Predicting NBA Players' Salaries

Allen Chun

6/15/2021

UCLA Department of Statistics

Course: STATS 101A – Introduction to Data Analysis and Regression

Table of Contents

Abstract **3**

Introduction **3**

Methodology **6**

Matrix and Correlation Plots 6

Variance Inflation Factor (VIF) 7

Diagnostic Plots 8

Transformations 9

Marginal Model Plots (MMPS) 10

Leverage Plots 11

Creating and Modifying Variables 11

Results **12**

Discussion **12**

Limitations and Conclusions **12**

References **13**

Feature	Response
Name	Allen Chun
SID	205308456
Kaggle Nickname	Allen Chun
Kaggle Rank	3
Kaggle R^2	0.73914
Total Number of Predictors	9
Total Number of Betas	24
BIC	13053.75
Complexity Grade	106

Abstract

The purpose of this project was to generate a multiple linear regression model that best predicted the salary of NBA players in a random year. We were to create a model that showed the relationship between NBA players's in-game statistics and their salaries. Through various regression techniques, a multiple linear regression model was built with a select number of variables to predict the salary of NBA players In total, 24 predictors were used. The model was first developed through a training dataset, and was then submitted to a class Kaggle competition, in which it placed third on the leaderboard with an R^2 value of 0.73914.

Introduction

The purpose of this project was to create a linear regression model to understand the factors on which the salary of NBA players depends on. Utilizing techniques of multiple linear regression, we were provided a dataset that contained 66 predictors and 420 observations of different basketball players across the NBA to try and find the best predictors for their salaries. The final regression model was then submitted to a class Kaggle competition, in which it was used to predict the salary of NBA players of a testing data set that contains 180 observations. Below are the set of initial predictors provided with the dataset:

Table 2: Type of Predictors

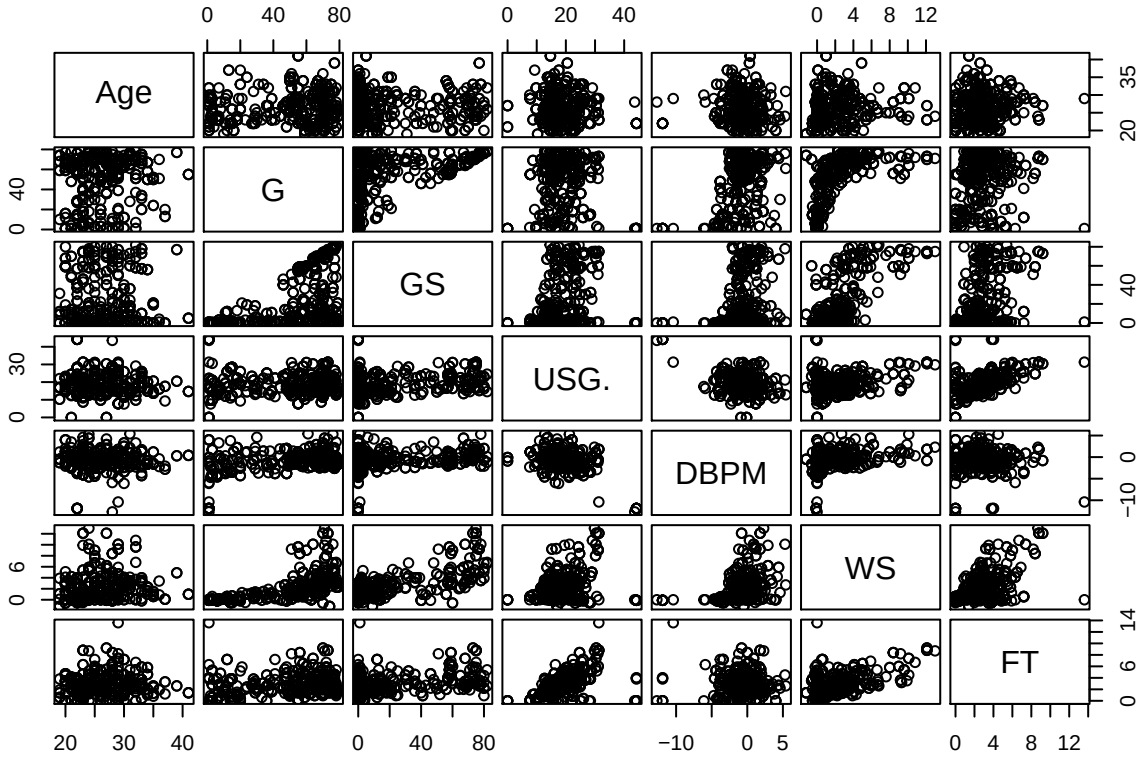
Variables	Type
NBA_Country	Categorical
Age	Numerical
TM	Categorical
G	Numerical
MP	Numerical
PER	Numerical
TS.	Numerical
X3PAr	Numerical
FTr	Numerical
ORB.	Numerical
DRB.	Numerical
TRB.	Numerical
AST.	Numerical
STL.	Numerical
BLK.	Numerical
TOV.	Numerical
USG.	Numerical
OWS	Numerical
DWS	Numerical
WS	Numerical
WS.48	Numerical
OBPM	Numerical
DBPM	Numerical
BPM	Numerical
VORP	Numerical
Rk	Numerical
Pos	Categorical
GS	Numerical
FG	Numerical
FGA	Numerical
FG.	Numerical
X3P	Numerical
X3PA	Numerical
X3P.	Numerical
X2P	Numerical
X2PA	Numerical
X2P.	Numerical
FT	Numerical
FTA	Numerical
FT.	Numerical
ORB	Numerical
DRB	Numerical
TRB	Numerical
AST	Numerical
STL	Numerical
BLK	Numerical
TOV	Numerical
PF	Numerical
PTS	Numerical

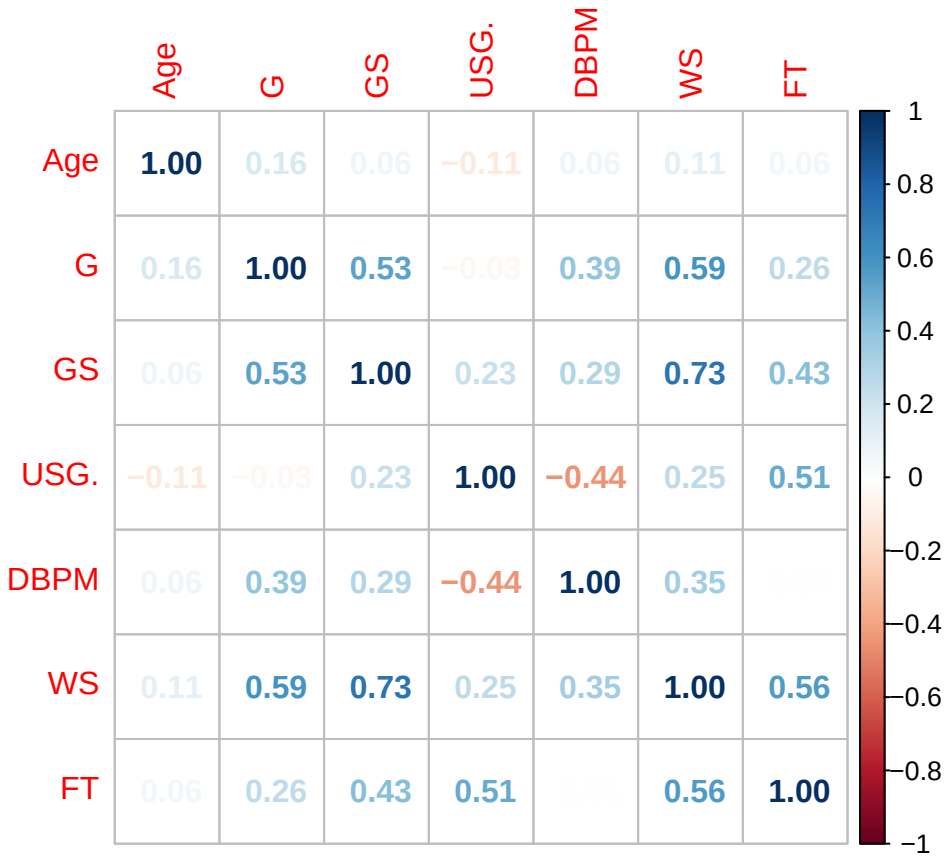
Variables	Type
Ortg	Numerical
DRtg	Numerical
Team.Rk	Numerical
Team	Categorical
T.Conf	Categorical
T.Div	Categorical
T.W	Numerical
T.L	Numerical
T.W.L.PERC	Numerical
T.MOV	Numerical
T.Ortg	Numerical
T.DRtg	Numerical
NRtg	Numerical
MOV.A	Numerical
Ortg.A	Numerical
DRtg.A	Numerical
NRtg.A	Numerical

Methodology

Matrix and Correlation Plots

Below are matrix plots and correlation plots for the numerical predictors used in the multiple linear regression model. The predictors with the lowest correlation are G and USG. while the predictors with the highest correlation are GS and WS. The variable MP is also very significant in regards to its relation to Salary, but I did not include it in my model due to possible violations in model assumptions.





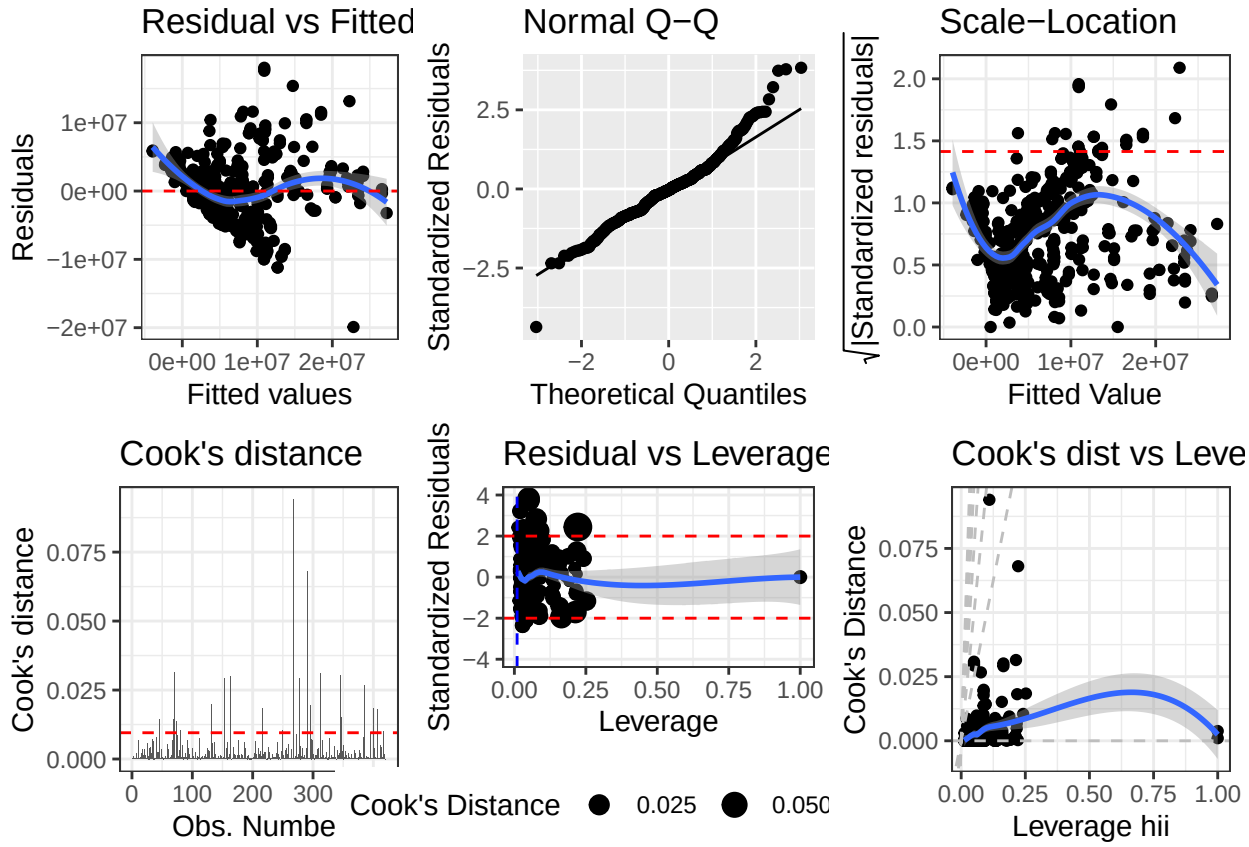
Variance Inflation Factor (VIF)

The VIF for all the predictors are shown below. As seen, none of the predictors have a VIF greater than 5. As a result, there is no serious problem with multicollinearity. Even though minutes played (MP) was one of the most significant variables that was related to a player's salary, I could not include the MP variable in the model because its VIF was too high.

```
##                  GVIF Df GVIF^(1/(2*Df))
## NBA_Country 2.608536 10        1.049107
## TM          2.738915  7        1.074622
## Age         1.171805  1        1.082500
## G           2.016470  1        1.420025
## GS          2.774404  1        1.665654
## USG.        2.207630  1        1.485809
## DBPM        1.949778  1        1.396345
## WS          3.961453  1        1.990340
## FT          1.977571  1        1.406261
```

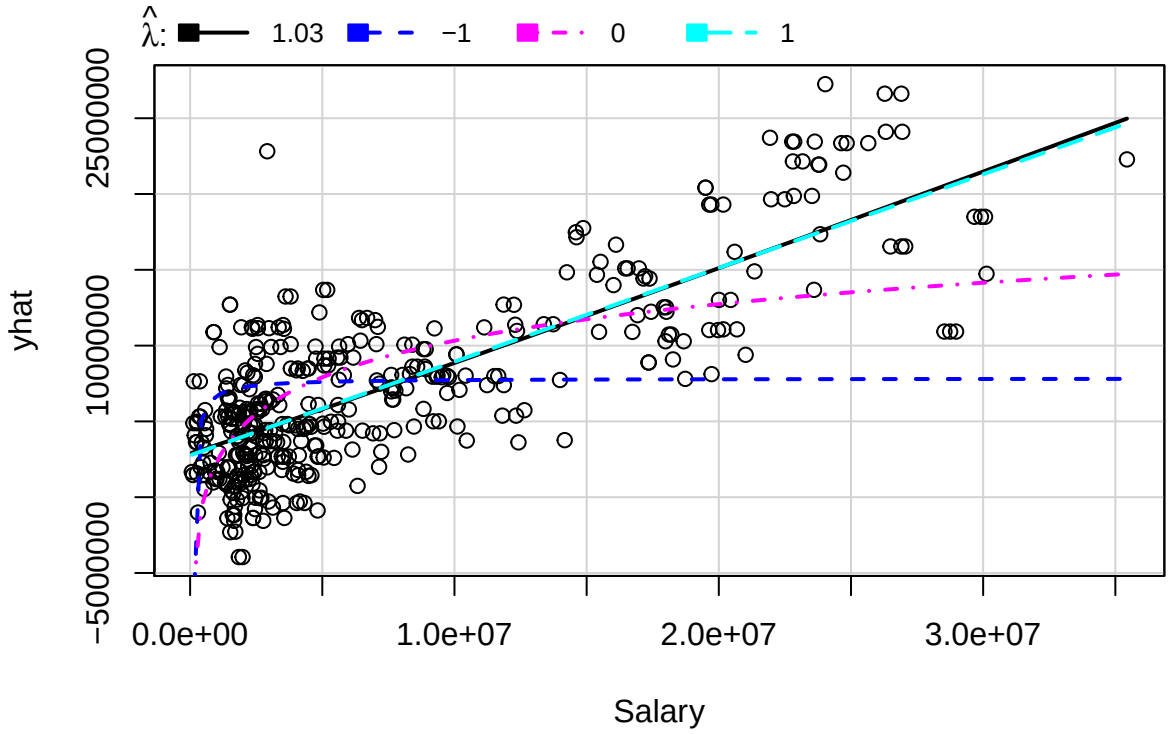
Diagnostic Plots

Below are the diagnostic plots for the model. The residuals show that constant variance isn't maintained throughout the entire plot. Looking at the leverage points, we can see that there are no bad leverage points. Instead, there are multiple outliers as well as a few good leverage points.



Transformations

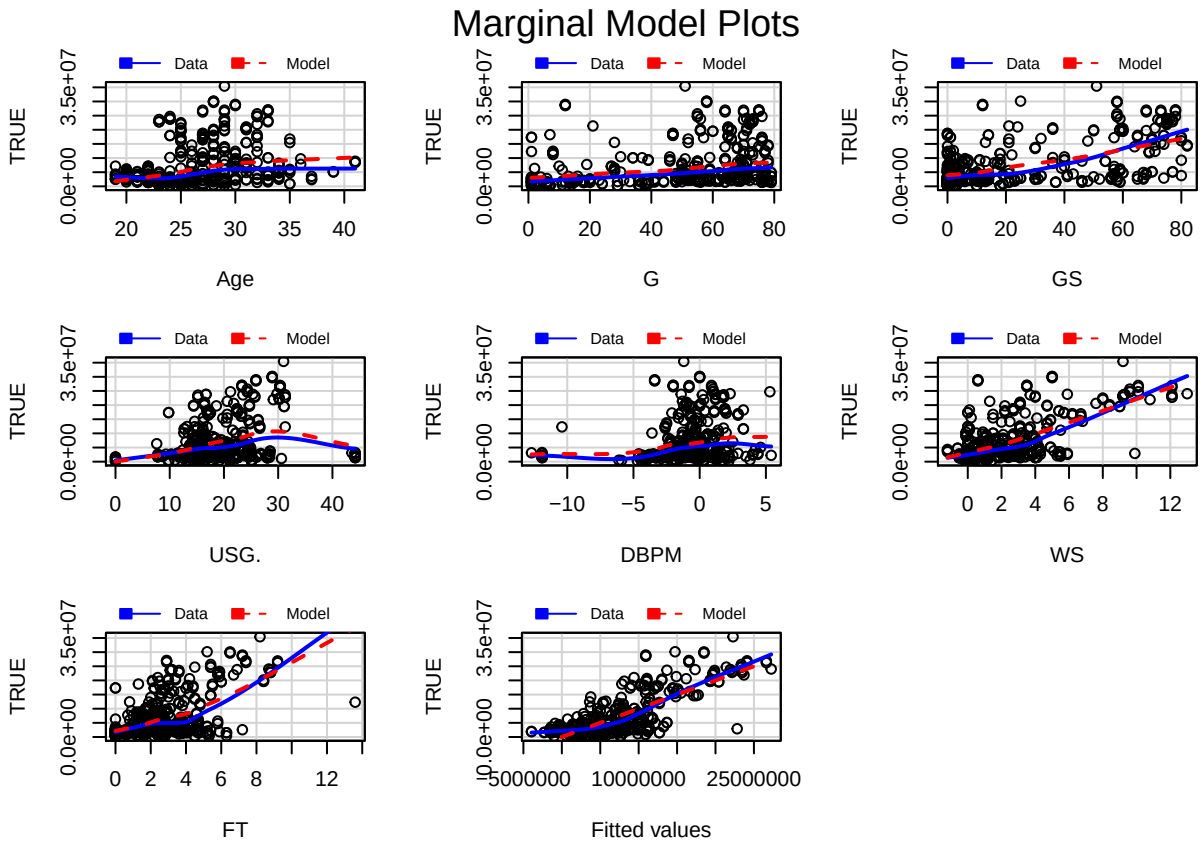
The `inverseResponsePlot()` function suggested a λ of 1.034336 to the response variable. After running the suggested λ to the regression model, I left the λ out of the model because it did not improve the R^2 by a lot. Using the `powerTransform()` also results in using powers that did not improve the R^2 by a lot.




```
##      lambda          RSS
## 1  1.034336 5.703071e+15
## 2 -1.000000 1.415652e+16
## 3  0.000000 8.296801e+15
## 4  1.000000 5.704668e+15
```

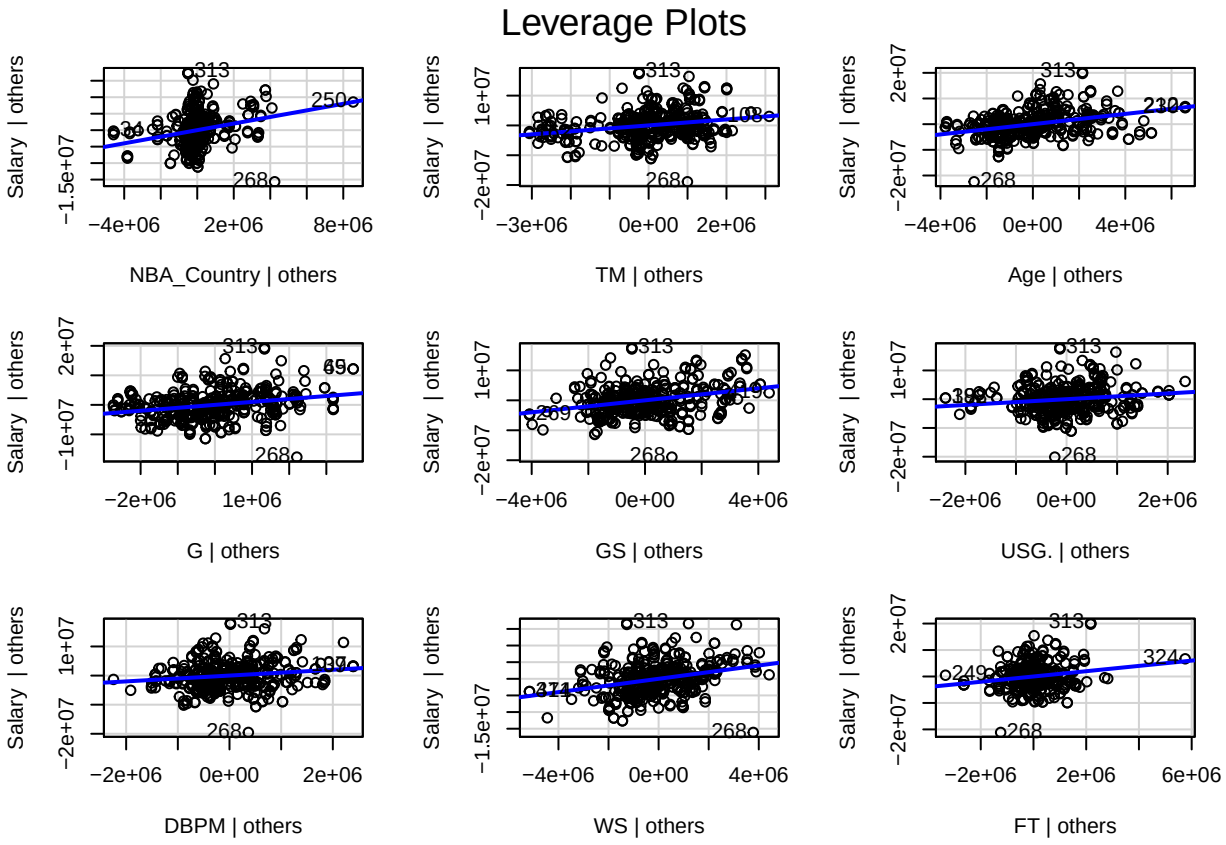
Marginal Model Plots (MMPS)

The marginal model plots show nearly identical lines between the blue data lines and the red model lines. This means that the relationship between the estimated linear regression model and the estimated non-parametric model is linear. Therefore, the multiple linear regression model used is an adequate one.



Leverage Plots

Leverage plots show that most of the plots follow a mostly linear pattern that is relatively positive. The plots show that there is no serious problem with multicollinearity and no clear violations or unusual patterns with the model assumptions.



Creating and Modifying Variables

I changed the NBA_Country predictor from the player's country or origin to the player's continent of origin in order to gain a better understanding of how salary is different for players from different parts of the world. This increased the R^2 by 0.0262. I also changed the TM predictor from each individual team and clustered them by salary cap. This increased the R^2 by 0.0228.

Results

My final multiple linear regression model had an R^2 of 0.76228 on the training data and an R^2 of 0.73914 on the testing data. While there was a slight violation in the residual patterns, the adequacy of the variance inflation factors, marginal model plots, and leverage plots show that this model is a sufficient one to use.

Discussion

The team (TM), games started (GS), and win shares (WS) seem to have the biggest effect on a player's salary. While age (Age) is also an important factor, it doesn't drastically affect their salary as much as the other factors listed above. The team that a player plays for largely affects their salary, primarily because each team has a different salary cap for their organization. The region from where a player is from also affects their salary because USA is known to have the best players in the world whereas players in Asia or Europe are not known to be as good as the Americans.

Limitations and Conclusions

In regards to the diagnostic plots, there looks to be a slight violation to the constant variance assumption, with a slight increasing trend as the fitted value increases.

In general, the biggest limitation is the inability of this model to accurately predict the salaryies of the testing dataset. Despite attempting to built many models with higher R^2, my final Kaggle R^2 was significantly lower compared to what was achieved through the training dataset. The models I built that had a higher R^2 had more violations, so it was difficult to justify the validity of the model. However, it isn't surprising to see that the testing dataset has a lower R^2, considering there are many other variables that make it nearly impossible to accurately predict an NBA player's salary.

Another problem with this model is the fact that there are 24 betas. There are probably other ways in which a model with less predictors can have a similar or higher R^2.

References

Almohalwas, Akram. 2021. *Chapter 5 updated Winter 2020.*

Almohalwas, Akram. 2021. *Statistics 101A Chapter 6.*

———. 2021. *STATS 101A Spring 2021 Kaggle Competition: Predicting Player's Salary Based on his Statistics.*

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