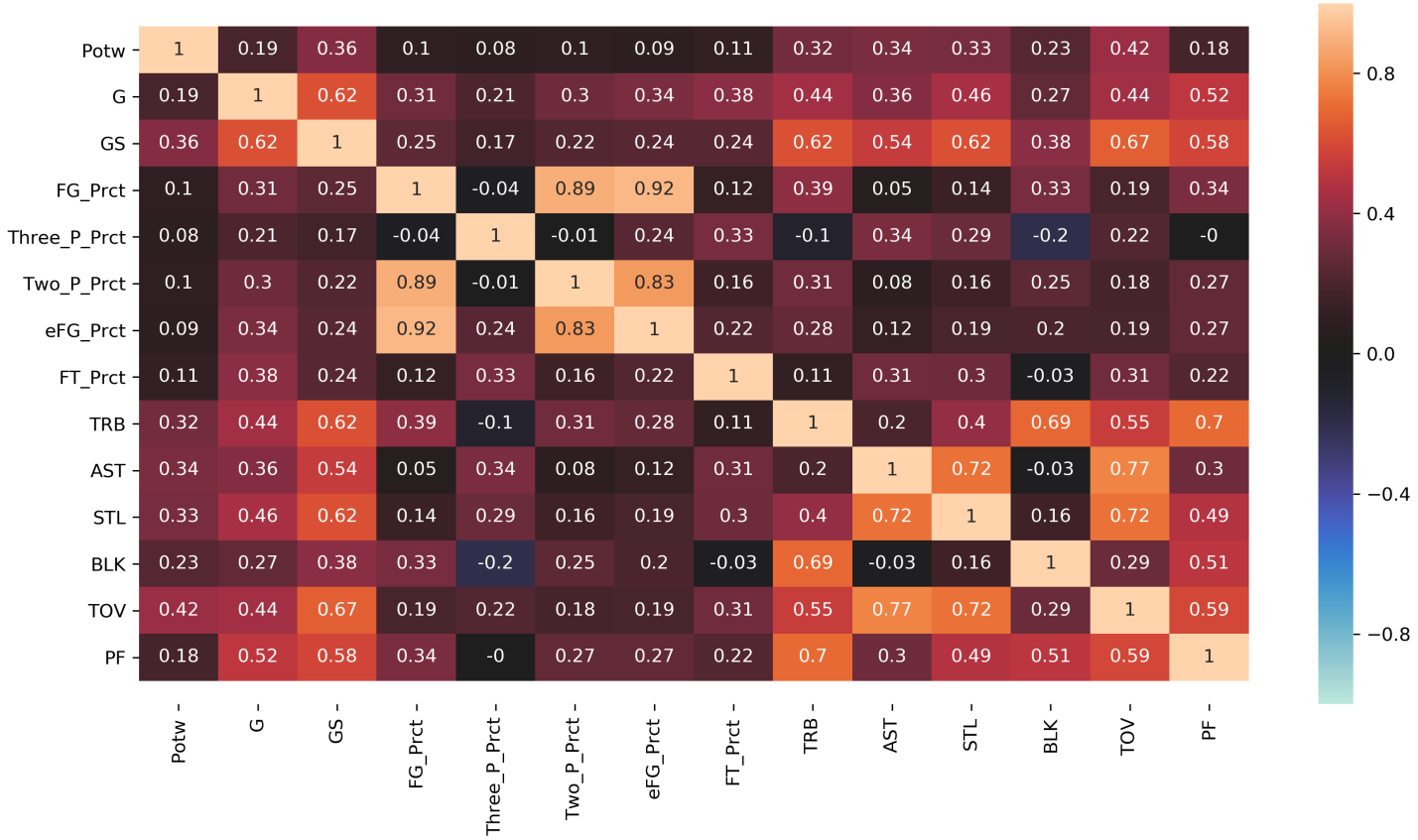
Observing the scatter plot, since `Potw` is a binary variable, the scatter plot did not give us a lot of useful information, apart from the differences in range of statistic values between the `Potw = 0` and `Potw = 1`. For every statistic, the range of values seems to be smaller for `Potw = 1`, with the most significant variable being `eFG_Prct`.

This discrepancy in range is also evident in the difference in frequency between `Potw = 0` and `Potw = 1`.

Potw	Count	Prct
0	8505	0.944685
1	498	0.0553149

The frequency table shows that `Potw = 0` accounts for 94% of the data, which is to be expected since the number of players receiving an award would always be significantly smaller than those who did not. However, we are not sure if this would effect the reliability of the models we would build in regression analysis.

We also plotted the correlation using a heatmap.



Observing the heatmap, there are evidence that multicollinearity might exist. For example, The most correlated variables are `Two_P_Prct` and `FG_Prct`, but this is to be expected since `FG_Prct` is derived from `Two_P_Prct`. Similarly, `eFG_Prct` is derived from `FG_Prct`, so the correlation is high between them. Hence, some of these variables, specifically those that have direct relationships, will need to be removed prior to regression analysis.

Meanwhile, `TOV`, `AST` and `STL` are highly correlated between one another. However, turnovers, assists and steals are basketball moves often performed by point guards, so there might be indirect relationships between these variables. Nonetheless, these correlations would need to be addressed in regression analysis.

Regression Analysis

Model Selection

- As `Pos` is categorical variables, we first get dummies for this predictors.
- Since some statistics are calculated by other statistics, there would be strong multicollinearity if we include all of them. Therefore, we drop these following statistics for our first model.

  `TRB = ORB + DRB`
  `FGA = FG * FG_Prct`
  `Three_PA = Three_P * Three_P_Prct`
  `Two_PA = Two_P * Two_P_Prct`
  `FTA = FT_P * FT_Prct`
  `PTS = Three_P + Two_P + FT_P`
  `FG = Three_P + Two_P`
- Then we fit the full model using all the remaining players' statistics, such as `Age`,`G`,`GS`,`MP`and etc.. Hence, we got the following logistic regression model as follows.

Full Model Summary - Model 1

Generalized Linear Model Regression Results

Dep. Variable:	Potw	No. Observations:	9003
Model:	GLM	Df Residuals:	8979
Model Family:	Binomial	Df Model:	23
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-804.36
Date:	Sun, 13 Oct 2019	Deviance:	1608.7
Time:	19:02:53	Pearson chi2:	5.37e+03
No. Iterations:	9		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-11.5546	1.662	-6.951	0.000	-14.813	-8.296
Age	0.0016	0.019	0.086	0.931	-0.035	0.039
G	0.0219	0.008	2.813	0.005	0.007	0.037
GS	0.0117	0.006	1.820	0.069	-0.001	0.024
MP	-0.0684	0.028	-2.405	0.016	-0.124	-0.013
FG_Prct	7.0209	9.942	0.706	0.480	-12.466	26.508
Three_P	1.0781	0.256	4.210	0.000	0.576	1.580
Three_P_Prct	-0.6691	0.666	-1.004	0.315	-1.975	0.637
Two_P	0.6287	0.074	8.493	0.000	0.484	0.774
Two_P_Prct	-4.7789	3.358	-1.423	0.155	-11.360	1.802
eFG_Prct	0.4554	7.372	0.062	0.951	-13.994	14.905
FT	0.4298	0.066	6.494	0.000	0.300	0.559
FT_Prct	0.7733	1.089	0.710	0.478	-1.362	2.908
ORB	-0.0114	0.132	-0.086	0.931	-0.270	0.248
DRB	0.3683	0.067	5.535	0.000	0.238	0.499
AST	0.2276	0.065	3.527	0.000	0.101	0.354
STL	0.4689	0.192	2.448	0.014	0.094	0.844
BLK	0.5223	0.141	3.691	0.000	0.245	0.800
TOV	-0.1118	0.171	-0.653	0.514	-0.447	0.224
PF	-0.3689	0.145	-2.537	0.011	-0.654	-0.084
Pos_PF	-0.3861	0.255	-1.513	0.130	-0.886	0.114
Pos_PG	0.4955	0.459	1.079	0.281	-0.405	1.396
Pos_SF	-0.4740	0.352	-1.345	0.179	-1.165	0.217
Pos_SG	0.0563	0.402	0.140	0.889	-0.732	0.844

Given that there are too many variables with high correlation from the heatmap above as well as the there is warnning on multicollinearity, we decided to first use both VIF Factors and Deviance Test to find removable predictors.

VIF analysis on full model

Features	VIF Factor
Age	23.7475
G	11.2644
GS	6.57655
MP	79.4739
FG_Prct	870.907
Three_P	8.40503
Three_P_Prct	6.21584
Two_P	22.851
Two_P_Prct	122.755
eFG_Prct	755.823
FT	10.2741
FT_Prct	21.833
ORB	11.5458
DRB	18.2727
AST	12.1883

Features	VIF Factor
STL	10.1388
BLK	3.92813
TOV	23.6123
PF	20.7312
Pos_PF	2.29591
Pos_PG	4.82529
Pos_SF	3.03752
Pos_SG	3.895

Using a function to remove a predictor with max VIF for each VIF test while deleting that predictor would not reject H0 in deviance test and thus choose reduced model.

Hence, we remove predictors `FG_Prct`, `eFG_Prct`, `TOV`, `Age`, `MP`, `FT_Prct`, `Two_P_Prct`, `ORB`, which both have high VIF factors and the reduced model with low ΔG in a Deviance Test.

With these remaining predictors, we run a logistic model again and here is our second model.

Reduced Model Summary - Model 2

Generalized Linear Model Regression Results

Dep. Variable:	Potw	No. Observations:	9003
Model:	GLM	Df Residuals:	8987
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-811.24
Date:	Sun, 13 Oct 2019	Deviance:	1622.5
Time:	19:02:56	Pearson chi2:	6.88e+03
No. Iterations:	9		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-10.8561	0.691	-15.702	0.000	-12.211	-9.501
G	0.0265	0.008	3.492	0.000	0.012	0.041
GS	0.0060	0.006	1.036	0.300	-0.005	0.017
Three_P	0.9353	0.116	8.065	0.000	0.708	1.163
Three_P_Prct	-0.5483	0.641	-0.855	0.392	-1.805	0.708
Two_P	0.5808	0.057	10.165	0.000	0.469	0.693
FT	0.4073	0.056	7.217	0.000	0.297	0.518
DRB	0.3356	0.060	5.629	0.000	0.219	0.452
AST	0.1738	0.052	3.340	0.001	0.072	0.276
STL	0.3502	0.183	1.909	0.056	-0.009	0.710
BLK	0.4701	0.138	3.414	0.001	0.200	0.740
PF	-0.4613	0.133	-3.475	0.001	-0.722	-0.201
Pos_PF	-0.5951	0.243	-2.447	0.014	-1.072	-0.118
Pos_PG	0.2976	0.440	0.676	0.499	-0.565	1.160
Pos_SF	-0.7959	0.328	-2.425	0.015	-1.439	-0.153
Pos_SG	-0.2993	0.375	-0.797	0.425	-1.035	0.436

Features	VIF Factor
Two_P	16.2202
DRB	12.2412
PF	12.0968
STL	9.59197
G	9.12068
FT	8.76224
AST	8.3415

Features	VIF Factor
GS	5.37419
Three_P_Prct	4.76798
BLK	3.78012
Pos_PG	3.45641
Three_P	3.4551
Pos_SG	2.54538
Pos_SF	2.13738
Pos_PF	1.90174

But still there are some remaining predictors with VIF Factor larger than 10.

To make sure whether reduced model is better than the full model, we do a deviance test.

Null Hypothesis: Reduced Model
Alternative Hypothesis: Full Model

ΔG = ΔG(Reduced Model) - ΔG(Full Model) = 13.7661
$\chi 2$ = 15.5073

On significant level of 0.05, $\Delta G > \chi 2$. Therefore, we cannot reject Null Hypothesis and then choose Model 2.

But as Wald test shows that there still seems some insignificant predictors with p-values larger than 0.05. Therefore, we continue to remove predictors using Deviance Test and Wald Test. Here are removable predictors based on Deviance Test.

Deviance test	GS	Three_P_Prct	Pos_PG	Pos_SG
delta_G	14.9021	14.5091	14.2247	14.3993
chi2_stat	16.9190	16.9190	16.9190	16.9190

However, we use position as dummies variables. So, if we drop `Pos_PG` and `Pos_SG`, we need to drop other 2 other variables. In this case, dropping too many predictors, Deviance Test would tell us to stick to the full model.

Hence, we only drop variables `GS` and `Three_P_Prct` and keep `Pos` dummies.

Reduced Model Summary - Model 3

Generalized Linear Model Regression Results

Dep. Variable:	Potw	No. Observations:	9003
Model:	GLM	Df Residuals:	8989
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-812.20
Date:	Sun, 13 Oct 2019	Deviance:	1624.4
Time:	19:02:58	Pearson chi2:	6.84e+03
No. Iterations:	9		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-11.2300	0.640	-17.551	0.000	-12.484	-9.976
G	0.0321	0.005	6.327	0.000	0.022	0.042
Three_P	0.9161	0.102	9.001	0.000	0.717	1.116
Two_P	0.5946	0.055	10.773	0.000	0.486	0.703
FT	0.4073	0.057	7.205	0.000	0.297	0.518
DRB	0.3506	0.059	5.968	0.000	0.235	0.466
AST	0.1781	0.051	3.464	0.001	0.077	0.279
STL	0.3595	0.184	1.958	0.050	-0.000	0.719
BLK	0.4913	0.137	3.581	0.000	0.222	0.760
PF	-0.4453	0.132	-3.363	0.001	-0.705	-0.186
Pos_PF	-0.6254	0.243	-2.574	0.010	-1.102	-0.149
Pos_PG	0.3103	0.438	0.708	0.479	-0.549	1.170
Pos_SF	-0.8116	0.328	-2.477	0.013	-1.454	-0.169
Pos_SG	-0.3024	0.375	-0.806	0.420	-1.037	0.432

So far, here is the main logistic model we'll use.

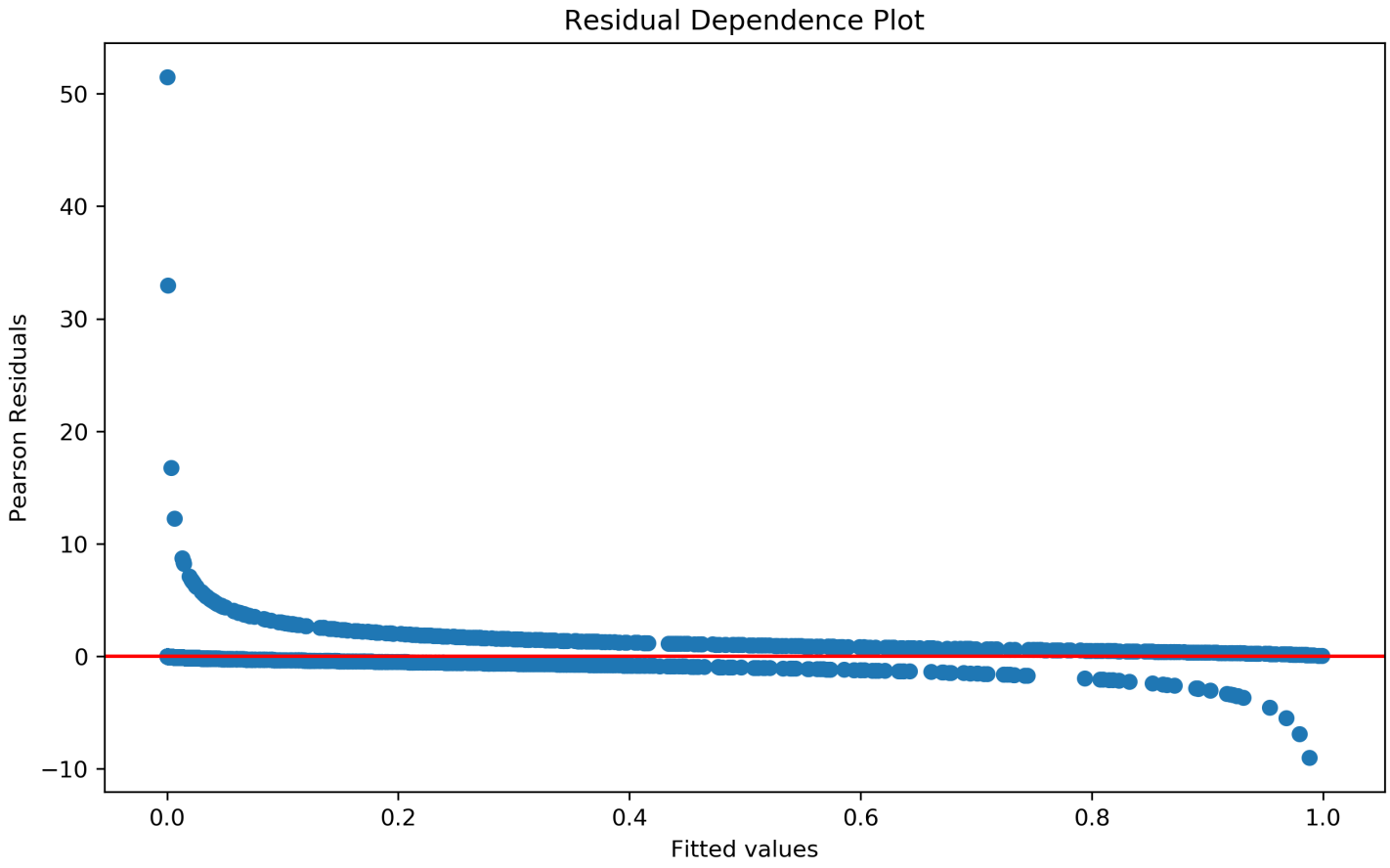
Model Diagnosis

Multicollinearity

Features	VIF Factor
Two_P	15.268
DRB	12.0013
PF	11.9826
STL	9.41759
FT	8.6996
G	8.27271
AST	8.09547
BLK	3.77999
Pos_PG	2.90391
Three_P	2.76944
Pos_SG	2.12558
Pos_SF	1.80177
Pos_PF	1.73616

The VIF table above indicates that there is multicollinearity problem in this model. But we don't choose to drop those predictors with high VIF as both Deviance test and Wald test consider them as significant. So we choose not to drop these predictors.

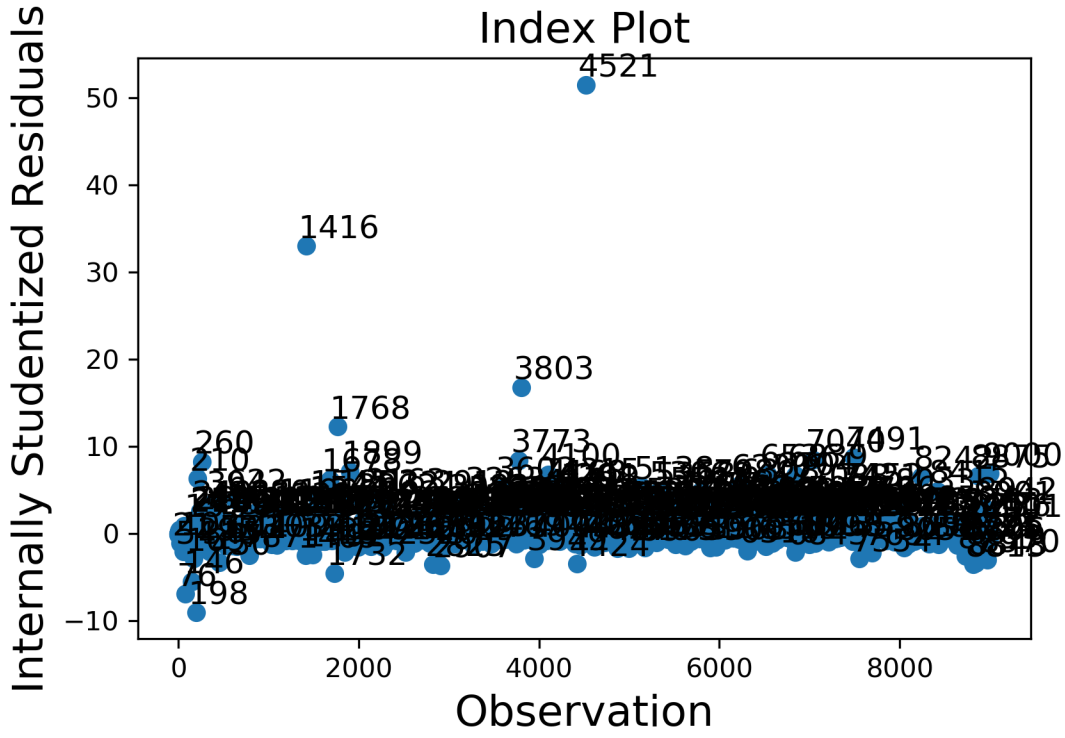
Pearson residuals Plot -- Test Heteroscedasticity



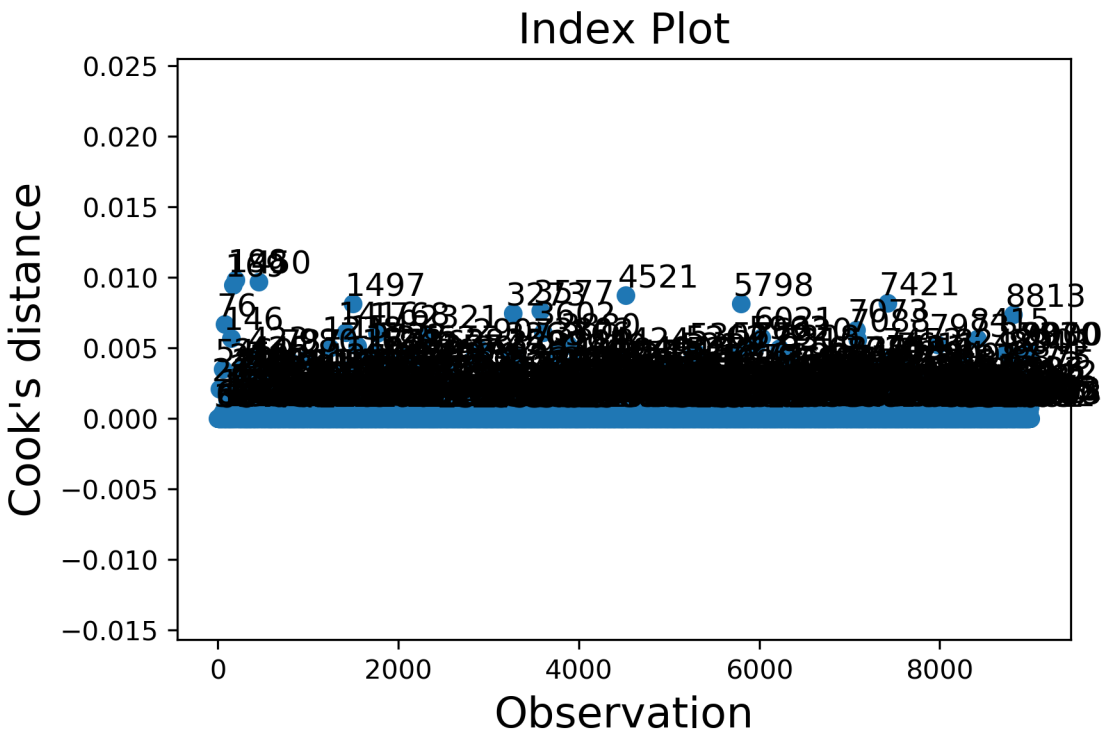
From the graph above, we can see there are some ”studentized residuals” with absolute values larger than 3, which indicates there may be outliers or influential points causing heteroscedasticity.

To find the outliers and influential points, here we plot residuals as well as cook’s distance.

Internally Studentized Residuals



Cook's Distance



Given cook's distance, Diffits and Studentized Residuals„ here we find 316 influential points. Since 316 observations take only about 5% of the total observations. Therefore, we drop these observations and rerun the model.

Final Model - Model 4

Reduced Model Summary - Model 4

Generalized Linear Model Regression Results

Dep. Variable:	Potw	No. Observations:	8687
Model:	GLM	Df Residuals:	8673
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-87.923
Date:	Sun, 13 Oct 2019	Deviance:	175.85
Time:	19:03:06	Pearson chi2:	204.
No. Iterations:	13		
Covariance Type:	nonrobust		

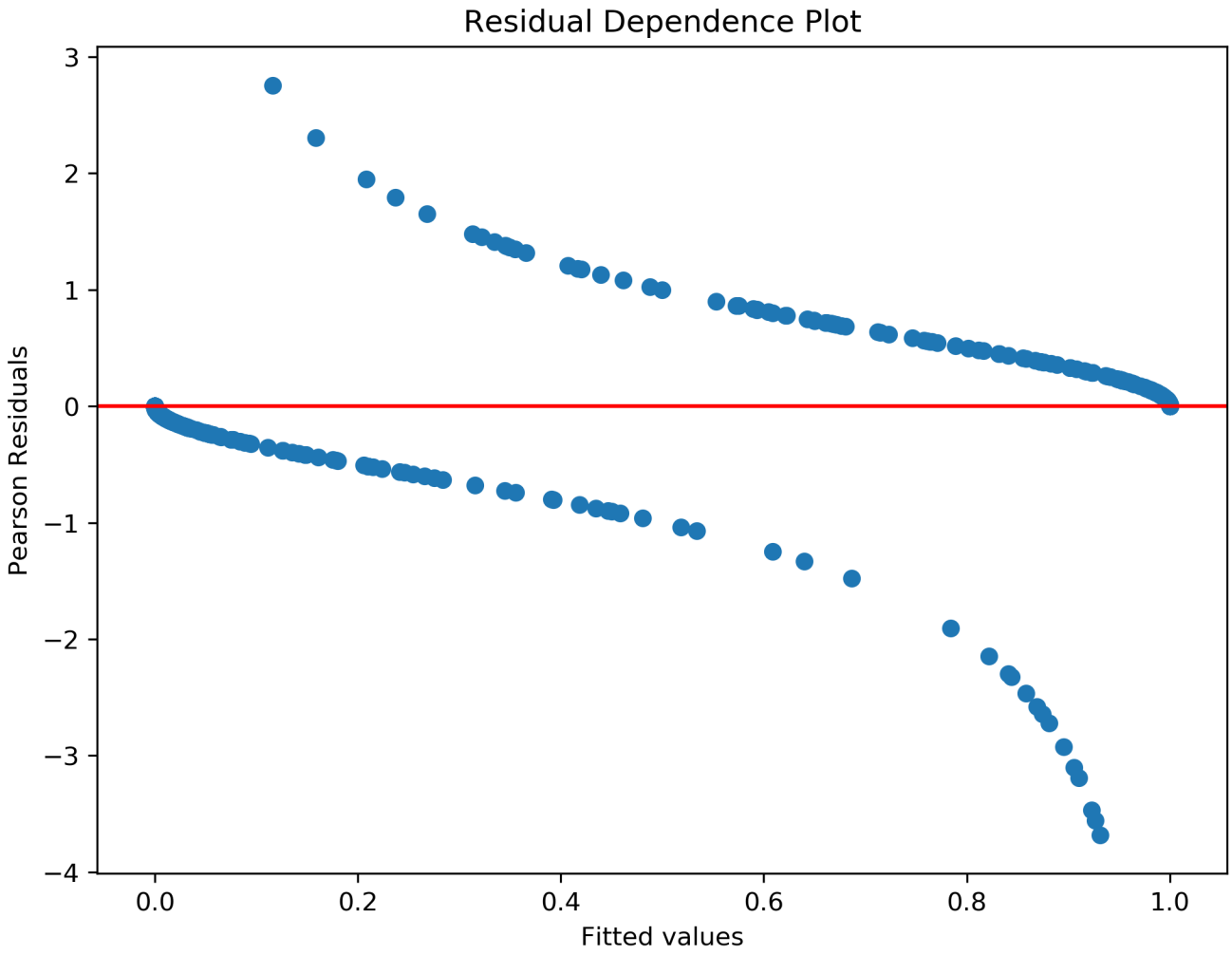
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-43.9893	5.114	-8.602	0.000	-54.013	-33.966
G	0.1668	0.029	5.704	0.000	0.110	0.224
Three_P	2.4891	0.353	7.052	0.000	1.797	3.181
Two_P	1.9363	0.269	7.186	0.000	1.408	2.464
FT	1.3466	0.200	6.738	0.000	0.955	1.738
DRB	1.0637	0.224	4.751	0.000	0.625	1.503
AST	0.4244	0.159	2.675	0.007	0.113	0.735
STL	1.9159	0.590	3.247	0.001	0.759	3.073
BLK	1.3372	0.455	2.940	0.003	0.446	2.229
PF	-1.4383	0.484	-2.970	0.003	-2.387	-0.489
Pos_PF	-1.6189	0.733	-2.209	0.027	-3.056	-0.182
Pos_PG	2.3514	1.533	1.534	0.125	-0.654	5.357
Pos_SF	-2.4764	1.128	-2.195	0.028	-4.687	-0.265
Pos_SG	-0.8487	1.288	-0.659	0.510	-3.372	1.675

Here is our final model. To confirm that whether it is the best model we have run, we compare AIC and BIC of the above 4 models.

Model	AIC	BIC
Model 1	1656.71	-80147.9
Model 2	1654.48	-80207
Model 3	1652.39	-80223.3
Model 4	203.846	-78484.6

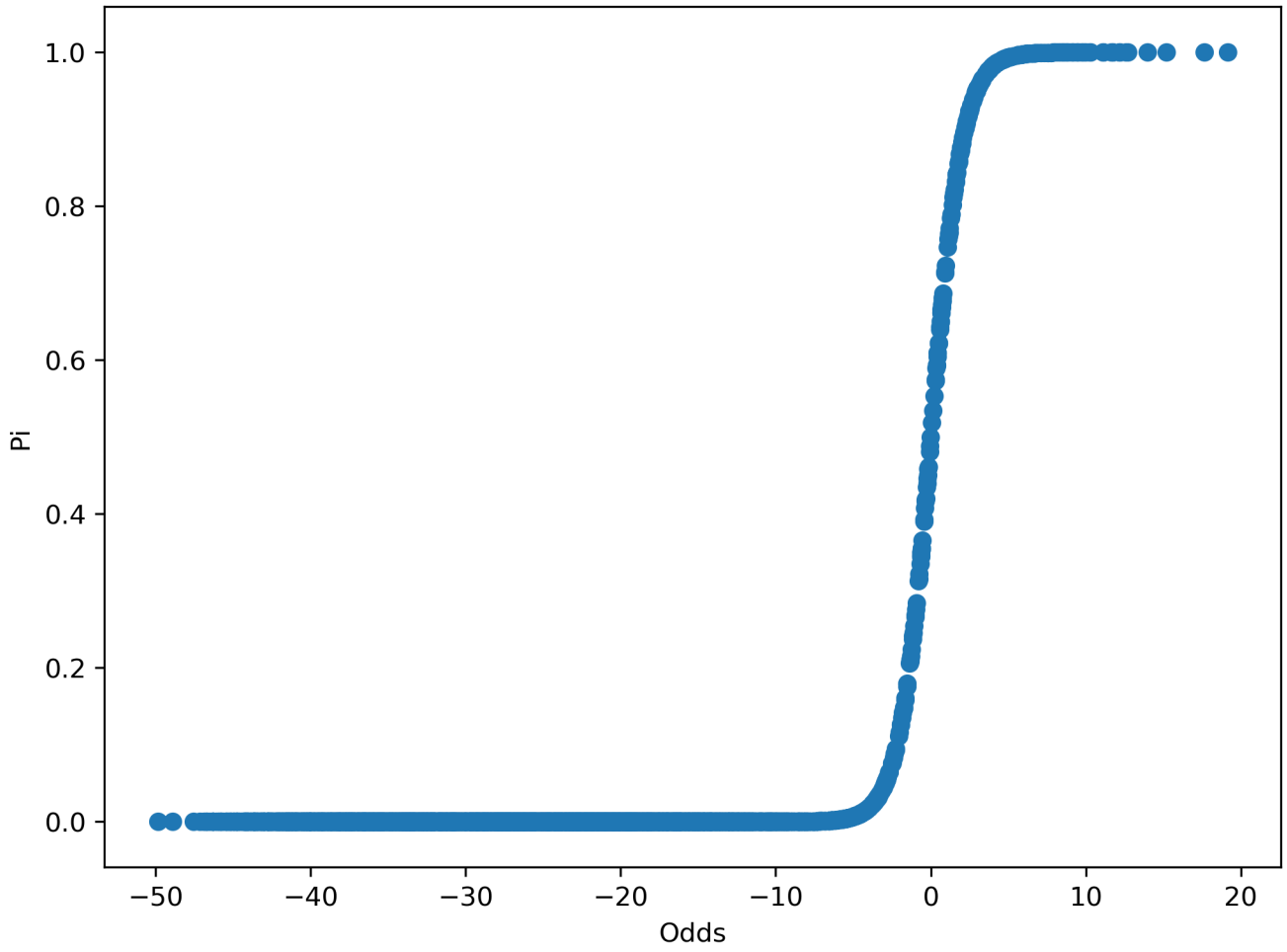
Clearly, before dropping outliers and influential points, Model 3 has the lowest AIC and BIC, showing Model 3 is better than Model 1 and Model 2. After we drop outliers and influential points, AIC of Model decreases a lot while BIC increases a little bit. So we will choose Model 4 as our final model.

Model 4 - Internally Studentized Residuals



After removing outliers, the residual plots seems better.

Model 4 - π Plot



Here we visualize how π changes with the model.

Final Model Summary

Variables

Predictors	βi	e^(βi)
Intercept	-43.98931819527114	7.86469427844486e-20
G	0.16684022971613738	1.181565471176185
Three_P	2.489105636067465	12.050493772492525
Two_P	1.9363248937885942	6.9332237580619385
FT	1.3465680005479181	3.8442095399977703
DRB	1.0637258193048331	2.89714514483543
AST	0.42443823739809583	1.52873139427614
STL	1.9159062275216838	6.793092095448223
BLK	1.3372375113410921	3.808507999802208
PF	-1.438304019319382	0.23732992451989693
Pos_PF	-1.6188977597716383	0.1981169512519482
Pos_PG	2.351391076003565	10.50016609913466
Pos_SF	-2.4763565102424177	0.08404889969899432
Pos_SG	-0.8486571509347313	0.4279892712297531

Formula

$$log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = \hat{\beta}_0 + \hat{\beta}_G \cdot X_G + \hat{\beta}_{ThreeP} \cdot X_{ThreeP} + \hat{\beta}_{TwoP} \cdot X_{TwoP} + \hat{\beta}_{FT} \cdot X_{FT} + \hat{\beta}_{DRB} \cdot X_{DRB} + \hat{\beta}_{AST} \cdot X_{AST} + \hat{\beta}_{STL} \cdot X_{STL} + \hat{\beta}_{BLK} \cdot X_{BLK} + \hat{\beta}_{PF} \cdot X_{PF} + \hat{\beta}_{PosPF} \cdot X_{PosPF} + \hat{\beta}_{PosPG} \cdot X_{PosPG} + \hat{\beta}_{PosSF} \cdot X_{PosSF} + \hat{\beta}_{PosSG} \cdot X_{PosSG}$$

$$= -43.9893 + 0.1668 \cdot X_G + 2.4891 \cdot X_{ThreeP} + 1.9363 \cdot X_{TwoP} + 1.3466 \cdot X_{FT} + 1.0637 \cdot X_{DRB} + 0.42444 \cdot X_{AST} + 1.9159 \cdot X_{STL} + 1.3372 \cdot X_{BLK} - 1.4383 \cdot X_{PF} - 1.6189 \cdot X_{PosPF} + 2.3514 \cdot X_{PosPG} - 2.4764 \cdot X_{PosSF} - 0.8487 \cdot X_{PosSG}$$

Interpretation of Model

- `Intercept`: the probability for a player win the award Player of the Week is **7.8647e-20**, which is super small.
- `G` : While controlling other variables, the odds for a player, who plays 1 more game, to win the POTW increase **18%**.
- `Three_P`: While controlling other variables, the odds for a player who can have one more 3-Point Field Goals per game, to win the POTW increase about **11** times.
- `Two_P`: While controlling other variables, the odds for a player, who can have one more 2-point field goals per game, to win the POTW increase about **6** times.
- `FT`: While controlling other variables, the odds for a player, who can have one more free throw per game, to win the POTW increase about **2.8** times.
- `DRB`: While controlling other variables, the odds for a player, who can have one more defensive rebounds per game, to win the POTW increase about **1.9** times.
- `AST`: While controlling other variables, the odds for a player, who can have one more assists per game, to win the POTW increase about **53%**.
- `STL`: While controlling other variables, the odds for a player, who can have one more steals per game, to win the POTW increase about **5.8** times.
- `BLK`: While controlling other variables, the odds for a player, who can have one more blocks per game, to win the POTW increase about **2.8** times.
- `PF`: While controlling other variables, the odds for a player, who can have one more personal fouls per game, to win the POTW decrease about **77%**.
- `Pos_PF`: While controlling other variables, the odds for a power forward is **80%** less than center.
- `Pos_PG`: While controlling other variables, the odds for a points guard is **9.5** times more than center.
- `Pos_SF`: While controlling other variables, the odds for a small forward is **92%** less than center.
- `Pos_SG`: While controlling other variables, the odds for a shooting guard is **57%** less than center.

To summarize, the model indicates that **3-Point Field Goals per game** attach the most importance to decide whether a player could get player of the week. Besides, the chance for a **point guard** to win player of the week is larger than other players. If a player wants to increase his chance of winning player of the week, increasing **2-point field goals per game**, **free throw per game**, **steals**, **assists**, **blocks** and **defensive rebounds** as well as decreasing **personal fouls** would be recommended.

Prediction of Model

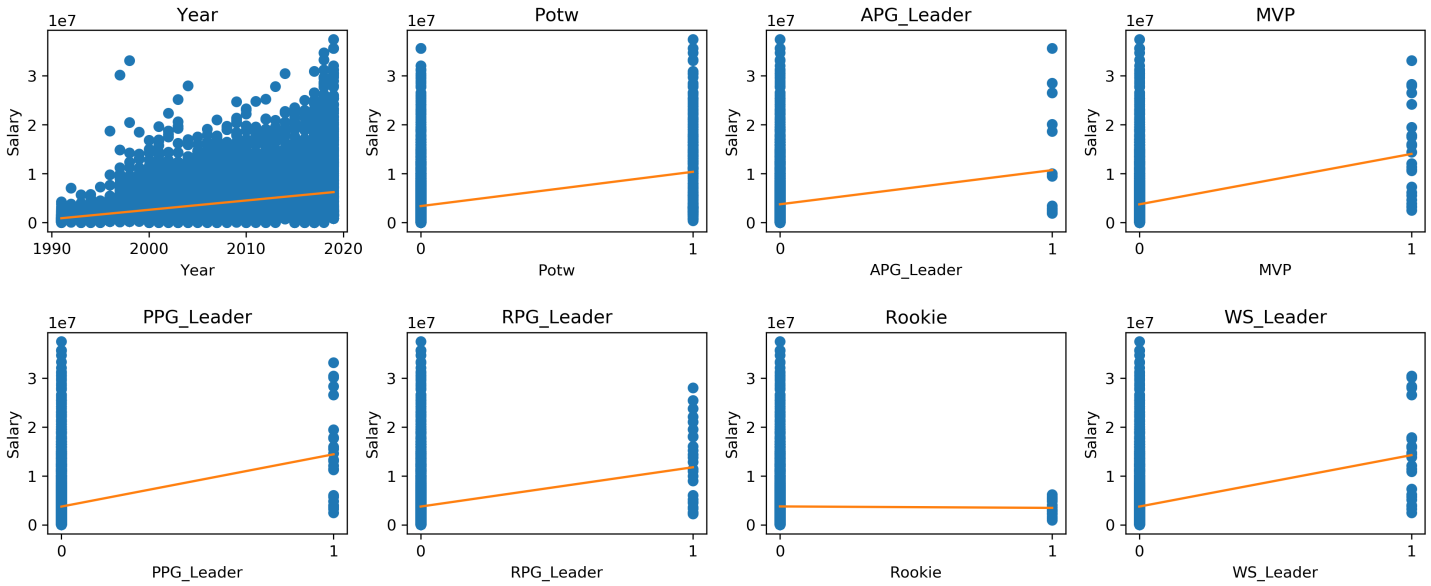
Intercept	G	Three_P	Two_P	FT	DRB	AST	STL	BLK	PF	Pos_PF	Pos_PG	Pos_SF	Pos_SG	Predicted πi
1	56	0.9	2.1	1	2.5	1.5	0.6	0.3	1.9	0	0	0	0	0
1	82	5.1	9.3	0	11.1	10.7	2.4	2.7	3.8	0	1	0	0	1

We use the median statistic of 2019 and max statistic of 2019 to do prediction. As a result, the probability of a player with median performance has 0% chance to win POTW while a player with max performance has 99.99% chance to win POTW. This prediction successfully indicates our model can predict whether a player could win POTW based on his performance to some extent.

Problem 2: Relationship between NBA Titles and Player Salary

Explanatory Analysis

After extracting the relevant player titles and salary from the dataset, we plotted the relationship between the titles and salary using a scatter plot.



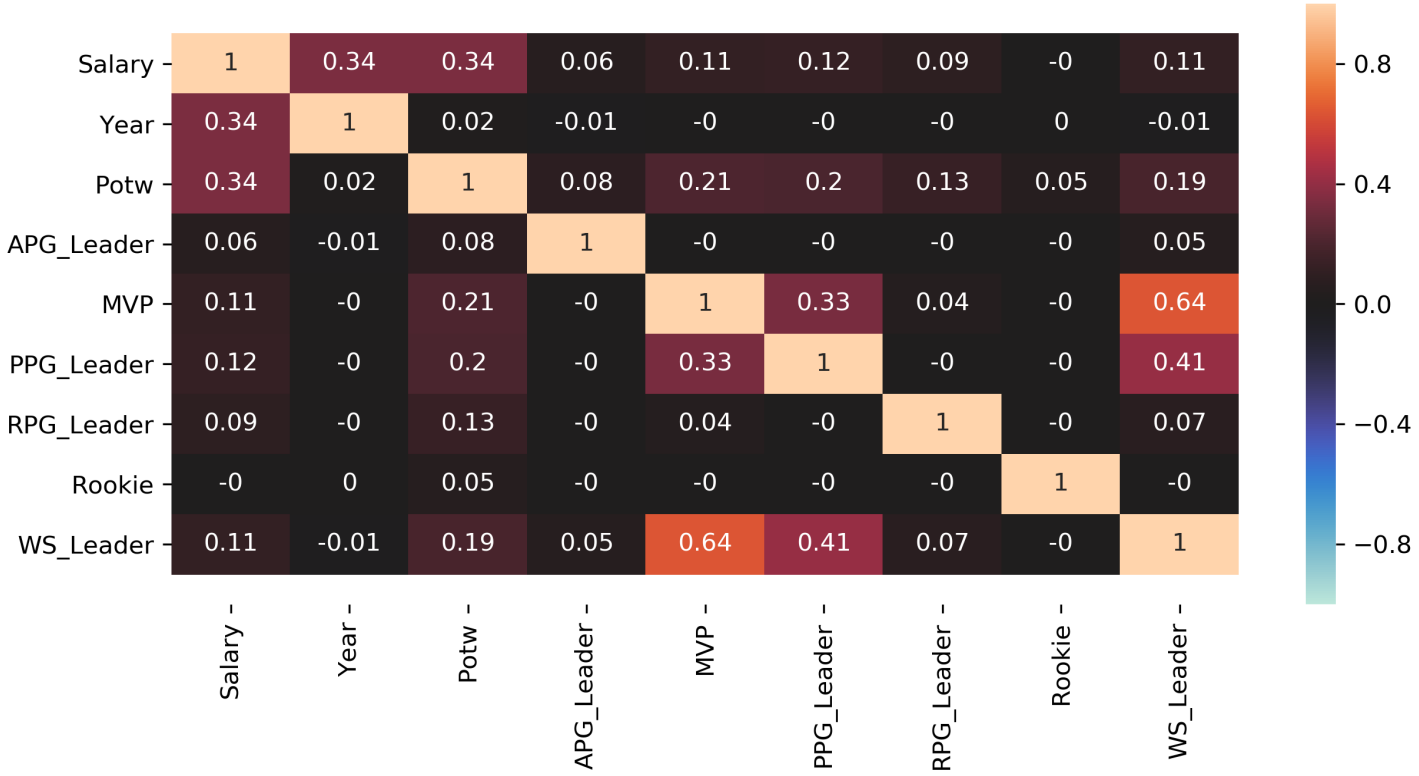
Observing the scatter plot, since all of the variables are binary except for `Year`, it is difficult to interpret the relationships using the scatter plot. Looking at the `Year` plot, there is evidently a positive linear relationship between `Year` and `Salary`. What is also interesting about the `Year` graph is that, despite having a positive relationship, the range of values also increased for every season.

This change in range may have been affected by the increase in observations over the years.

Decade	Count
1990	2416
2000	2998
2010	3589

From the frequency table, we can clearly see the increase in observations over the seasons. The cause of this is unknown; either there is a steady increase in players, or there is a steady increase in data collected. Nonetheless, this might be worth looking into and be cautious about during regression analysis.

We also plotted the correlation using a heatmap.



Observing the heatmap, the overall correlation seems pretty low, except for `WS_Leader` and `MVP`. This means that, except for MVP and Win Shares Leader, having one NBA title does not automatically entitled you to another. It also meant that multicollinearity is likely not an issue in regression analysis.

Regression Analysis

We first fitted the full model with variables `Year`, `Potw`, `APG_Leader`, `MVP`, `PPG_Leader`, `RPG_Leader`, `Rookie`, `WS_Leader`.

Model Summary

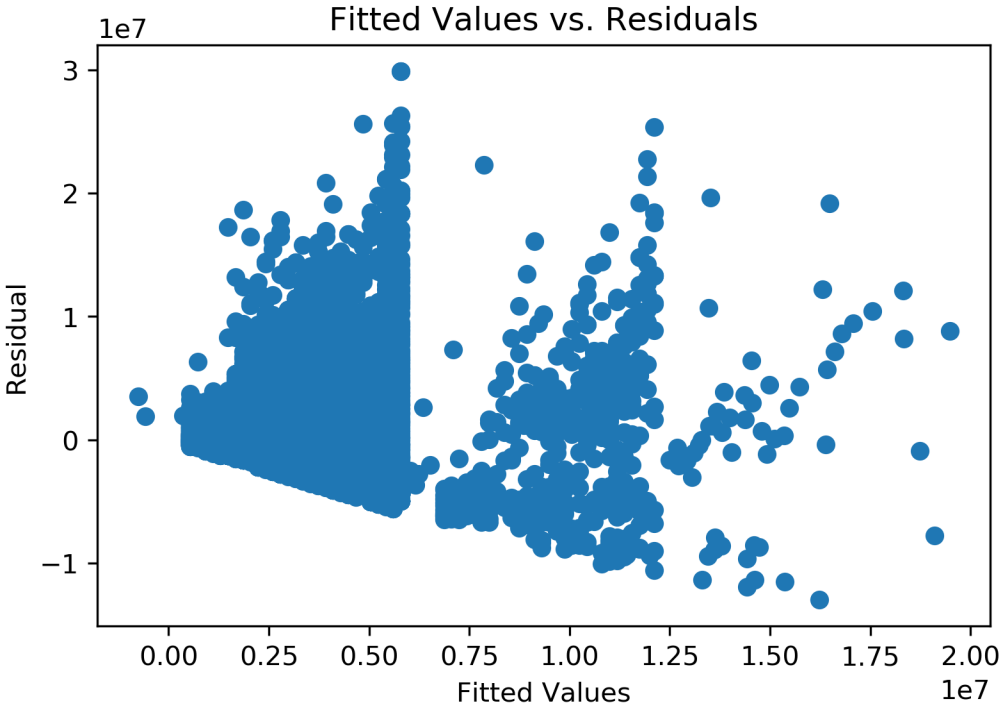
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Salary   R-squared:                       0.231
Model:                            OLS   Adj. R-squared:                  0.231
Method:                 Least Squares   F-statistic:                     338.4
Date:                Sun, 13 Oct 2019   Prob (F-statistic):               0.00
Time:                        19:01:39   Log-Likelihood:            -1.5001e+05
No. Observations:                9003   AIC:                         3.000e+05
Df Residuals:                    8994   BIC:                         3.001e+05
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -3.726e+08   1.05e+07    -35.504      0.000   -3.93e+08   -3.52e+08
Year        1.874e+05   5231.501     35.823      0.000    1.77e+05    1.98e+05
Potw        6.327e+06   2.01e+05     31.404      0.000    5.93e+06    6.72e+06
APG_Leader  4.374e+06   1.05e+06      4.165      0.000    2.32e+06    6.43e+06
MVP         1.353e+06   1.14e+06      1.182      0.237    -8.9e+05     3.6e+06
PPG_Leader  3.988e+06   9.28e+05      4.295      0.000    2.17e+06    5.81e+06
RPG_Leader  4.677e+06   7.99e+05      5.856      0.000    3.11e+06    6.24e+06
Rookie     -2.054e+06    9.6e+05     -2.140      0.032   -3.94e+06   -1.73e+05
WS_Leader   2.213e+06   1.16e+06      1.916      0.055   -5.16e+04    4.48e+06
==============================================================================
Omnibus:                     3646.378   Durbin-Watson:                   1.894
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            18043.676
Skew:                           1.922   Prob(JB):                         0.00
Kurtosis:                       8.773   Cond. No.                     4.79e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.79e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

ANOVA Table

index	df	sum_sq	mean_sq	F	PR(>F)
Year	1	2.3164e+16	2.3164e+16	1330.82	7.34954e-272
Potw	1	2.21228e+16	2.21228e+16	1271	1.67026e-260
APG_Leader	1	2.89658e+14	2.89658e+14	16.6414	4.55432e-05
MVP	1	3.69627e+14	3.69627e+14	21.2359	4.11687e-06
PPG_Leader	1	4.0067e+14	4.0067e+14	23.0193	1.6296e-06
RPG_Leader	1	6.31608e+14	6.31608e+14	36.2872	1.76961e-09
Rookie	1	8.00524e+13	8.00524e+13	4.59918	0.032014
WS_Leader	1	6.38674e+13	6.38674e+13	3.66932	0.0554546
Residual	8994	1.56548e+17	1.74058e+13	nan	nan

From the model summary and ANOVA table, it is evident that `MVP` and `WS_Leader` are not statistically significant variables based on both t-test and F-test. However, since they are highly correlated, as mentioned above, it is likely that only one of them would need to be remove.

Upon checking for non-linearity of the model, we found signs of non-linearity from the residual plot.



Hence, we performed a log transformation on `Salary` to attempt to correct the problem. We refitted the full model using the log-transformed data.

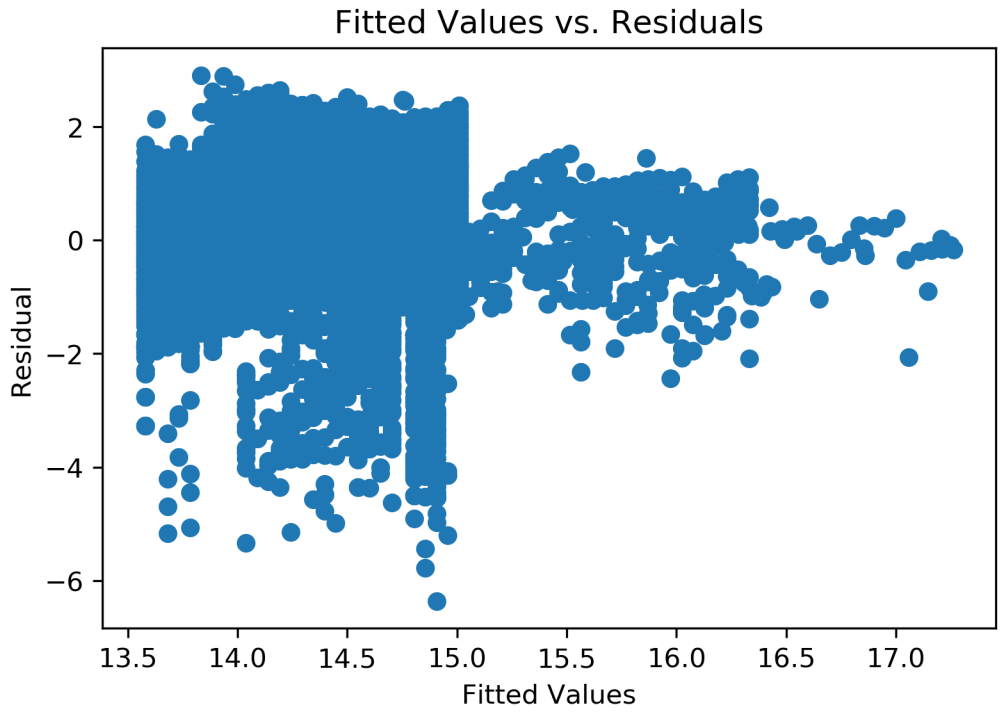
Model Summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Salary   R-squared:                       0.177
Model:                            OLS   Adj. R-squared:                  0.176
Method:                 Least Squares   F-statistic:                     241.3
Date:                Sun, 13 Oct 2019   Prob (F-statistic):               0.00
Time:                        19:01:40   Log-Likelihood:                -14245.
No. Observations:                9003   AIC:                         2.851e+04
Df Residuals:                    8994   BIC:                         2.857e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -88.1389      2.963    -29.744      0.000     -93.948     -82.330
Year           0.0511      0.001     34.585      0.000       0.048       0.054
Potw           1.3214      0.057     23.226      0.000       1.210       1.433
APG_Leader     0.6687      0.297      2.255      0.024       0.087       1.250
MVP            0.0916      0.323      0.284      0.777      -0.542       0.725
PPG_Leader     0.5126      0.262      1.955      0.051      -0.001       1.026
RPG_Leader     0.8803      0.225      3.904      0.000       0.438       1.322
Rookie         0.1315      0.271      0.485      0.627      -0.400       0.663
WS_Leader      0.3642      0.326      1.116      0.264      -0.275       1.004
==============================================================================
Omnibus:                      842.715   Durbin-Watson:                   1.831
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1294.483
Skew:                          -0.709   Prob(JB):                    8.06e-282
Kurtosis:                       4.200   Cond. No.                     4.79e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.79e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

ANOVA Table

index	df	sum_sq	mean_sq	F	PR(>F)
Year	1	1709.68	1709.68	1231.95	4.70618e-253
Potw	1	926.665	926.665	667.731	4.35078e-142
APG_Leader	1	6.77177	6.77177	4.87956	0.0272016
MVP	1	5.36321	5.36321	3.86459	0.0493459
PPG_Leader	1	6.68263	6.68263	4.81533	0.0282331
RPG_Leader	1	22.0516	22.0516	15.8898	6.76709e-05
Rookie	1	0.323805	0.323805	0.233326	0.629081
WS_Leader	1	1.72901	1.72901	1.24588	0.26437
Residual	8994	12481.7	1.38778	nan	nan

From the model summary, the log-transformed data confirmed the statistical insignificance of `MVP`. Meanwhile, the ANOVA table shows that `Rookie` and `WS_Leader` are also statistically insignificant. We kept that in mind for model selection.



The log-transformed data had seemingly reduced the severity of non-linearity.

Model Selection

We then proceeded to Model Selection using Adjusted R^2, Mallow's CP, AIC, and BIC.

Best Subset Regression Table

index	Predictors	Adjusted R^2	Mallows CP	Predictors	AIC	BIC
225	6	0.176143	5.31422	Year, Potw, APG_Leader, PPG_Leader, RPG_Leader, WS_Leader	28505.1	28554.8
166	5	0.176011	5.75517	Year, Potw, APG_Leader, PPG_Leader, RPG_Leader	28505.5	28548.1
218	6	0.176037	6.47921	Year, Potw, APG_Leader, MVP, PPG_Leader, RPG_Leader	28506.2	28556
250	7	0.176073	7.08048	Year, Potw, APG_Leader, PPG_Leader, RPG_Leader, Rookie, WS_Leader	28506.8	28563.7
247	7	0.176059	7.23564	Year, Potw, APG_Leader, MVP, PPG_Leader, RPG_Leader, WS_Leader	28507	28563.8

From this table, we can see that, the models with 5 and 6 predictors performed relatively similar. They differ in whether `WS_Leader` is included as a variable. It is noted above that `WS_Leader` might be insignificant as shown in the F-test result. We chose to regress both models and compare.

Regressing the model with 6 variables, we got the following results.

Model Summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Salary   R-squared:                       0.177
Model:                            OLS   Adj. R-squared:                  0.176
Method:                 Least Squares   F-statistic:                     321.8
Date:                Sun, 13 Oct 2019   Prob (F-statistic):               0.00
Time:                        19:01:47   Log-Likelihood:                -14246.
No. Observations:                9003   AIC:                         2.851e+04
Df Residuals:                    8996   BIC:                         2.855e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -88.1433      2.963    -29.748      0.000     -93.951     -82.335
Year           0.0511      0.001     34.590      0.000       0.048       0.054
Potw           1.3246      0.056     23.453      0.000       1.214       1.435
APG_Leader     0.6636      0.296      2.241      0.025       0.083       1.244
PPG_Leader     0.5178      0.261      1.983      0.047       0.006       1.030
RPG_Leader     0.8778      0.225      3.894      0.000       0.436       1.320
WS_Leader      0.4165      0.267      1.563      0.118      -0.106       0.939
==============================================================================
Omnibus:                      843.220   Durbin-Watson:                   1.831
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1295.231
Skew:                          -0.709   Prob(JB):                    5.55e-282
Kurtosis:                       4.200   Cond. No.                     4.79e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.79e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

ANOVA Table

index	df	sum_sq	mean_sq	F	PR(>F)
Year	1	1709.68	1709.68	1232.18	4.21791e-253
Potw	1	926.665	926.665	667.856	4.09393e-142
APG_Leader	1	6.77177	6.77177	4.88048	0.0271872
PPG_Leader	1	10.0137	10.0137	7.21698	0.00723497
RPG_Leader	1	22.3129	22.3129	16.0811	6.11763e-05
WS_Leader	1	3.38751	3.38751	2.44141	0.118205
Residual	8996	12482.1	1.38752	nan	nan

Regressing the model with 5 variables, we got the following results.

Model Summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Salary   R-squared:                       0.176
Model:                            OLS   Adj. R-squared:                  0.176
Method:                 Least Squares   F-statistic:                     385.6
Date:                Sun, 13 Oct 2019   Prob (F-statistic):               0.00
Time:                        19:01:48   Log-Likelihood:                -14247.
No. Observations:                9003   AIC:                         2.851e+04
Df Residuals:                    8997   BIC:                         2.855e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -88.1045      2.963    -29.734      0.000     -93.913     -82.296
Year           0.0511      0.001     34.575      0.000       0.048       0.054
Potw           1.3347      0.056     23.783      0.000       1.225       1.445
APG_Leader     0.6846      0.296      2.314      0.021       0.105       1.265
PPG_Leader     0.6752      0.241      2.802      0.005       0.203       1.147
RPG_Leader     0.9018      0.225      4.010      0.000       0.461       1.343
==============================================================================
Omnibus:                      842.383   Durbin-Watson:                   1.830
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1293.724
Skew:                          -0.709   Prob(JB):                    1.18e-281
Kurtosis:                       4.200   Cond. No.                     4.79e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.79e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

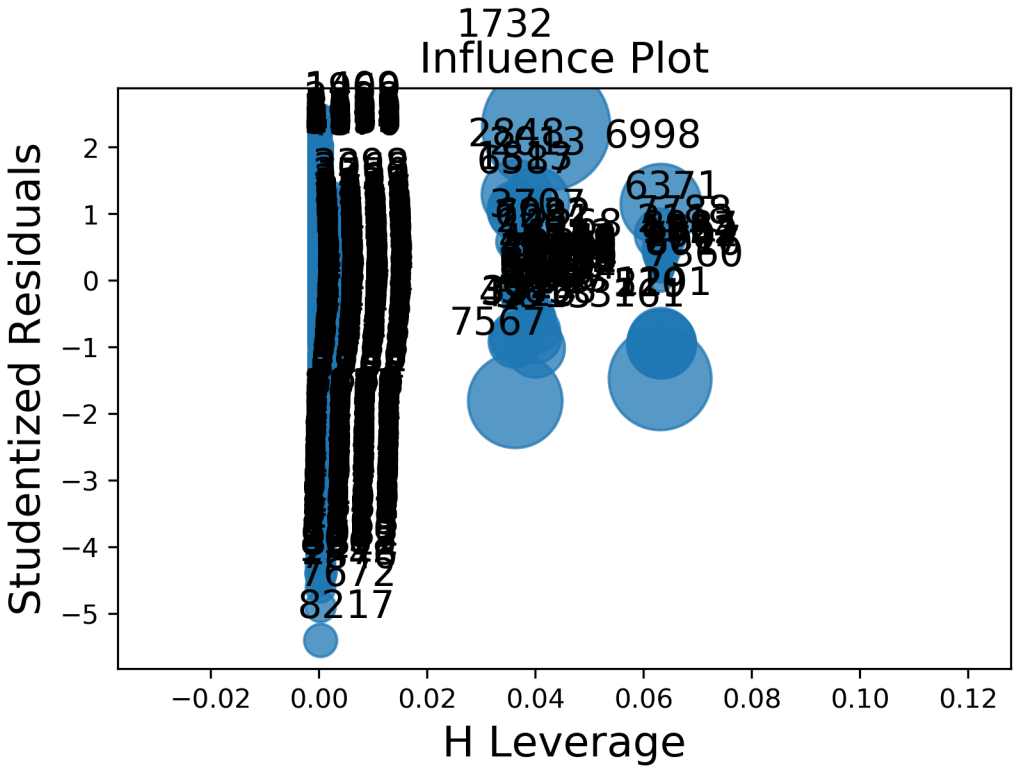
ANOVA Table

index	df	sum_sq	mean_sq	F	PR(>F)
Year	1	1709.68	1709.68	1231.98	4.58256e-253
Potw	1	926.665	926.665	667.749	4.29787e-142
APG_Leader	1	6.77177	6.77177	4.8797	0.0271995
PPG_Leader	1	10.0137	10.0137	7.21582	0.00723963
RPG_Leader	1	22.3129	22.3129	16.0786	6.12594e-05
Residual	8997	12485.5	1.38774	nan	nan

It is clear that the model `Salary ~ Year + Potw + APG_Leader + PPG_Leader + RPG_Leader` performed better without `WS_Leader`, as shown in t-test and F-test results in the model summary and ANOVA table.

Model Diagnosis

We first checked whether influential points exist in the model.



From the influence plot, it is evident that some of the observations are influential. After calculating the Cook's Distance for each observation, we found that 134 observations are influential under the `4 / n - p` heuristic threshold, which is 1.49% of the data. We attempted remove these outliers and refit the model.

Model Summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Salary   R-squared:                       0.202
Model:                            OLS   Adj. R-squared:                  0.201
Method:                 Least Squares   F-statistic:                     447.8
Date:                Sun, 13 Oct 2019   Prob (F-statistic):               0.00
Time:                        19:02:13   Log-Likelihood:                -13599.
No. Observations:                8869   AIC:                         2.721e+04
Df Residuals:                    8863   BIC:                         2.725e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -92.9023      2.854    -32.550      0.000     -98.497     -87.307
Year           0.0535      0.001     37.586      0.000       0.051       0.056
Potw           1.4459      0.056     25.768      0.000       1.336       1.556
APG_Leader     0.6319      0.650      0.972      0.331      -0.642       1.906
PPG_Leader     0.5933      0.295      2.013      0.044       0.016       1.171
RPG_Leader     0.7595      0.283      2.684      0.007       0.205       1.314
==============================================================================
Omnibus:                      391.092   Durbin-Watson:                   1.861
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              462.930
Skew:                          -0.496   Prob(JB):                    2.99e-101
Kurtosis:                       3.518   Cond. No.                     4.81e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.81e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

ANOVA Table

index	df	sum_sq	mean_sq	F	PR(>F)
Year	1	1866.2	1866.2	1483.66	2.92382e-300
Potw	1	935.444	935.444	743.696	2.5721e-157
APG_Leader	1	1.04749	1.04749	0.832772	0.361497
PPG_Leader	1	4.77119	4.77119	3.79319	0.051493
RPG_Leader	1	9.06309	9.06309	7.20534	0.00728222
Residual	8863	11148.1	1.25783	nan	nan

Observing the model summary and the ANOVA table, it seems that `APG_Leader` and `PPG_Leader` were rendered statistically insignificant after the removal of outliers. We don't believe that dropping more variables is the right approach, so we kept the outliers in the final mode and proceeded.

We then checked for heteroscedasticity using the Breusch-Pagan test.

Breusch-Pagan Test Results

LM Statistic	LM-Test p-value	F-Statistic	F-Test p-value
131.12	1.3766e-26	26.5939	8.87478e-27

From the p-values of LM-test and F-test in the Breusch-Pagan test, we determined that, it is unlikely that the model suffers from heteroscedasticity.

We then checked for multicollinearity using both the Breusch-Godfrey test and VIF.

Breusch-Godfrey Results

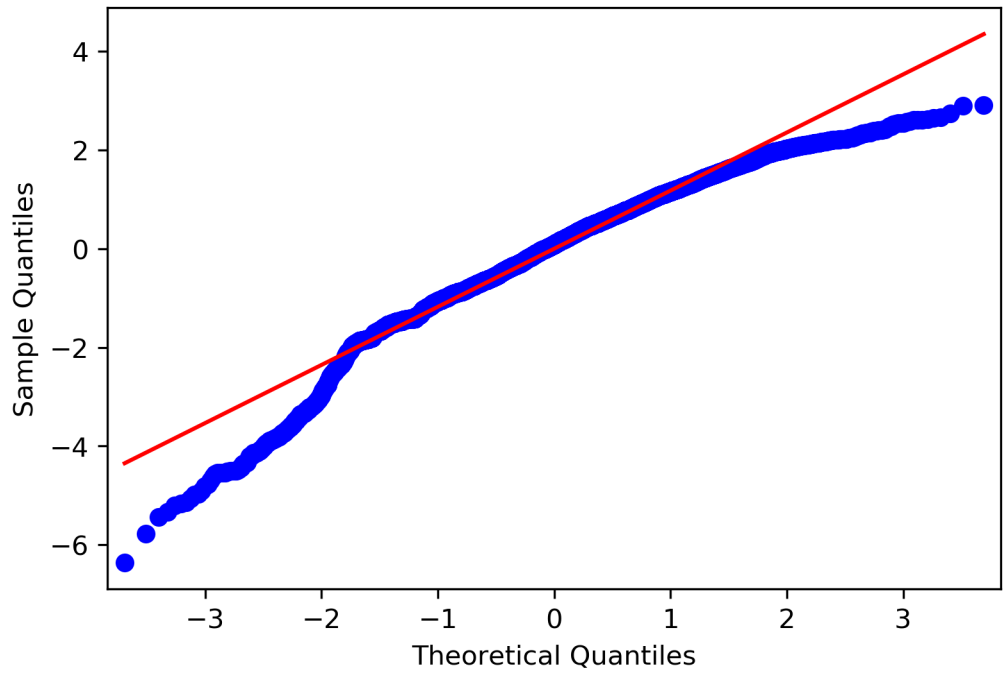
LM Statistic	LM-Test p-value	F-Statistic	F-Test p-value
209.138	2.76112e-26	5.91981	1.25427e-26

VIF Test Results

Features	VIF Factor
Year	1.0598
Potw	1.12948
APG_Leader	1.00912
PPG_Leader	1.04567
RPG_Leader	1.02037

From the p-values of LM-test and F-test in the Breusch-Godfrey test, as well as the VIF factors from the VIF test result, we determined that, it is unlikely that the model suffers from multicollinearity.

We finally checked for non-normality using QQ plot.



From the QQ plot, we determined that, it is unlikely that the model suffers from non-normality.

Final Model Summary

Variables

Type	Intercept	Year	Potw	APG_Leader	PPG_Leader	RPG_Leader
βi	-88.1045	0.0510719	1.33466	0.684587	0.675161	0.901754
e^(βi)	5.45361e-39	1.0524	3.79872	1.98295	1.96435	2.46392

Formula

$$\begin{aligned} log(\hat{y}_{Salary}) = & \quad \hat{\beta}_0 + \hat{\beta}_{Year} \cdot X_{Year} + \hat{\beta}_{Potw} \cdot X_{Potw} + \hat{\beta}_{APGLeader} \cdot X_{APGLeader} + \hat{\beta}_{PPGLeader} \cdot X_{PPGLeader} + \hat{\beta}_{RPGLeader} \cdot X_{RPGLeader} \\ = & -88.1045 + 0.0510719 \cdot X_{Year} + 1.33466 \cdot X_{Potw} + 0.684587 \cdot X_{APGLeader} + 0.675161 \cdot X_{PPGLeader} + 0.901754 \cdot X_{RPGLeader} \end{aligned}$$

Interpretation of Model

- `Year`: Regardless the award, a player would tend to earn **5.24%** more salary than last year.
- `Potw`: If a player is a Player of the Week (POTW), he would tend to earn **2.8** times more than non-POTW.
- `APG Leader`: If a player is an Assists Per Game Leader (APG Leader), he would tend to earn **98.3%** more than non-APG_Leader.
- `PPG Leader`: If a player is a Points Per Game Leader (PPG Leader), he would tend to earn **96.44%** more than non-PPG_Leader.
- `RPG Leader`: If a player is a Points Per Game Leader (RPG Leader), he would tend to earn **1.46** times more than non-RPG_Leader.

To summarize, the model indicates that NBA player's salary will naturally increase each year by 5.24%. If an NBA player can earn an award such as POTW, APG Leader, PPG Leader, or RPG Leader, his salary would significantly higher than those who don't receive awards. Within these awards, **POTW** mostly reflect a player's value since POTW earns most. Besides, **RPG Leader** also earns much maybe because these players can make use of their body to play basketball and thus their advantage is stable and hard to be replaced by other guys. Therefore, their salaries tend to be higher.

Prediction of Model

Intercept	Year	Potw	APG_Leader	PPG_Leader	RPG_Leader	Predicted Salary
1.00	2020.00	0.00	0.00	0.00	0.00	3473656.83
1.00	2020.00	1.00	0.00	0.00	0.00	13195447.58
1.00	2020.00	1.00	0.00	1.00	0.00	25920463.52

We predict three situations based on this model.

- Predict in year 2020 a player's salary when he doesn't have any awards.
- Predict in year 2020 a player's salary when he only wins POTW.
- Predict in year 2020 a player's salary when he wins POTW as well as PPG Leader.

The result shows that the average salary for NBA player is about 3.47 millions if he doesn't win any awards. However, if a player wins at least one time POTW, it means it's predicted salary could be 13.2 millions, which is a lot more higher than non_POTW. What's more, PPG Leader could also indicate a player's higher salary.

Summary

The first model shows us how player's performance can predict whether a player could win POTW. Despite unfixable multicollinearity, this logistic model is fitted with many important predictors such as 3-points field goals and 2 points field goals, quantitatively giving us a way to predict the chance of a player to get POTW based on his statistics.

The second model fits well and indicate the relationship between player's awards and his salaries. Unexpectedly, NBA player's salary would be higher if he could win POTW, PPG Leader, RPG Leader and so on. However, MVP and WS Leader seems not so significant and thus are excluded in the model. We assume that it might be due to time lag and awards might better reflect next year's salary. This is what we could improve for the further research.