Can an NBA player's main position and in-game performance statistics predict the player's salary?

Chenxu (Robin) Mao, Yakun Wang, Hugo Feng

December 11, 2023

Introduction

NBA teams often hire professionals to quantify the relationship between player salary, in-game performance, and wins. Quantifying the relationship between these variables enables teams to make objective decisions regarding player contracts, player salaries, and free-agent signings. Similar to the sports economists and statisticians NBA teams hire, we will quantify the relationship between player salary, in-game performance, and main position through linear regression. However, the main objective of our research is not to quantify the relationship between the variables (player salary, in-game performance, and player position) but to develop a model capable of predicting an NBA player's salary based on the player's position and in-game performance.

The relationship between player salary and in-game performance has been extensively researched and examined in much of the modern sports economics literature. In Determinants of NBA Player Salaries, the authors examined the relationship between a player's on-court data and the player's salary. The authors found that points per game and field goal percentage were the two main determinants of NBA player salaries (Lyons et al., 2018)[(2)]. The authors also found that other in-game statistics such as rebounds, assists, and fouls were statistically significant contributors to player salary (Lyons et al., 2018). In NBA Players' Pay and Performance: What Counts?, the researchers examined variables related to an NBA player's salary such as points, rebounds, and three-point attempts. The researchers found that points, assists, and rebounds were statistically significant, and three-pointers made and PER(player efficiency rating) were statistically insignificant (Sigler & Compton, 2018)[(4)]. In Pay Discrimination in the NBA Revisited, Hill explored a previous study on pay discrimination in the NBA by comparing the statistics of different players (Hill, 2004). Hill quantified the coefficients for explanatory variables such as rebounds, assists, and blocks in his OLS model (Hill, 2004) [(3)].

The background literature above focuses primarily on determining the statistical significance of predictor variables and uncovering the main determinants of player salary. Our analysis and methods will differ from the literature since our main objective is to develop a predictive model based on NBA player salary, player position, and in-game performance.

Methods

The dataset contains in-game statistics and positions for NBA players between 2016 and 2019 and has 1409 observations. After importing the dataset into R, we will check variable types in the dataset and select multiple numerical and categorical variables to fit a Multiple Linear Regression (MLR) model. We will then clean the dataset by removing observations with missing entries and mutating variables with incorrect data types or values. After cleaning the dataset, we will conduct an Exploratory Data Analysis (EDA) on the cleaned dataset. For each numerical variable, we will generate a histogram. After concluding there is no skewness in the histograms, we will randomly split the cleaned dataset into a training set (60%) and a test set (40%). We will use the training dataset to fit an MLR model, generate scatterplots of response variable vs. fitted values, and produce all pairwise scatterplots of predictors. Using the scatterplots, we will check the two additional conditions for MLR: conditional mean response condition and conditional mean predictor condition. If one/both of these two additional conditions have been violated, we will use variable transformation, interaction terms, or residual analysis to correct it.

We will move on to checking for normality assumption, constant variance assumption, linearity assumption, and uncorrelated errors assumption. We will generate a QQ plot to check the normality assumption. A stark deviation in the QQ plot indicates that the normality condition is violated. If the normality assumption is violated, we will use residual transformations (Box-Cox) to correct it. We will generate a scatterplot of residuals vs. fitted values to check for constant variance assumption, linearity assumption, and uncorrelated errors assumption. A systematic pattern display in the scatterplot will indicate the violation of one or more of the assumptions. If the constant variance assumption is violated, we will perform a variance stabilizing transformation. If the linearity assumption is violated, we will apply linearity transformation. If the uncorrelated errors assumption is violated, we will apply the Weighted Least Squares method.

After checking for the regression assumptions and correcting any violations, we will remove the predictors that are statistically insignificant according to the MLR. We will continue performing MLR, checking regression assumptions/conditions, and removing statistically insignificant predictors until we find a reduced model where all the predictors are significant. After arriving at the reduced model based on the training set, we will perform an overall F test, and find the outliers, high-leverage points, and influential points. We will repeat the whole process again but instead with the test dataset. We will validate the reduced model based on the training set by comparing it to the reduced model based on the test set. Criteria for model validation comparison include minimal differences in estimated coefficients, the same number of predictors, similar adjusted R-squared, similar numbers and types of problematic observations, and a similar amount of multicollinearity. Through comparing these values, we can conclude if our final model is valid.

Results

Numerical Summaries:

Training /Test	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
FT.	0.0000 /0.0000	0.6890 /0.6677	0.7690 /0.7600	0.7463 /0.7378	0.8330 /0.8197	1.0000 /1.0000
FG.	0.1000 /0.0000	0.4050 /0.4052	0.4450 /0.4440	0.4485 /0.4518	0.4830 /0.4930	0.7310 /0.7500
TRB	0.200 /0.000	2.100 /2.000	3.300 /3.200	3.843 /3.791	5.100 /4.800	15.200 /16.000
AST	0.000 /0.000	0.800 /0.800	1.500 /1.300	2.095 /1.913	2.700 /2.400	11.200 /10.700
BLK	0.0000 /0.0000	0.1000 /0.1000	0.3000 /0.3000	0.4152 /0.4247	0.5000 /0.5000	2.4000 /2.7000
TOV	0.000 /0.000	0.600 /0.600	1.000 /0.900	1.220 /1.131	1.600 /1.500	5.700 /5.000
PF	0.000 /0.000	1.300 /1.300	1.800 /1.800	1.759 /1.789	2.300 /2.300	3.900 /3.800
PTS	0.400 /0.500	4.700 /4.600	7.700 /7.400	9.225 /8.994	12.700 /12.175	31.600 /36.100
Salary	77250 /56845	1505942 /1512601	4187599 /3362124	7242215 /6750262	11709603 /10000000	33285709 /37457154

Figure 1: Numerical Summaries of Variable in Both Training Set and Test Set

The chart above summarizes the statistics for all numerical predictors and the response among the training and test datasets. For the numerical predictors, we observe a balance among the statistics between the training and test datasets. However, we observe some statistical differences in the response variable. For example, there is a significant difference in the mean. This can be reasoned with the great variation in the original dataset, which will possibly have the differences (as above) even if the split was random.

Overall, we conclude that the training and test datasets are similar.

Exploratory Data Analysis (EDA):

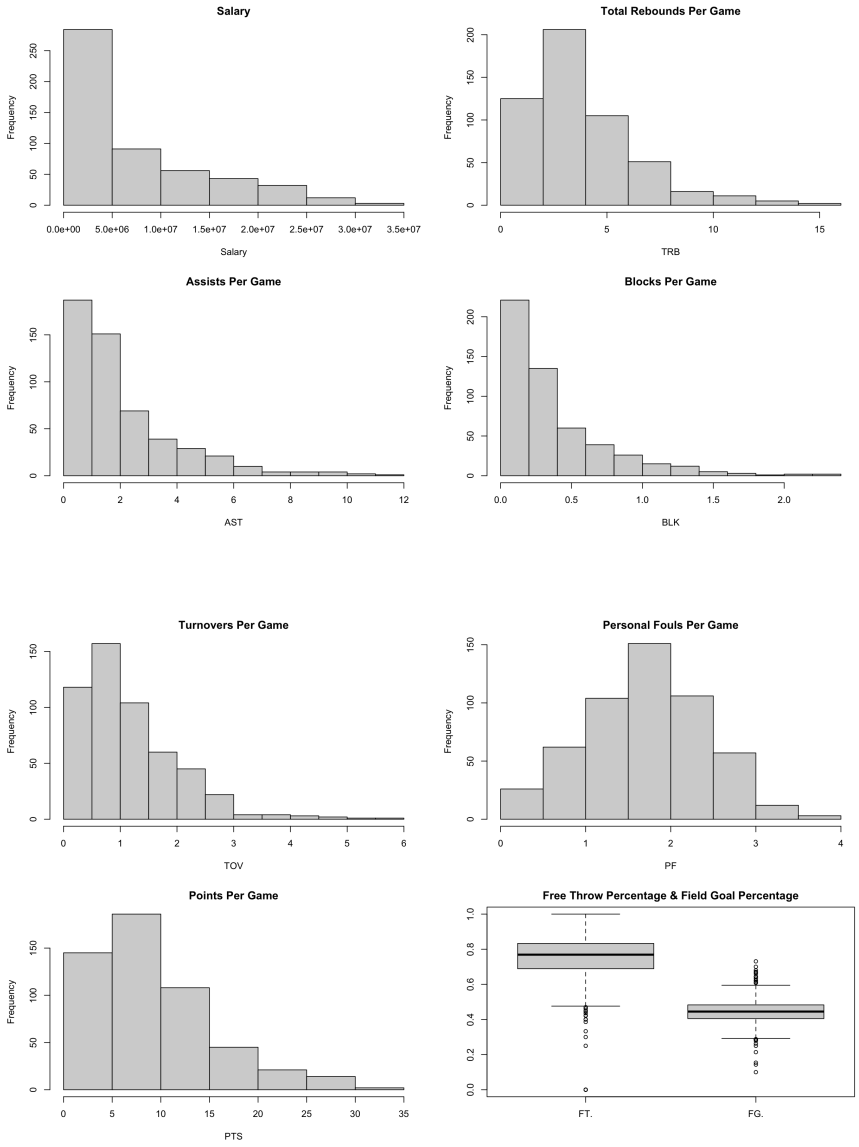

Figure 2: Histograms & Boxplot for Numerical Variables in Training

Figure 2 above includes the histograms and the boxplot for the predictors and the response. The Salary histogram seems to have a noticeable right-skewness, which contradicts the normality assumption or the linearity assumption. We also observe several skewnesses in the histogram for TRB, AST, BLK, TOV, and PTS, which also contradicts the two assumptions as before. However, the boxplot for $FG.$ and the PF histogram presents no issue.

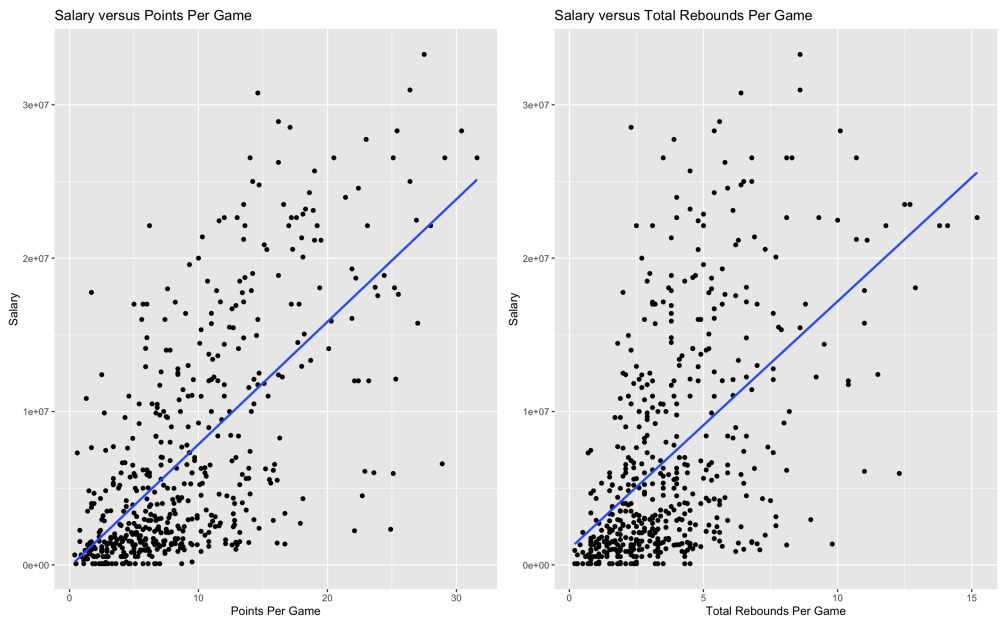

Figure 3: Scatterplots of Salary versus PTS or TRB

Then, we generate the scatterplots for PTS and TRB. The scatterplots show no significant violation due to the presence of a linear relationship between salary the the predictors. However, the spread is uneven and has a concentration in the left-bottom corner, which could indicate violations of the normality or the linearity assumption. This is consistent with the result from the histograms.

Model Fitting:

To begin with, we fit a Multiple Linear Regression Model (MLR) with the training set, denoted Model 1: Salary $\sim FT. + FG. + TRB + AST + BLK + TOV + PF + PTS + Pos1$.

Note that for simplicity, we modify the categorical predictor $Pos1$ to the following as stated in the proposal,

$$Pos1 = \begin{cases} 1 & \text{if Point Guard} \\ 2 & \text{if Power Forward} \\ 3 & \text{if Centre} \\ 4 & \text{if Shooting Guard} \\ 5 & \text{if Small Forward} \end{cases}$$

The five positions are listed by order of their average earning in the NBA, with $Pos1 = 5$ being the highest. (Spotrac., 2023)[(1)]

Then, we conduct model-checking to examine the Conditional Mean Response Condition (Condition 1) and the Conditional Mean Predictor Condition (Condition 2). To do that, we need to examine the scatterplot of the salary (response variable) versus the fitted values. (see Figure 4) and the pairwise scatterplots of predictors from Model 1 (see Figure 5).

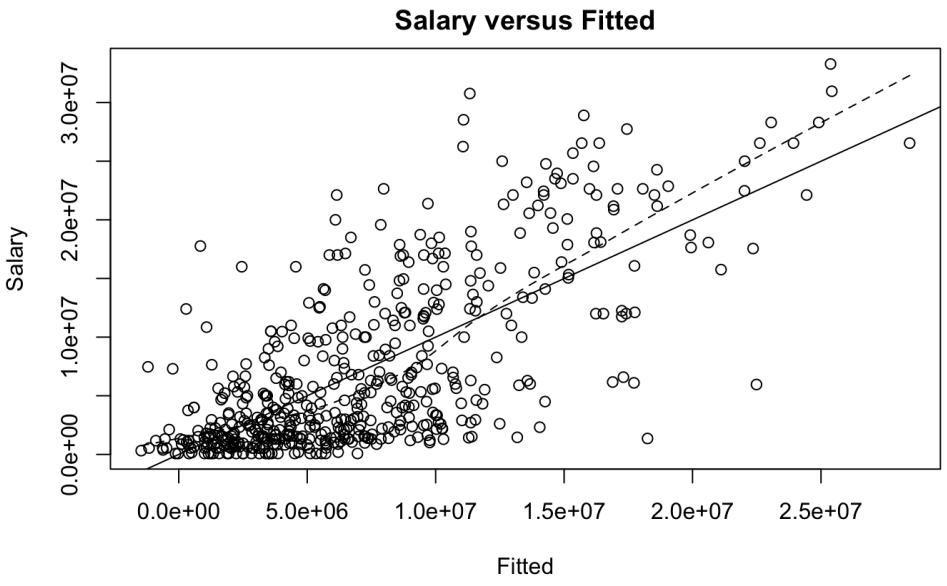

Figure 4: Scatterplot of Salary versus Fitted Values

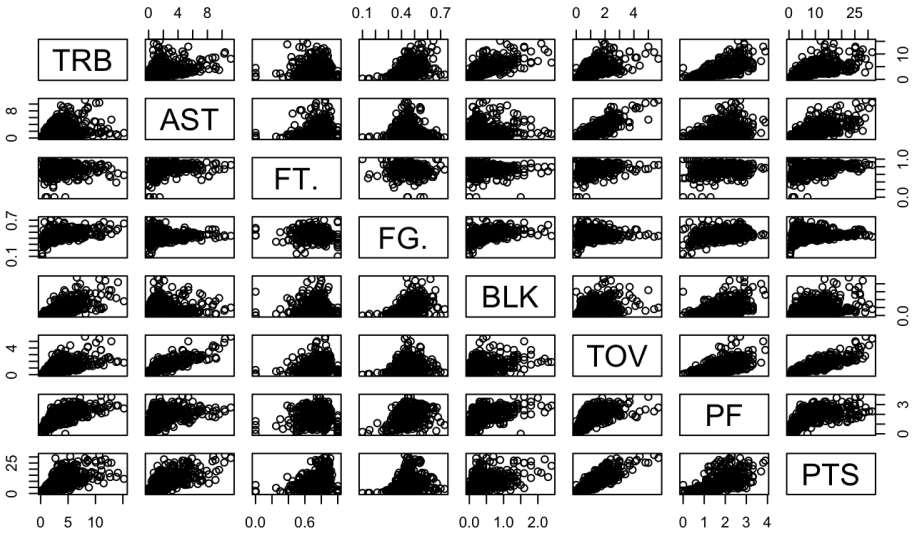

Figure 5: Pairwise Scatterplots of Predictors in Model

Figure 4 shows no identifiable pattern/function, thus Condition 1 does not hold. In Figure 5, there is no presence of any non-linear relation between the predictors, thus Condition 2 holds.

Therefore, we need to do a Box-Cox transformation on the response. As a result of the log-likelihood plot (shown in the rmd), we apply a fourth root transformation to the response and generate a new MLR model, denoted Model 2, where Transformed Salary $\sim$

$FT. + FG. + TRB + AST + BLK + TOV + PF + PTS + Pos1$. Again, we need to conduct a check on Model 2.

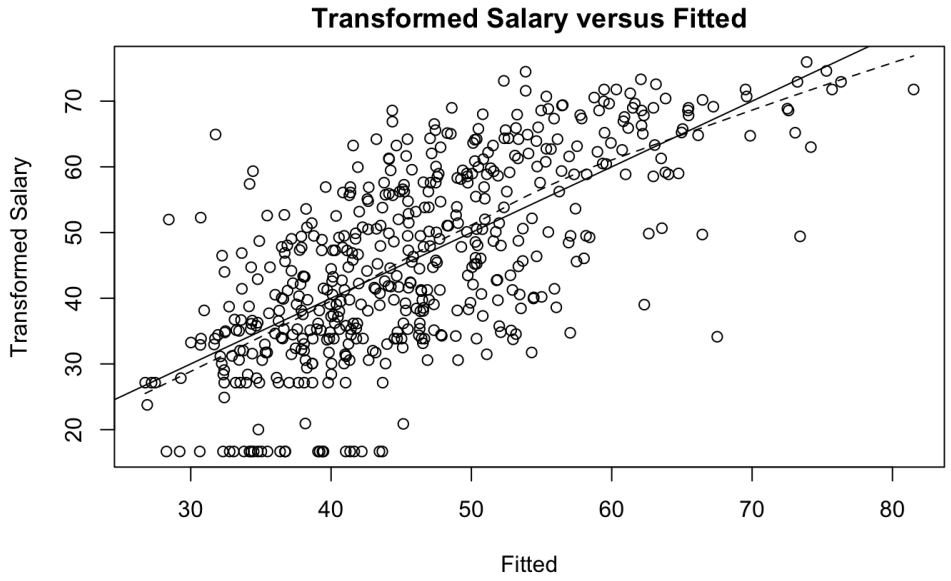

Figure 6: Scatterplot of Transformed Salary versus Fitted Values

Figure 6 shows a relatively symmetrical pattern around the fitted line, which is an identifiable pattern/function that makes Condition 1 hold. As before, Figure 5 ensures that Condition 2 holds. Then, we begin checking model assumptions by examining the scatterplot of residuals vs fitted values and the QQ plot.

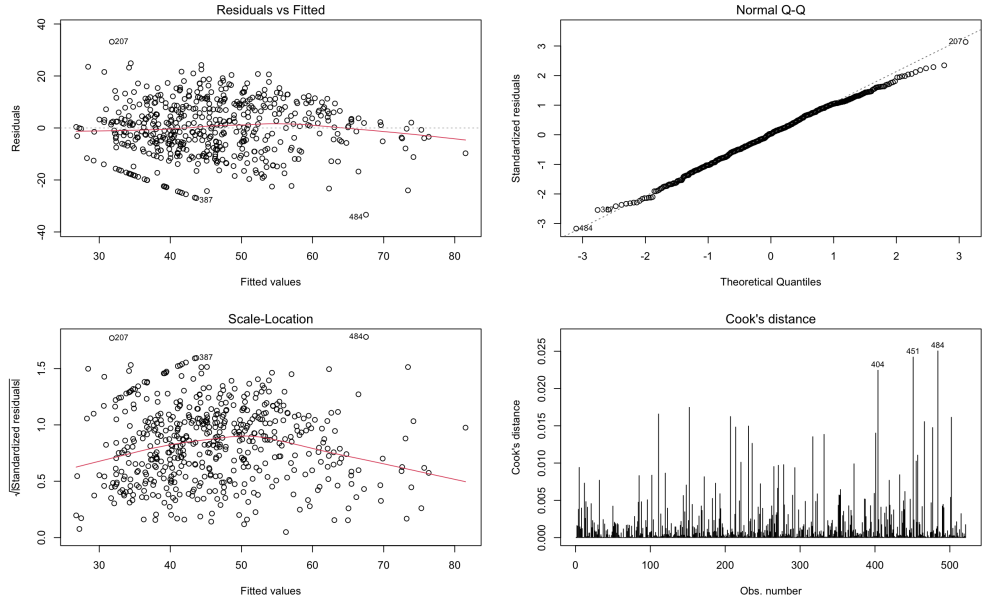

Figure 7: Fitted Models of Training Set and Test Set

Figure 7 presents the scatterplots, the QQ plot, and the Cook's Distance plot. As the scatterplot of residuals versus fitted values has no systematic pattern such as curves or fanning patterns, and with no clustering, all linearity assumption, constant variance assumption, and uncorrelated error assumptions are met. The QQ plot of the standardized residuals shows no stark deviation nor curving from the diagonal line, thus the normality assumption is met.

Now we have all the model conditions and assumptions checked. We begin to reduce Model 2 by removing insignificant predictors to help with the interpretation.

Coefficients	Estimate	Standard Error	t value	Pr (> |t|)
Intercept	21.1521	4.6818	4.518	7.76×10^{-6}
FT.	7.6374	3.8904	1.963	0.05017
FG.	-2.7614	7.3398	-0.376	0.70691
TRB	1.8678	0.3428	5.449	7.87×10^{-8}
AST	2.1329	0.5370	3.972	8.16×10^{-5}
BLK	-0.7280	1.7130	-0.425	0.67103
TOV	-3.2672	1.6290	-2.006	0.04542
PF	1.3384	0.9941	1.346	0.17880
PTS	0.8475	0.1616	5.243	2.32×10^{-7}
Pos1	0.9882	0.3672	2.691	0.00735

Figure 8: Model 2 Coefficient Statistics

Figure 8 above shows the predictor statistics. Define significant level $\alpha = 0.05$, which makes predictors with a p-value greater than α to be insignificant. Significant otherwise. From the table, we observe $FT., FG., BLK$, and PF to be insignificant, thus we remove them from Model 2 and fit a new model with only the significant predictors, denoted as Model 3: Transformed Salary $\sim TRB + AST + TOV + PTS + Pos1$.

We have to check the predictor significance and delete insignificant coefficients again in Model 3. Then, we conduct an overall F test on models to test whether there is a linear association between the response variable and all predictors. Our final model is denoted Model 4: Transformed Salary $\sim TRB + AST + PTS + Pos1$, which in detail, is

$$\begin{aligned}\text{Transformed Salary} = 26.8177 &+ 1.5845[\text{Total Rebounds Per Game}] \\ &+ 1.3794[\text{Assists Per Game}] + 0.8071[\text{Points Per Game}] \\ &+ 0.9553[\text{Main Position}]\end{aligned}$$

Where Transformed Salary $= \text{Salary}^{\frac{1}{4}}$.

Finally, we search for problematic observations, such as leverage points, outliers, or influential observations in the final model. We also examine multicollinearity to check for potential linear trends among the predictors (see Figure 10 in Appendix).

Model Validation:

To validate the final model generated by the training set, we use the test set to examine whether our final model overfits.

Same as the methods applied on the training set, we fit an MLR model on the test set with the same preliminary model, denoted as Model 1: Salary $\sim FT. + FG. + TRB + AST + BLK + TOV + PF + PTS + Pos1$. We then check the model conditions and model assumptions of Model 1. Conduct a transformation on the response variable if any conditions or assumptions are violated. In this case, the linearity assumption is violated. We conduct a Box-Cox transformation on Salary same as before. Then, we fit a new model with the transformed Salary as the response. Model 2: Transformed Salary $\sim FT. + FG. + TRB + AST + BLK + TOV + PF + PTS + Pos1$. Again, we check the model conditions and assumptions of Model 2. We remove insignificant predictors and fit new models until we have a model with only significant predictors. The final model can be denoted as Model 3: Transformed Salary $\sim TRB + AST + PTS$. More specifically,

$$\begin{aligned}\text{Transformed Salary} = 29.9090 &+ 1.4692[\text{Total Rebounds Per Game}] \\ &+ 1.2910[\text{Assists Per Game}] + 0.7970[\text{Points Per Game}]\end{aligned}$$

To check model validation, we compare the final models generated by both the training set and test set. Criteria include minimal differences ($< 2s.e.'s$) in estimated coefficients, the same number of predictors, similar R^2_{adj}, similar numbers and types of problematic observations, and a similar amount of multicollinearity. Through comparing these values, we can conclude that our final model is valid (see Figures 10 & 11 in the Appendix).

Training Set	
Model 1	Salary ~ FT. + FG. + TRB + AST + BLK + TOV + PF + PTS + Pos1
Model 2	Transformed Salary ~ FT. + FG. + TRB + AST + BLK + TOV + PF + PTS + Pos1
Model 3	Transformed Salary ~ TRB + AST + TOV + PTS + Pos1
Model 4	Transformed Salary ~ TRB + AST + PTS + Pos1
Test Set	
Model 1	Salary ~ FT. + FG. + TRB + AST + BLK + TOV + PF + PTS + Pos1
Model 2	Transformed Salary ~ FT. + FG. + TRB + AST + BLK + TOV + PF + PTS + Pos1
Model 3	Transformed Salary ~ TRB + AST + PTS

Figure 9: Fitted Models of Training Set and Test Set

Discussion

Conclusion:

As stated in the result section, our final model states the relation between an NBA player's salary and the performance variables as the following,

$$\begin{aligned}\text{Transformed Salary} = 26.8177 &+ 1.5845[\text{Total Rebounds Per Game}] \\ &+ 1.3794[\text{Assists Per Game}] + 0.8071[\text{Points Per Game}] \\ &+ 0.9553[\text{Main Position}]\end{aligned}$$

Where Transformed Salary $= \text{Salary}^{\frac{1}{4}}$.

We can explain the model as:

On average, the fourth square root of an NBA player's salary equals 27.773 when the player is a Point Guard with no rebound, no assist, and no points. The fourth salary square root increases by 1.5845 for each increase in total rebound per game, increases by 1.3794 for each increase in assist per game and increases by 0.8071 for each increase in points per game. The fourth salary square root of the player increases by another 0.9553 once the player becomes a power forward from a point guard, and increases by another 0.9553 once the player becomes a centre from a power forward. The play will have this increase upon any ordered change in its main position until it becomes a small forward from a shooting guard.

This concludes that an NBA player's main position and in-game performance statistics can predict the player's salary as the answer to the research question.

Compared to others' works, for example, Determinants of NBA Player Salaries (The Sport Journal, 2015)[(2)], this paper provides a more elegant model trained by more recent data (up to the 2018-19 season, instead of where the former used only the data from 2013-2014). However, the former final model yields a similar result with similar predictors. This paper's model has a stronger focus on the optimal model complexity rather than the prediction outcome than other works such as Discrimination in the NBA Revisited (Hill, 2004)[(3)].

Limitations:

This study contains potential limitations as we transformed the $Pos1$ variable into a continuous variable during the fitting section which could cause bias in the final model for the coverage of more potential positions not presented in the dataset. Moreover, although we concluded that there exists no statistically significant violation to the constant variance assumption, there does exist somewhat violation that we chose to ignore based on its magnitude. The predictor number of the final model created by the training set is different from the test set final model, which could also cause bias.

Appendix

1. Spotrac. (2023, October). NBA Player Earnings. https://www.spotrac.com/nba/positional/

2. Lyons Jr, R., Jackson Jr, E. N., & Livingston, A. (2015). Determinants of NBA Player Salaries. The Sport Journal. https://doi.org/10.17682/sportjournal/2015.019

3. Hill, J. R. (2004). Pay Discrimination in the NBA Revisited. Quarterly Journal of Business and Economics, 43(1/2), 81–92. http://www.jstor.org/stable/40473375

4. Sigler, K., & Compton, W. (2018). NBA Players' Pay and Performance: What Counts? The Sport Journal. https://thesportjournal.org/article/nba-players-pay-and-performance-what-counts/

	Model of Training Set	Model of Test Set
Outlier		
Leverage points	32 37 38 69 76 85 87 88 90 98 112 115 134 152 180 189 192 194 195 222 231 234 290 291 332 342 352 371 385 404 413 415 417 428 440 441 466 472 474 481 498 499 501 503 517 518	31 33 66 74 113 137 140 357 372 401 404 431 457 469 489 522 540 552 555 606 615 665 685 703 704 747
Influential on all fitted values		
Influential on own fitted values	32 85 88 152 221 231 236 270 290 317 323 332 352 401 404 440 455 457 466 475 484	32 85 88 152 221 231 236 270 290 317 323 332 352 401 404 440 455 457 466 475 484
Influential on at least one estimated coefficient	"Beta 0" 4 12 50 96 102 104 111 152 193 207 221 247 265 271 290 354 360 368 372 401 419 422 425 439 447 451 455 457 463 475 477 484 "Beta 1" 37 85 152 178 185 187 192 231 244 290 307 332 352 365 399 404 415 433 436 440 455 475 484 501 "Beta 2" 14 16 21 25 32 50 88 112 120 134 146 152 189 192 221 246 261 262 279 291 323 332 340 401 408 413 425 440 447 463 466 477 484 490 "Beta 3" 14 21 37 50 85 88 120 152 169 178 192 207 231 262 266 278 279 307 323 332 352 401 404 413 440 455 466 484 "Beta 4" 5 12 14 25 39 50 85 88 94 102 214 219 221 236 262 263 265 270 271 279 284 317 321 323 351 353 368 422 425 447 451 455 457 463 466 475 483	"Beta 0" 51 113 218 219 229 246 250 262 320 334 372 381 384 408 425 459 527 537 553 566 567 582 601 631 705 736 755 "Beta 1" 7 21 31 77 95 113 125 135 155 174 184 195 204 219 246 306 317 372 431 444 451 452 457 461 500 532 549 553 582 668 679 726 761 "Beta 2" 36 135 240 246 276 304 358 372 401 428 457 469 477 498 540 553 606 747 "Beta 3" 7 16 21 33 84 132 174 222 246 275 358 360 372 388 401 403 444 452 457 469 474 477 498 540 599 605 606 679 724 739 761
Variance inflation factor	TRB: 1.673402 AST: 2.264794 PTS: 3.009965 Pos1: 1.168712	TRB: 1.663052 AST: 1.990210 PTS: 2.884788

Figure 10: Model Comparison by Problematic Observations and Variance Inflation Factor

Standard Error:	Training Model/Test Model
Intercept	1.4802/0.7780
TRB	0.2430/0.2032
AST	0.3668/0.3248
PTS	0.1318/0.1114
Pos1	0.3660/NULL
R^2_{adj}	0.4507/0.4018

Figure 11: Model Comparison by Standard Error and Adjusted R-squared