STAT 410 Final Project:
A Multiple Linear Regression Approach for Analyzing NBA Player Salary via Performance

Alexander Xiong

Abstract

Understanding the relationship between National Basketball Association (NBA) player salary and performance is very important to setting up contracts that fairly measure a player's worth, as well as, are manageable and fit the needs for a team. This project is concerned with the problem of determining the relationship between a basketball player's salary and the player's performance profile. For this purpose, this project is to investigate the relationship between player salary and player features including various player per game performance metrics (e.g., points, rebounds, and shooting performance) and non-performance statistics (e.g., age and position). Specifically, we first study and extract eight years of NBA player data from 2012 to 2020, and then apply linear regression models to determine which features play a vital role in player salary. That is, we fitted simple linear regression models and multiple linear regression models on the individual performance-salary data for each year. Due to the cross-correlation of performance statistics, we used ridge and lasso regression to minimize multicollinearity and predictors, respectively. Using the subset of predictors obtained from lasso regression, we set up a generalized additive model for each performance-salary data. Our extensive data analysis shows that the proposed regression model provides insight into the relationship between a player's performance and salary.

1 Project Statement

I grew up watching a lot of basketball and this past year, I played fantasy basketball for the first time. For fantasy basketball, I got a chance to look more closely at player performance to maximize my own roster's production. Having found datasets for the salaries of NBA players, I wanted to examine the relationship between a player's performance and salary. The purpose of this analysis is to determine what performance metrics can predict a player's salary. I would want to extend this to determining if a player is worth their contract for a specific year. After computing a model that would fit player data to salary, I would be able to see which players were above the fitted line (outperform) and which were below the fitted line (underperform). Furthermore, this could provide insight into the process of how general managers/front offices determine a player's contract. As there have been several players whose production have drastically declined after receiving a large contract as well as several players who start playing at a much higher level than what they are paid for, it would be impractical to say that player salary could solely determine the skill level of a player. Rather, a player's contract reflects the determined "worth" of a player to a specific team. Teams offer contracts to players that they believe are valuable assets (e.g. young players, veteran presence, trade assets, etc.). Therefore, this analysis could be summarized as determining if a player had been appropriately evaluated when receiving a contract offer.

To keep this project as complete as possible, we are investigating the relationship between all recorded performance statistics on a per-game basis with a player's salary. While advanced statistics can be more representative of a player's true value, we think that it would be a more fruitful analysis to forego advanced metrics.

2 Data Collection

We obtained the player performance data from the 2013 season to the 2020 season using [2] (to access previous seasons, change the year in the url as necessary). We obtained the player salary data from [3]. Unfortunately, Basketball Reference only hosts the contracts for the current year; thus, to examine previous years contracts, we entered the link above into the Wayback Machine [4] to access past snapshots of the url. This was the limiting factor as to why we only processed eight years of data. Basketball Reference had no data regarding

contracts prior to the 2012-2013 NBA season.

We cleaned the data using preprocess.py that is located at the link given in Section 6. This program processes salary data and matches it to the correct player on the correct team in the correct year. We implemented a baseline for a player's performance statistics to be included in this study. A player had to have played in more than five games and more than one minute per game. The player had to have also attempted more than one field goal per game and more than 0.2 free throws per game and averaged more than one point per game over the course of one season. The processed csv files are located in the stats_sal directory at the link given in section 6.

A comprehensive explanation of the performance statistics can be found in [1]:
Pos, Age, Tm, G, GS, MP, FG, FGA, FG%, X3P, X3PA, X3P%, X2P, X2PA, X2P%, eFG%, FT, FTA, FT%, ORB, DRB, TRB, AST, STL, BLK, TOV, PF, PTS.

3 Exploratory Data Analysis

The breakdown of the data is shown in Table 1:

	2020	2019	2018	2017	2016	2015	2014	2013	Total
C	90	95	101	97	92	89	87	89	740
PF	95	104	94	105	105	109	93	99	804
SF	85	87	80	88	90	94	97	97	718
SG	115	129	115	100	97	116	104	100	876
PG	81	106	101	99	103	116	92	91	789
Total	446	521	491	489	487	524	473	476	3927

Table 1: Data breakdown

As shown in Figure 1, we plotted a histogram of player salary, age, minutes played, and points. We were recommended and decided to work with the salary log base-10 scaled rather to help with accounting for the skewness of the salary data. Otherwise, super-max contracts (the largest contract a player could be offered) would heavily skew salary data right. As seen above, the plots for salary, age, and points per game are fairly normally distributed while the plot for minutes per game plateaus. We would have also included a scatter plot of the total performance-salary data, but given the number of parameters, there was no feasible method of producing the data on one graph.

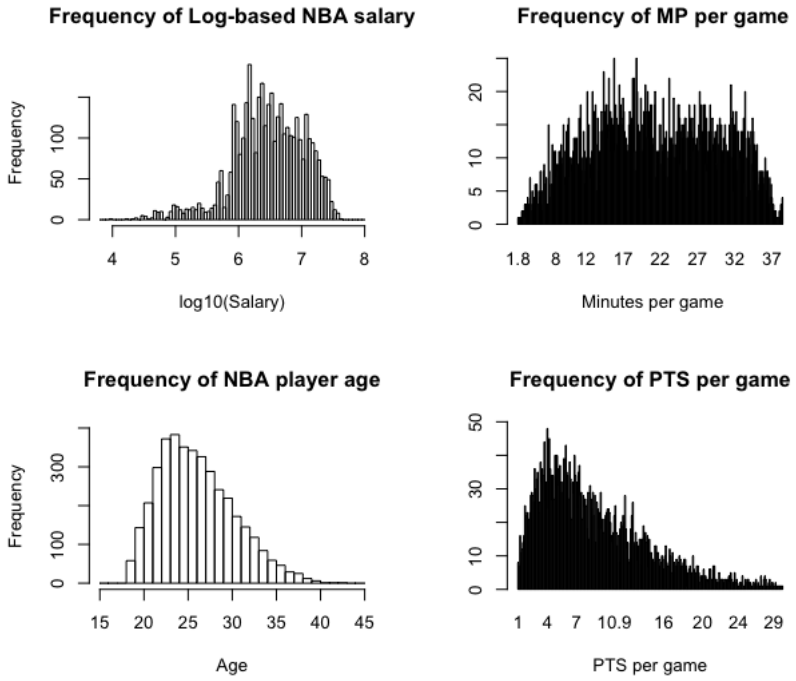


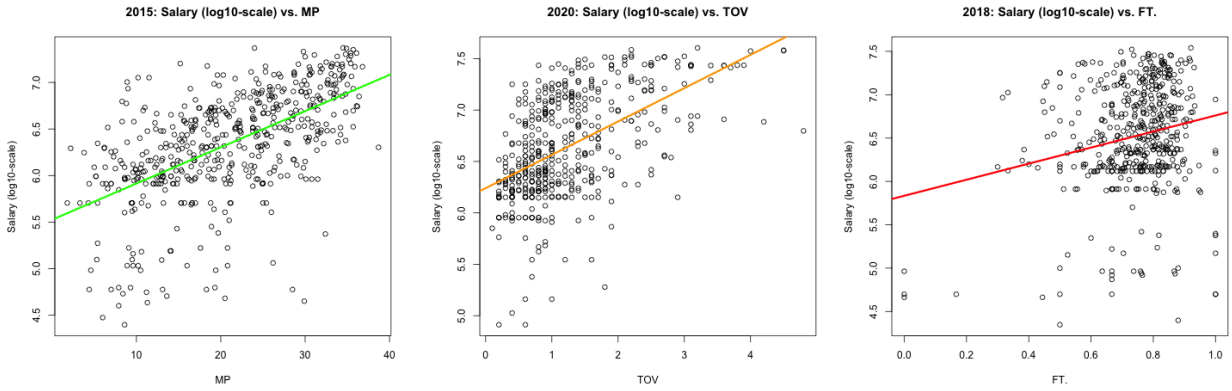
Figure 1: Histograms of Salary, Minutes Played, Age, and Points over 8 years

4 Data Analysis

We apply simple and multiple linear regression as studied in [5, 6].

4.1 SLR Analysis

Working with data from each season individually, we decided to set up a linear regression model determining the correlation between a single performance statistic with the log-based salary data. Thus, for each of the eight years and for the 26 performance statistics (excluding position), we generated the SLR model plot and the diagnostic plot for that model. These figures can be found in each year directory under the specified performance statistic 6. Moreover, we determined if the SLR model was a satisfactory fit by using the mean of the 75% quartile values of all the adjusted R-squared values from the models. All the models with green SLR lines were in the 75% quartile of adjusted R2 values, all the models with orange SLR lines were in the inter-quartile range, and all the models with red SLR lines were below the 25% quartile. The models are found under each year and performance statistic in Section 6. We can see in Figure 2. below that even though MP is in the 75% quartile of fitting salary, the data is too scattered to be fitted very well.



(a) 2015: Salary vs. MP (b) 2020: Salary vs. TOV (c) 2018: Salary vs. FT%

Figure 2: The relationship between salary and MP/TOV/FT% from different years

	2020	2019	2018	2017
Significant	MP, FG, FGA X2P, X2PA, FT FTA, TOV, PTS	MP, FG, PTS	GS	GS, MP, FG FGA, X2PA PTS
Insignificant	G, FG%, X3P% X2P%, eFG% FT%	FG%, X3P% X2P%, eFG% FT%, ORB, BLK	FG%, X3P% X2P%, eFG% FT%	FG%, X3P% X2P%, eFG% FT%

	2016	2015	2014	2013
Significant	G, GS, MP, FG FGA, X2P, X2PA DRB, PTS	MP, FG, FGA X2P, X2PA DRB, PTS	MP, FG FGA, X2P X2PA, PTS	MP, FG, FGA, X2P X2PA, FT, FTA, DRB TRB, TOV, PTS
Insignificant	Age, FG% X3P%, X2P% eFG%, FT%	FG%, X3P, X3PA X3P%, X2P% eFG%, FT%, BLK	FG%, X3P, X3PA X3P%, X2P% eFG%, FT%	Age, FG%, X3P X3PA, X3P%, X2P% eFG%, FT%

Table 2: SLR

We can see from Table 2 that counting statistics, such as MP or X2P appear as significant while percentage statistics such as eFG% and FT% appear as insignificant. Although we cannot make any explicit generalizations regarding the correlation between salary and performance statistics as a whole, we can hypothesize that an increase in one specific counting statistic is more likely to increase salary than an increase in one specific percentage statistic. This hypothesis makes logical sense as there are players with high percentage statistics but are not paid much because these percentages come from low volume/output. A higher paid player is more likely to take more shots and very likely could have a lower FG% than the bench player.

4.2 SLR Adjusted by Position Analysis

Our next step was to replicate the process of applying SLR models on individual statistics but filtered by position. This would allow us to examine a the relationships between the 26 performance statistics on a positional level. We could determine the significance of an individual performance statistic in correlation with salary.

		2020	2019	2018	2017
C	Sig	GS, MP, FG, FGA, X2PA DRB, TRB, TOV, PTS			
	Insig	FG%, X3P%, X2P% eFG%, FT%	Age, G, FG%, X3P X3PA, X3P%, X2P% eFG%, FT%, BLK, PF	G, FG%, X3P, X3PA X3P%, X2P%, eFG% FT%, STL, BLK	FG%, X3P, X3PA X3P%, X2P%, eFG% FT%, AST, BLK
PF	Sig	FGA	MP	GS, MP, FG, FGA X2P, X2PA, DRB TRB, TOV, PTS	MP, FG, FGA DRB, PTS
	Insig	Age, G, FG%, X3P% X2P%, eFG%, FT% ORB	FG%, X3P%, X2P% eFG%, FT%, ORB BLK	G, FG%, X2P%, eFG%	FG%, X3P%, X2P% eFG%, FT%, ORB
SF	Sig	MP, FG, FGA, X2P X2PA, FT, FTA DRB, AST, TOV, PTS	MP, FG, FGA, PTS		MP, FGA
	Insig	G, FG%, X3P%, X2P% eFG%, ORB, BLK	FG%, X3P%, X2P% eFG%, ORB, BLK FT%, ORB, PF	Age, X3P%, X2P% ORB, DRB, TRB BLK, PF	Age, FG%, X3P% X2P%, eFG%, FT% ORB, BLK, PF
SG	Sig		MP, FG, FGA, PTS	GS	G, MP, FGA, PTS
	Insig	G, X3P%, X2P% eFG%, FT%, ORB	Age, G, FG%, X3P% X2P%, eFG%, FT%	Age, X3P%, eFG% FT%, PF	FG%, X3P%, X2P% eFG%, ORB
PG	Sig	MP, FG, FGA, X2P X2PA, FT, FTA, DRB AST, TOV, PTS	MP, FGA, AST TOV, PTS	GS, MP, FG, FGA FT, AST, TOV, PTS	G, GS, MP, FG, FGA, X2P X2PA, FT, FTA, ORB, DRB TRB, AST, STL, TOV, PTS
	Insig	G, FG%, X3P% X2P%, eFG%, FT%	FG%, X3P%, X2P% eFG%, FT%, BLK	FG%, X3P%, X2P% eFG%, FT%, ORB, BLK	X2P%, BLK

Table 3: SLR filtered by position

The information for the years that is not included in the table below is in the project folder at the link given in Section 6. As seen below in Table 3, there is a distinct change in the results for the statistics filtered by position. There are some positions where there are no significant predictors and other positions where there are many significant predictors, in comparison with the table generated without a position filter. The same hypothesis from the plain SLR model is observed to still hold.

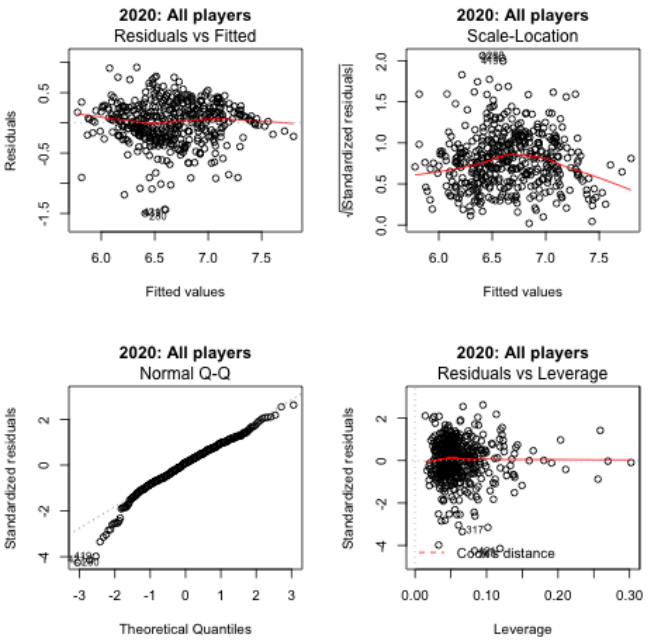
Although using a SLR model allows us to examine specific relationships between a single predictor variable and salary, it is unable to bring the data together to use as much of the data as possible, in conjunction, to predict salary. Thus, we continue our analysis to MLR models.

4.3 MLR Analysis

The analysis for MLR and the following sections are located in analysis.R at the link given in section 6. For each year's performance-salary data, we use an MLR model to predict the performance statistics effect on salary. Since the predictor Pos (Position) is non-numeric, an indicator variable is set up such that each position corresponds to a 1 with the value in all other positions equal to 0.

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.9775632  0.3211697  18.612  < 2e-16 ***
xdataPosC    0.2090367  0.0840500   2.487   0.0133 *
xdataPosPF   0.0310325  0.0617000   0.503   0.6153
xdataPosPG  -0.0623073  0.0667101  -0.934   0.3508
xdataPosSF   0.0132969  0.0574607   0.231   0.8171
xdataPosSG          NA         NA      NA       NA
xdataAge     0.0372862  0.0046436   8.030 9.83e-15 ***
xdataG       0.0009039  0.0012333   0.733   0.4640
xdataGS     -0.0003116  0.0013264  -0.235   0.8144
xdataMP      0.0125069  0.0067348   1.857   0.0640 .
xdataFG     -0.1198357  0.3710355  -0.323   0.7469
xdataFGA    -0.2921204  0.3364097  -0.868   0.3857
xdataFG.     2.0093880  1.1293926   1.779   0.0759 .
xdataX3P     0.3164620  0.3771424   0.839   0.4019
xdataX3PA    0.2641747  0.3336496   0.792   0.4289
xdataX3P.    0.1836445  0.2347707   0.782   0.4345
xdataX2P     0.2437486  0.3711088   0.657   0.5117
xdataX2PA    0.2464347  0.3391355   0.727   0.4678
xdataX2P.   -0.3988619  0.4300368  -0.928   0.3542
xdataeFG.   -3.2456284  1.1605309  -2.797   0.0054 **
xdataFT      0.0230313  0.1062972   0.217   0.8286
xdataFTA    -0.0089683  0.0859166  -0.104   0.9169
xdataFT.    -0.1580361  0.2148292  -0.736   0.4624
xdataORB     0.5200125  0.3699402   1.406   0.1606
xdataDRB     0.5727033  0.3717968   1.540   0.1242
xdataTRB    -0.5632235  0.3708446  -1.519   0.1296
xdataAST     0.0332415  0.0208566   1.594   0.1117
xdataSTL     0.0705080  0.0682148   1.034   0.3019
xdataBLK     0.1250066  0.0606731   2.060   0.0400 *
```

(a) 2020: MLR lm output



(b) 2020: MLR Diagnostic plot

Figure 3: MLR: lm output and diagnostic plot

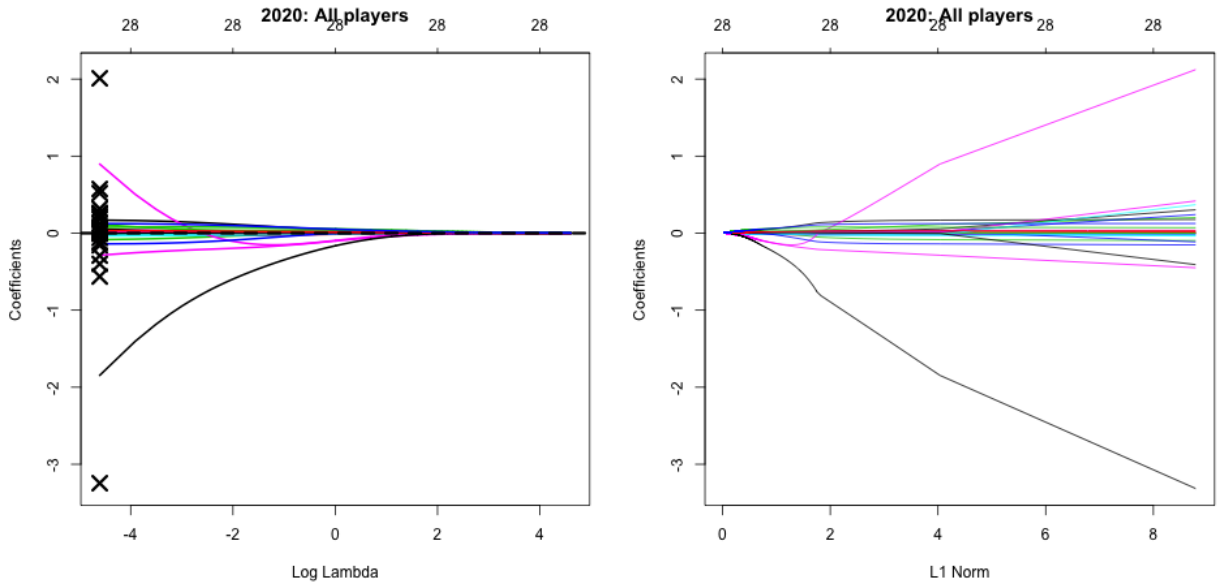
As shown in Figure 3, the predictors that have p-values less than 0.05 are PosC, Age, MP, FG%, eFg% and BLK. This output is interesting because it is contradictory to the hypothesis. Although they are percentage statistics, FG% and eFg% are deemed to be significant. Furthermore, Age, a predictor that was only deemed as insignificant in the SLR model, has the lowest p-value of any predictor. Looking through past years, Age is consistently the predictor with the lowest p-value and many of the significant predictors are percentage statistics. What could have cause this shift in dynamic? We theorize that it is because of the structure of an MLR model. When using SLR, an individual predictor was examined in relation to salary. Now, all predictors are contributing to predicting salary; therefore, the other predictors are helping account for the individual flaws that Age and

percentage statistics had in predicting salary. As mentioned earlier before, we now know and can account for a player shooting less/scoring less to have a high eFG%. Since the predictors are working in tandem, we can now reward a player who has both high PTS and high eFG%.

When examining the MLR Diagnostic plot, we see that the Residuals vs. Fitted is roughly a flat line and that there is no distinctive pattern among the residuals. This indicates that it is unlikely there are non-linear relationships between performance statistics and salary. The Q-Q plot demonstrates that the residuals are normally distributed. The Scale-Location plot indicates that the residuals are about equally spread along the ranges of the predictors. Finally, the Residuals vs. Leverage Plot indicates that since Cook's distance is not even observable on the plot, there are no influential cases. This demonstrates that a linear model is satisfactory and is not problematic for the four plots shown.

4.4 MLR Ridge Regression Analysis

After completing each year's the OLS estimates, we wanted to perform a ridge regression model on the OLS estimates.



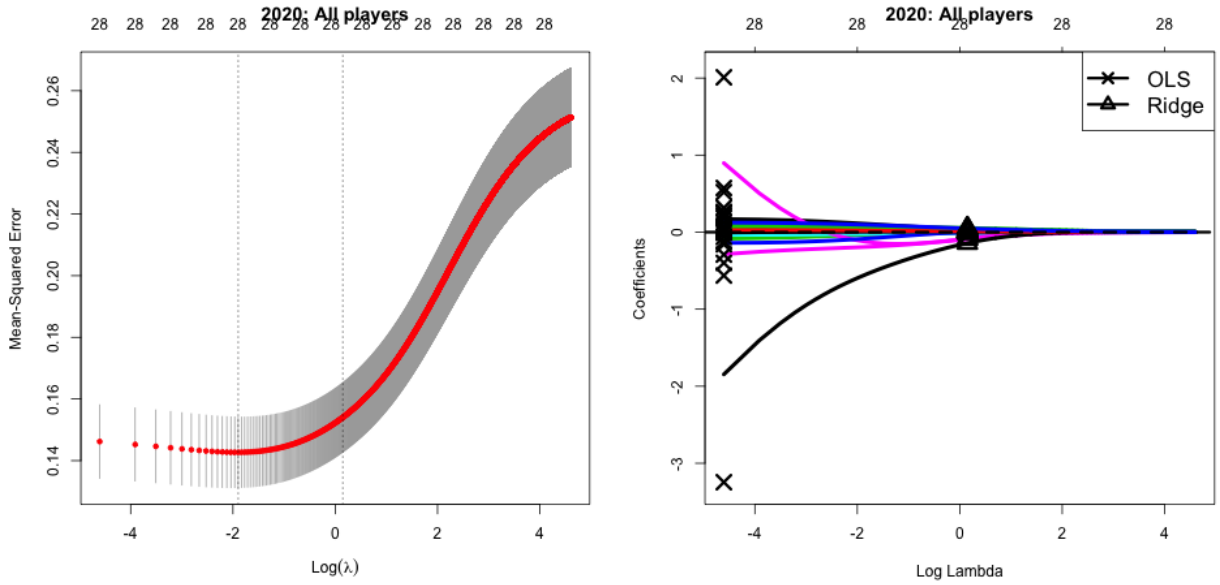
(a) 2020: Ridge regression lambda output

(b) 2020: Ridge regression plot

Figure 4: Ridge regression estimates without optimizing for cross-validation

As shown in Figure 4a, over each iteration of lambda, the predictors shrink towards 0.

At first glance, we see that there are no significant predictors that remain constant. We can also examine the L1 Norm plot, Figure 4b, to view the path of each coefficient against the L1-norm of the whole coefficient vector as lambda iterates.



(a) 2020: Ridge regression lambda output

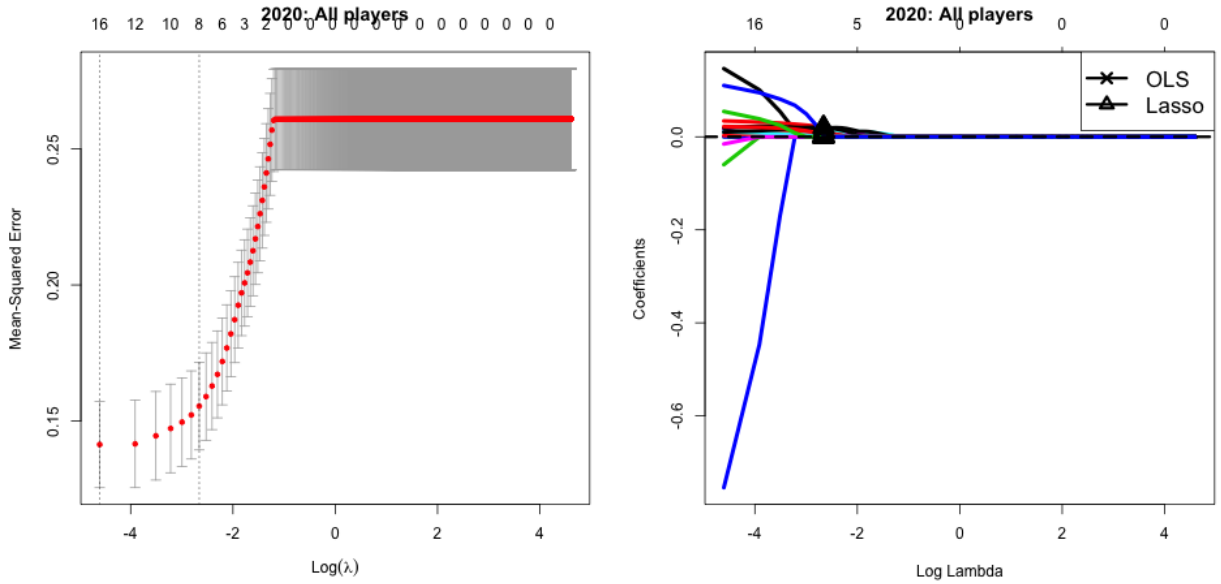
(b) 2020: Ridge regression plot

Figure 5: Ridge regression estimates with optimizing for cross-validation

For computing Figure 5, we picked the lambda value that is the largest within 1 SE of lambda.min. Since a ridge regression model helps minimize multi-collinearity by shrinking the predictor parameters, we can compare the output of the coefficients from the ridge regression model with the output from the OLS model. For the parameters deemed significant by the OLS model, the coefficients for the Age, MP, BLK predictors decreased, and the coefficient for the FG% predictor increased (less negative). This highlights the effect of ridge regression as the coefficients became less extreme accounting for the multi-collinearity between the parameters.

4.5 MLR Lasso Regression Analysis

For our lasso regression analysis, we set up the program as we did for ridge regression, except that we specified for lasso regression. Instead of minimizing the extremity of the coefficients as ridge regression does, lasso regression removes predictors from the model.



(a) 2020: Lasso regression lambda output

(b) 2020: Lasso regression plot

Figure 6: Optimized lasso regression model for the 2020 data.

We can see in Figure 6b that many of the lasso predictors go to zero, but some of them do not. The parameters that do not go to zero are stored in lasso_select.txt in Section 6. For the 2020 season, Age, MP, FGA, X2PA, FTA, DRB, AST, and BLK are the predictors that are non-zero. Interestingly enough, this supports the original hypothesis of having counting statistics be most valuable in predicting salary. Based on the premise of lasso regression, we can determine that this is the simplest model while accounting for the cross-correlation of the parameters.

4.6 MLR GAM Analysis

For the generalized additive model (GAM), we only used the parameters that were obtained from the lasso regression model as there was an error with not having enough unique data points from the 26 predictors. GAM's are useful for adding non-linear components through the use of smooth functions. This accounts for the "curviness" of data.

We see in Figure 7 that there is some non-linearity in the model. Using common sense, we can even justify the reason for the non-linearity. For Age, the initial peak comes from 1-and-done college basketball stars that are usually drafted high first round when they are 18 or 19, resulting in the max rookie contract. The trough comes from college players that

are drafted when they are older but are less likely to be drafter high, thus smaller contracts. Usually, players reach their primes before their 30's, where they get their biggest contracts.

5 Summary and Discussion

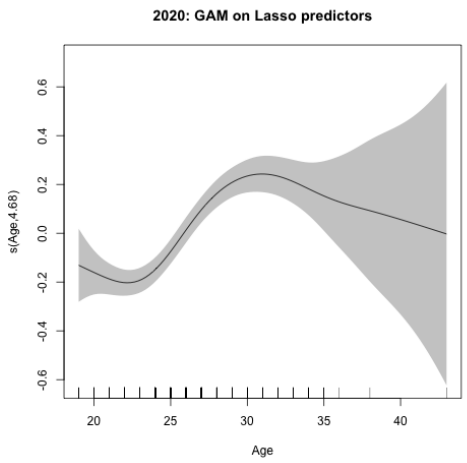


Figure 7: 2020: GAM on Age

The main conclusions of my analysis are that there should be further analysis on the effects of counting statistics vs. percentage statistics. Although there is an inherent correlation between them, it could be useful to split the predictors into these two separate categories (for the ones that are directly related, e.g. X2PA and X2P%) and run OLS, ridge, and lasso models to see if the adjustments would result in observing what we saw here. It was interesting that the ridge regression model favored percentage statistics more so than the lasso regression model did.

We had trouble setting up F-tests and partial F-tests for these models, but there is a necessity to determine whether or not the coefficients and predictors selected are significant, beyond p-values. Another limitation is the amount of data we are working with. Ideally, I would have wanted to examine the entirety of the 2000's. Moreover, I initially wanted to filter the salaries by the type of contracts, however, there were not enough unique data values to justify splitting the data by contract type.

Although the examples presented throughout this report have been centered on the 2020 season data, all analysis examined here have been applied to all years of data. We specifically tried to avoid comparing one year of data to another (i.e. change over time) because the salary cap has changed over time, affecting how contracts are given.

A future direction would be to succinctly quantify which contracts are good vs. bad, as well as examine which predictors could be more likely to result in a bad contract. This direction could transform this project from an analysis-focused project to a prediction-based project that could predict the future contracts of players.

Overall, I had a very fun experience with this project!

References

[1] Basketball Reference. Glossary.

[2] Basketball Reference. NBA ABA League Index.https://www.basketball-reference.com/league

[3] National Basketball Association. 2019-20 NBA Player Contracts. https://www.basketball-refere

[4] Wayback Machine. Wayback machine.

[5] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis, Fifth Edition.* Wiley, Reading, Massachusetts, 2012.

[6] S. J. Sheather. *A Modern Approach to Regression with R.* Springer, New York, 2009.

6 Code and Data

The code and data can be found in the zip file attached with this project or at `https://github.com/awx1/stat410-final`.

The structure of the stat410-final directory is as follows:
There is a folder for every year. All folders named as years are of the same structure Within the folder. Each year folder contains 26 folders, each with two plots for the SLR models. Each year folder also contains 5 position folders, each containing 26 folders for the positional SLR models. Each year folder also contains a MLR folder where the MLR, Ridge, Lasso, and GAM outputs are stored. There is a salary folder and a stats folder. There is a salary_stats folder than combines the two datasets. Within the stat410-final directory, outside of the inner folders are the code files and other miscellaneous generated data.

Note: The code is functionalized and can be run by executing the statements at the bottom of the page.