Predicting 2022-2023 NBA Player Salaries

Group 4: Emma Aller Vidal[a,b,c], Allison Lynn[a,c], Olivia Motmans[a,c], Dillon Maheshwari[a,c], Jaelyn Watson[a,b,c], Jesper White[a,c,d]

[a]University of California, Los Angeles (UCLA); [b]Department of Mathematics; [c]Department of Statistics and Data Science; [d]Department of Economics

W24-LEC2-STATS 101A: Introduction to Data Analysis and Regression | Dr. Maria Cha | TA: Navin Souda

1. Introduction

In this project, we develop a multiple linear regression model to predict the salaries of NBA players during the 2022-2023 season. Specifically, we will investigate the role of age, total rebounds, points per game, and value over replacement in determining the salary of an NBA player.

1.1. Background and Dataset

The dataset used is from Kaggle [4] and measures all 466 NBA player salaries (in USD) from the 2022-2023 season along with 50 other categorical and quantitative variables. The categorical variables are composed of player name, position, and team, while the quantitative variables consist of age and player performance metrics, both offensively and defensively.

The predictor variables chosen are Age, PTS, VORP, and TRB because of their statistically significant slope coefficients when compared to the other quantitative variables in the dataset. **Age** is a player's age during the start of the 2022-2023 season, **Points (PTS)** are a player's average number of points per game, **Total Rebounds Per Game (TRB)** are a player's average offensive and defensive rebounds per game, and **Value Over Replacement Player (VORP)**, is calculated using [BPM – (-2.0)] * % of minutes played * games played/82, where BPM is the change of the score when a player is on the court.

1.2. Methods

We use R and the `car`, `leaps`, and `MASS` library packages. We start by fitting a full multiple linear regression model using Age, PTS, VORP, and TRB as predictors, and Salary as the response variable. Then, we check model assumptions (linearity, normality, and constant variance of the error terms) using standardized residuals and diagnostic plots. We also explore the interesting outliers in our data. We transform our model and check assumptions again to verify an improvement.

In an attempt to simplify our model and improve interpretability, we try variable subsetting by considering all possible subsets and performing stepwise regression. We analyze the resulting models and once again confirm model assumptions.

Lastly, to choose our final model, we calculate VIFs to discover potential multicollinearity issues, we use goodness-of-fit measures (adjusted R-squared, AIC, AICc, BIC), and Added-Variable plots. We then connect our research to other studies about NBA salaries.

We conclude our report with a discussion of the limitations of our findings and recommendations for possible improvements in future research.

2. Data Description

	Salary (USD)	Age	PTS	VORP	TRB
Mean	8,416,599	25.82	9.13	0.54	3.53
SD	10,708,118	4.28	6.91	1.17	2.28

Table 1. *Summary Statistics for each variable*

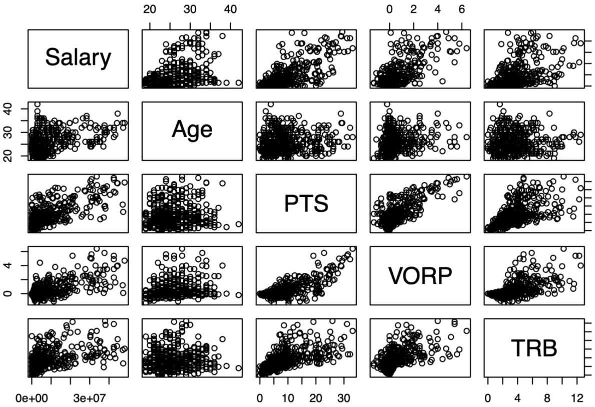
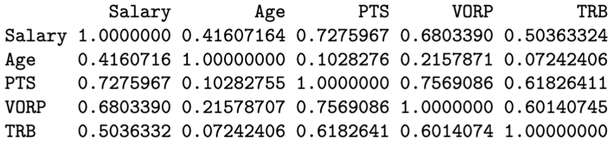


Figure 1. *Scatterplot Matrix*



```
          Salary        Age       PTS      VORP        TRB
Salary 1.0000000 0.41607164 0.7275967 0.6803390 0.50363324
Age    0.4160716 1.00000000 0.1028276 0.2157871 0.07242406
PTS    0.7275967 0.10282755 1.0000000 0.7569086 0.61826411
VORP   0.6803390 0.21578707 0.7569086 1.0000000 0.60140745
TRB    0.5036332 0.07242406 0.6182641 0.6014074 1.00000000
```

Figure 2. *Correlation Matrix*

None of the distributions for each variable follow a normal distribution (histograms can be found in the R markdown file). Every variable is skewed right, with a prominent skew in the response variable, Salary, which takes on a range of USD 5,849 to USD 48 million, has a high standard deviation, (Table 1) and a striking difference between the mean and median (for full summary statistics, please see the R markdown file). Salary is unequally distributed amongst the players, which can explain issues relating to non-constant variance. Minimum salaries near USD 6,000 are seen in players on a 10-day contract and minimum full-season contracts are seen in bench players (the average bench player makes USD 1.5 million [5], barely over the minimum USD 1.1 [6] million full-season contract). However, suppose a player can make the jump into a starter role. In that case, their average salary can go up to USD 8.5 million [6] and can reach figures up to the maximum contract of USD 48 million.

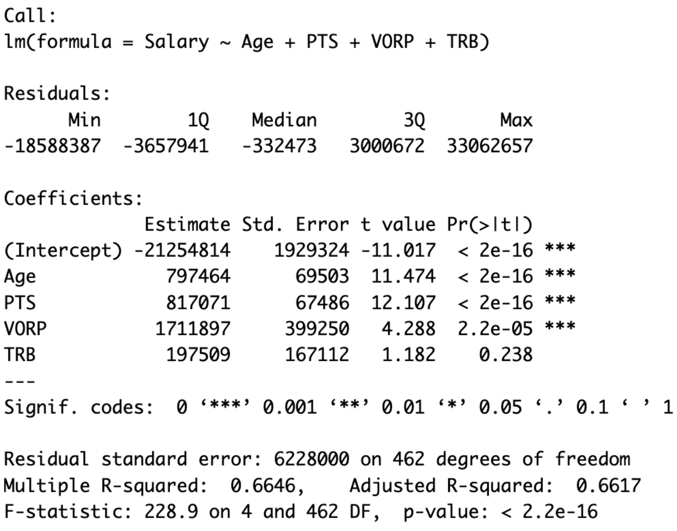
The predictor variable Age has the smallest variance out of our predictor variables but still follows a right-skewed distribution. This is because one must be at least 19 to be in the NBA and once players get past their years of peak body development (ages 23-27), it becomes increasingly hard to stay in the NBA. There are still some outliers to this general rule, which we will explore later. The longer a player can stay in the NBA, the greater the compensation in their contracts, so age has a relatively fair positive correlation with salary.

On the other hand, the predictor variables Points, TRB, and VORP appear to have positive correlations with one another (Figures 1 and 2) since they all measure player performance. While none of the correlations exceed 0.8, as shown in Figure 2, we will still check for multicollinearity using VIF.

3. Results and Interpretation

3.1. Model 0: Full MLR

$$\widehat{Salary} = -21254814 + 797464Age + 817071PTS + 1711897VORP + 197509TRB \quad (1)$$



```
Call:
lm(formula = Salary ~ Age + PTS + VORP + TRB)

Residuals:
      Min        1Q    Median        3Q       Max 
-18588387  -3657941   -332473   3000672  33062657 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -21254814    1929324 -11.017  < 2e-16 ***
Age            797464      69503  11.474  < 2e-16 ***
PTS            817071      67486  12.107  < 2e-16 ***
VORP          1711897     399250   4.288  2.2e-05 ***
TRB            197509     167112   1.182    0.238    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6228000 on 462 degrees of freedom
Multiple R-squared:  0.6646,	Adjusted R-squared:  0.6617 
F-statistic: 228.9 on 4 and 462 DF,  p-value: < 2.2e-16
```

Figure 3. *R Summary for Model 0*

Figure 3 indicates significant slopes for all predictors but TRB. 66.46% of the variation in Salary is explained by the model. The p-value of the overall F-test also suggests that at least one of the slopes of the predictors is significant. Because this is the full untransformed model, interpretation is very simple: for each unit increase in each predictor, predicted salary increases by the slope of the corresponding predictor.

3.1.1. Exploration of outliers.

In our statistical analysis of NBA player salaries, certain outliers have emerged that do not conform to the expected relationships predicted by our regression model. These outliers can be primarily attributed to the NBA's salary structure, particularly the salary cap, which creates a unique economic environment where player salaries do not always correlate directly with their statistical output. The NBA's salary cap is a pivotal factor that dictates player salaries. Players like LeBron James, Luka Dončić, and Nikola Jokić, who are at the pinnacle of the sport, always command maximum salaries due to their exceptional talent and influence on the game. However, because the salary cap limits the total amount that teams can allocate to player salaries, franchises must strategically balance paying for superstar talent while also securing capable players to fill out the roster.

Fringe All-Stars and elite role players represent a category of players who command significant salaries, not because they match the statistical output of players like LeBron or Jokić, but because their roles are critical to team success in a capped economy. A majority of the outliers were these types of players. Players such as D'Angelo Russell and Kris Middleton may not produce the same level of statistics as the league's superstars, but their ability to contribute significantly to a team's performance makes them highly valued.

For example, the Milwaukee Bucks have a generational talent in Giannis Antetokounmpo, but to construct a championship-caliber team, they also need high-caliber players like Middleton. Despite the discrepancy in statistical output compared to a superstar like Giannis, Middleton's salary is only slightly less due to the limitations imposed by the salary cap. In an uncapped system, such as Major League Baseball, the salary landscape would likely look very different, with a greater disparity in pay that correlates more closely with player statistics.

Other outliers included All-star players who are recovering from injury (Klay Thompson and Kevin Love) who have produced at a high caliber in the past and are still paid significantly, just not having the same output. Additionally, 42-year-old Udonis Haslem represents an outlier as he was able to contribute beyond play statistics, offering intangible assets such as leadership and mentorship. Additionally, Louis King, a two-way player (someone from the minor leagues) who only played in one game, was also an outlier. For a complete set of outliers, and for the process we used to find them, please see the R markdown file.

3.1.2. Model 0 Model Assumptions.

We will now check whether our model assumptions are validated by looking at the Diagnostic plots.

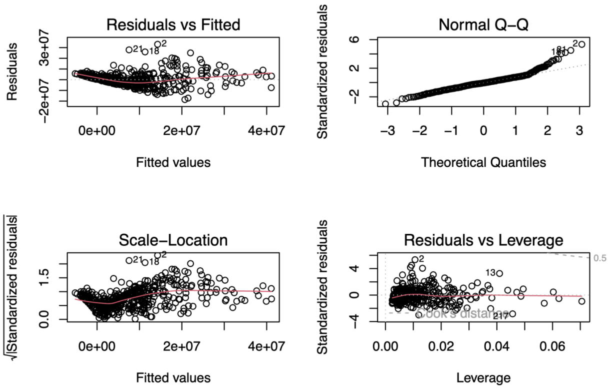


Figure 4. *R Diagnostic Plots for Model 0*

Figure 4 shows that our model assumptions are not completely validated. In the residuals vs. fitted plot as well as the scale-location plot, the points are not randomly scattered showing non-constant variance and the red line is curved showing non-linearity. The Normal QQ plot also shows points that deviate from the general trend which shows there may not be normality of the errors. In the residuals vs. leverage plot, we can see there are potential outliers or influential points. We will try transformations to try to fix the model assumptions of constant variance, linearity, and normality.

3.2. Model 1: Full Transformed MLR

Because our data has negative values, we cannot run a Box-Cox transformation. Therefore, we opted for trying the inverse response method to transform Salary.

$$\widehat{Salary}^{0.71} = -118685.8 + 4993.7Age + 5078.2PTS + 7230VORP + 2333.2TRB \quad (2)$$

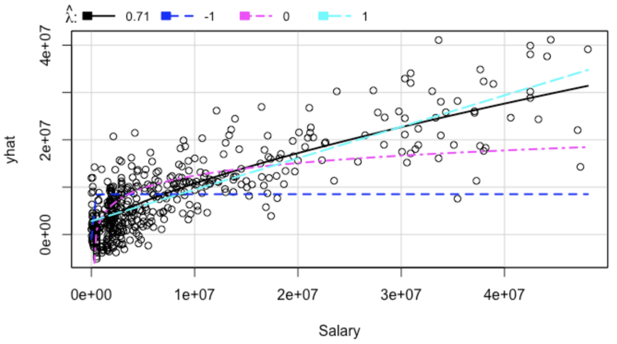
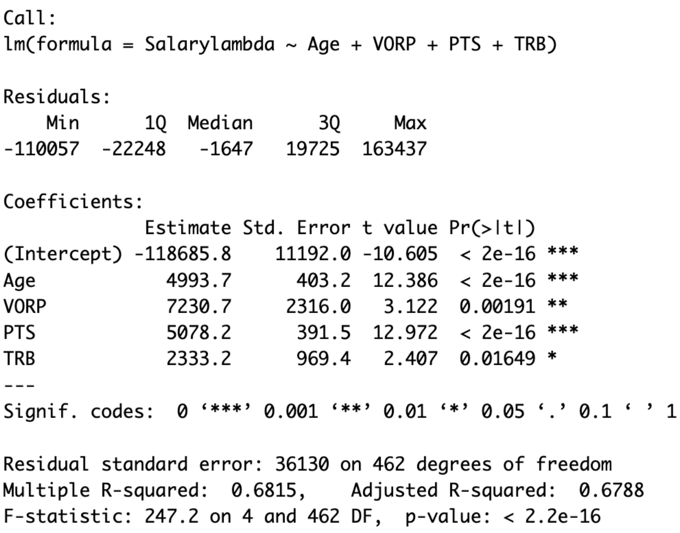


Figure 5. *Inverse Response Plot*



```
Call:
lm(formula = Salarylambda ~ Age + VORP + PTS + TRB)

Residuals:
    Min      1Q  Median      3Q     Max 
-110057  -22248   -1647   19725  163437 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -118685.8    11192.0 -10.605  < 2e-16 ***
Age            4993.7      403.2  12.386  < 2e-16 ***
VORP           7230.7     2316.0   3.122  0.00191 ** 
PTS            5078.2      391.5  12.972  < 2e-16 ***
TRB            2333.2      969.4   2.407  0.01649 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36130 on 462 degrees of freedom
Multiple R-squared:  0.6815,	Adjusted R-squared:  0.6788 
F-statistic: 247.2 on 4 and 462 DF,  p-value: < 2.2e-16
```

Figure 6. *R Summary for Model 1*

After transforming the response variable to the power of 0.71 as suggested by Figure 5, Figure 6 shows that all the predictors have become significant to Salary. The p-value for the overall F-test again shows that at least one of the slopes of the predictors is statistically significant.

Since we have transformed the variable, the interpretation is a bit trickier. For every 1-year age increase, predicted salary transformed to the power of .71 increases by $4993.7 on average.

For every 1-point increase, predicted salary transformed to the power of .71 increases by 5078.2 on average.

For every 1 unit increase in VORP, the predicted salary transformed to the power of .71 increases by 7230 on average.

For every 1 rebound increase, the predicted salary transformed to the power of .71 increases by 2333.2 on average.

It is also necessary to check whether the model assumptions have improved.

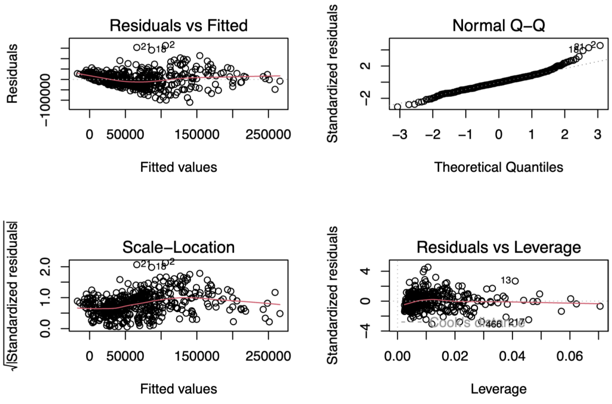


Figure 7. *R Diagnostic Plots for Model 1*

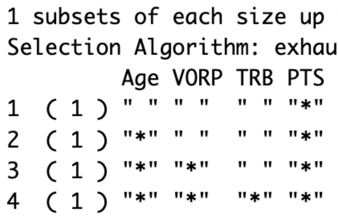
The model assumptions have slightly improved after the transformation of the response variable. We can see in the residuals vs. fitted plot that the points are more randomly scattered than in Model 0 and the red line is much straighter showing more linearity. The Normal QQ plot also has improved from model 0 with less deviation of the points from the general trend. The scale-location plot shows some model assumptions are not validated as the red line is curved and the points are not randomly scattered. The residuals vs. leverage plot also shows possible outliers or influential points.

While not all the model assumptions are completely validated, they have improved from Model 0, and in practice, we will take these assumptions. Ideally, they would be better fulfilled, but at least they have improved from our original model. We will continue with variable selection to try to simplify our model.

3.3. Reducing Model 1 (Full Transformed MLR)

Since simpler models are preferred because of their interpretability, we wanted to try variable selection to improve our transformed model. We did this using the two methods for model selection: considering all possible subsets, and stepwise regression (in both directions and for both AIC and BIC). Note that all of these methods are performed on the transformed model.

3.3.1. Method 1: Consider all possible subsets.



```
1 subsets of each size up to 4
Selection Algorithm: exhaustive
          Age VORP TRB PTS
1  ( 1 ) " " " "  " " "*"
2  ( 1 ) "*" " "  " " "*"
3  ( 1 ) "*" "*"  " " "*"
4  ( 1 ) "*" "*"  "*" "*"
```

Figure 8. *R best predictors for each subset size*

This process (Figure 8) suggested the following; for a model with only one predictor we should use PTS, for two predictors we should use PTS and Age, for three predictors we should use PTS, Age, and VORP, and trivially, for four predictors, we use the full model.

#	R^2_{Adj}	AIC	AICc	BIC
1	0.528	14769.21	14769.26	14777.50
2	0.646	14636.79	14636.88	14649.23
3	0.675	9810.99	9811.12	9827.58
4	0.679	9807.18	9807.36	9827.91

Table 2. *Number of predictors and goodness-of-fit measures*

Looking at the goodness-of-fit measures in Table 2, we select the models with 3 and 4 predictors for further study, as they have the highest R^2_{adj}, and lowest AIC, AICc, and BIC values. The model with 4 predictors corresponds to the full transformed model (Model 1), which we discussed earlier. We will call the model with 3 predictors "Model 2."

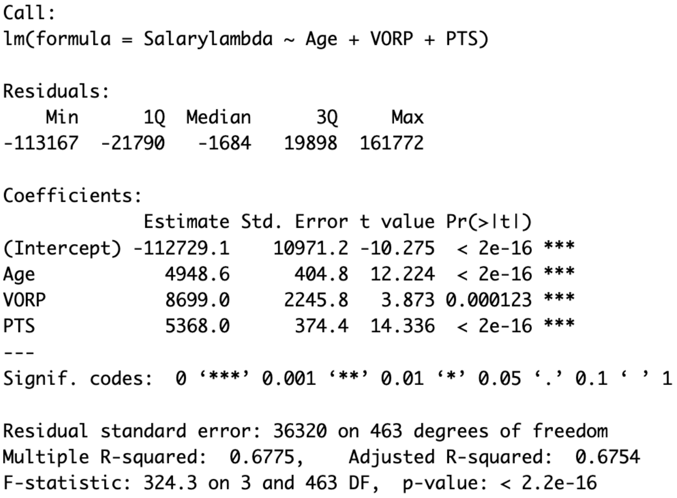
3.3.2. Method 2: Stepwise Regression.

Before looking into Model 2, we will analyze whether stepwise regression agrees with Method 1.

The process for stepwise regression can be found in the R markdown file. We determined that stepwise AIC suggests Model 1, and stepwise BIC suggests Model 2. It makes sense that BIC suggests the reduced model since BIC has a greater penalty for complexity when $\log n > 2$, where n is the number of rows of data. We found that $\log n = 6.146$, which is greater than 2.

3.4. Model 2: Reduced Transformed MLR

$$\widehat{Salary}^{0.71} = -112729.1 + 4948.6Age + 5368.0PTS + 8699VORP \qquad (3)$$



```
Call:
lm(formula = Salarylambda ~ Age + VORP + PTS)

Residuals:
    Min      1Q  Median      3Q     Max 
-113167  -21790   -1684   19898  161772 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -112729.1    10971.2 -10.275  < 2e-16 ***
Age            4948.6      404.8  12.224  < 2e-16 ***
VORP           8699.0     2245.8   3.873 0.000123 ***
PTS            5368.0      374.4  14.336  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36320 on 463 degrees of freedom
Multiple R-squared:  0.6775,	Adjusted R-squared:  0.6754 
F-statistic: 324.3 on 3 and 463 DF,  p-value: < 2.2e-16
```

Figure 9. *R Summary for Model 2*

Figure 9 indicates significant slopes for all predictors. 67.75% of the variation in Salary is explained by the model. The p-value of the overall F-test also suggests that at least one of the slopes of the predictors is significant.

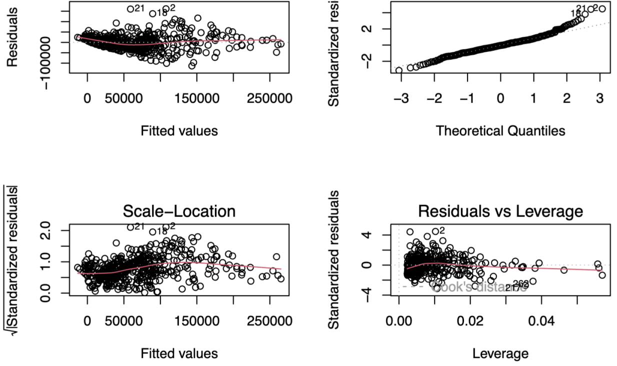


Figure 10. *R Diagnostic Plots for Model 2*

The diagnostic plots in Figure 10 appear very similar to those of Model 1 (Figure 7), which suggests our validation of model assumptions has not changed, as expected.

3.5. Final Model Choice

The final candidate models are Model 1 (Regression equation 2) and Model 2 (Regression equation 3).

All of the predictors are significant in both models, as indicated by Figures 6 and 9, and the model assumptions are validated in similar ways, as seen in Figures 7 and 10. None of the VIF_j's are greater than 5, so neither model has multicollinearity issues.

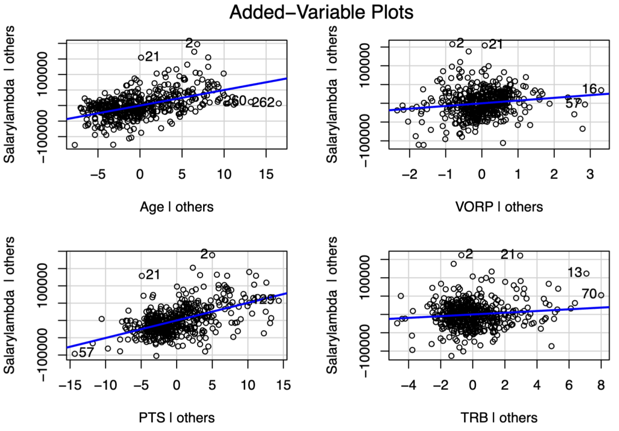


Figure 11. *R Added-Variable Plots for Model 1*

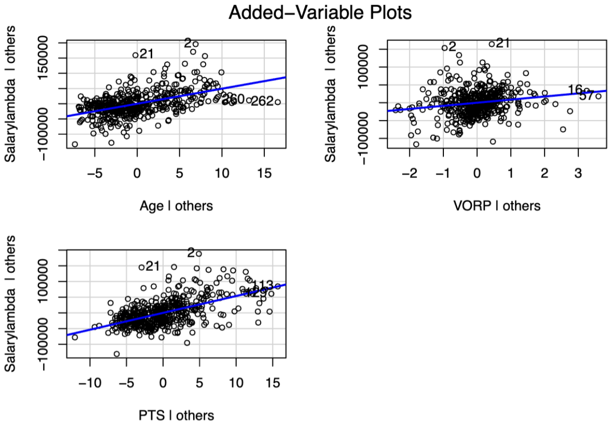


Figure 12. *R Added-Variable Plots for Model 2*

The Added-Variable plots (Figures 11 and 12) are appropriate in both models as they exhibit positive slopes, consistent with the R summaries in Figures 6 and 9.

We will choose our model based on goodness-of-fit measures because of the similarities between both models regarding other factors (predictor significance, the validity of model assumptions, VIF, and Added-Variable plots).

Although Model 2 has a lower BIC, Model 1 has higher R^2_{adj}, and lower AIC and AICc (Table 3). This inclined us towards opting for the full model. In addition, the TRB Added-Variable plot for Model 1 (Figure 11) shows a significant slope, indicating that TRB adds unique information to the model. Because of these reasons, we chose Model 1, the full transformed MLR model, as our final model.

Model	R^2_{Adj}	AIC	AICc	BIC
1	0.675	9810.99	9811.12	9827.58
2	0.679	9807.18	9807.36	9827.91

Table 3. *Goodness-of-fit measures for Models 1 and 2*

4. Discussion

4.1. Project Summary

Our goal in this project was to predict NBA player salaries in the 2022-2023 season using points per game (PTS), value over replacement player (VORP), total rebounds per game (TRB), and Age as predictor variables.

We first looked at the scatterplot matrix, summary statistics, and correlation matrix to understand the nature of our data. We started by fitting a full MLR model (Model 0) and exploring its outliers, leverage, and influential points. The diagnostic and standardized residual plots suggested that the model assumptions were not valid so we needed to transform our model. In addition, TRB was not a significant variable.

Using the inverse response method (Model 1), we were able to make all of our predictors significant and improve our model assumptions. For the sake of simplicity and interpretability, we wanted to see if a reduced model would be more appropriate. To do this, we used the subset selection method, as well as backward and forward stepwise regression for both AIC and BIC. These processes suggested that the best reduced model would be one without TRB, i.e., one with PTS, Age, and VORP as predictors. We called this model, Model 2. The diagnostic plots were similar to those of Model 1, and all of the predictors were significant.

Our best candidates are Model 1 and Model 2. None of the VIFs are greater than 5, so we can deduce that we do not have multicollinearity issues in either model. The Added-Variable plots also show positive slopes for both models. To choose between the candidates, we selected Model 1 as our final model because it has a higher Adjusted R-squared, lower AIC, lower AICc, and nearly equal BIC.

4.2. Our Project in a Broader Context

Our final model attempts to quantify the impact of specific performance metrics—Age, Points Per Game (PTS), Value Over Replacement Player (VORP), and Total Rebounds (TRB)—on the salaries of NBA players. The strength of our model, as indicated by the multiple R-squared value, suggests that these variables do indeed have a significant relationship with salary, which makes sense in a real-world context.

NBA teams operate in a competitive market where player performance metrics are key considerations for contract negotiations. Age is often associated with both experience and potential physical decline and thus its inclusion in salary considerations is intuitive. Points per game (PTS) is a fundamental measure of a player's offensive contribution, while total rebounds (TRB) are a vital statistic for understanding a player's defensive capabilities and overall game impact.

VORP is a more complex metric that captures a player's overall contribution to the team compared to a theoretical replacement player. While it is less significant in our model, it is still a relevant factor, indicating that teams value players who can do more than what a replacement-level player could offer. Literature from sports economics, such as an article from Basketball Noise [3], support the idea that teams need to consider not just the player's current performance but also their strategic value, how their skills complement the team's playing style, and their potential for future contribution.

Another study [2] examined a range of variables, including points per game, rebounds per game, and experience, using multiple regression to determine their significance in predicting NBA player salaries. It was found that experience, points, and rebounds are significant predictors of player pay. Interestingly, personal fouls were also significant, potentially because players who are more involved in the game have more opportunities to commit fouls.

An additional study [1], focusing on the determinants of NBA player salaries suggested that points per game and field goal percentage were the most statistically significant factors. This finding is consistent with our model, as it also includes points per game (PTS) as a key variable. The same study found rebounds to be a significant non-scoring contributor, supporting our inclusion of TRB in the model.

However, while our model does make sense in the real-world context of the NBA, it's crucial to acknowledge that it doesn't capture the entirety of the complexities involved in player salary determination. This model is only a surface-level analysis of these complexities that go into how much a player is paid. Factors such as a player's marketability, personal brand, leadership, their agent, and the potential to sell merchandise and tickets also play a significant role. These are often hard to quantify but can still affect a player's salary as much as their on-court performance.

4.3. Limitations and future improvement

The limitations of our model lie primarily in the model assumptions. We ideally want all of our model assumptions to be guaranteed before we move forward in our model selection, however, in practice, this is not always tractable when using real-world data.

In our final model, normality and linearity are mostly satisfied but the variance of the errors does not seem to be validated as seen in large clusters in the lower x-values of the diagnostic plots.

As statisticians, we would not want to go forward with models that do not validate all assumptions (linearity, normality of error terms, constant variance of error terms).

The non-constancy can be explained by the nature of our dataset, which shows a high number of outliers. Furthermore, basketball salaries themselves have a large variance as the more well-renowned players earn exponentially more than a rookie would. In addition, the NBA league has particular rules around salaries such as minimum and maximum salary caps (starting at 1 million and going all the way up to 56 million).

We can work to improve this variance in the future by possibly analyzing salaries in two sectors: rookie players and veteran players. By breaking up our data, we could potentially see the assumption of variance validated as the distribution of salary would not be as extreme.

Our dataset did not offer any statistics on player fame level, which could affect salary, and perhaps have been measured through Instagram followers or advertisement appearances. Another potential improvement could have been to find and test such predictors.

Overall, we believe we followed proper statistical procedures throughout our analysis and chose our final model to the best of our abilities.

References

[1] R. Lyons, N. Jackson, and A. Livingston, "Determinants of NBA Player Salaries", *The Sport Journal*, 2014. [Online]. Available: https://thesportjournal.org/article/determinants-of-nba-player-salaries/.

[2] K. Sigler and W. Compton, "NBA Players' Pay and Performance: What Counts?", *The Sport Journal*, 2018. [Online]. Available: https://thesportjournal.org/article/nba-players-pay-and-performance-what-counts/.

[3] James. "NBA Contracts: A Complete Guide". (2021), [Online]. Available: https://basketballnoise.com/nba-players-contracts-a-complete-guide/.

[4] J. Welsh. "NBA Player Salaries (2022-23 Season)". (2023), [Online]. Available: https://www.kaggle.com/datasets/jamiewelsh2/nba-player-salaries-2022-23-season.

[5] "Do euroleague stars earn more than nba benchwarmers?" (), [Online]. Available: https://basketnews.com/news-176287-do-euroleague-stars-earn-more-than-nba-benchwarmers.html#:~:text=On%20average%20a%20player%20loses,which%20really%20changes%20the%20picture.&text=Taking%20that%20NBA%20benchwarmers%20are,But%20what%20about%20Europe%3F.

[6] "How much do nba players make? average salary from 1990-2022". (), [Online]. Available: https://www.thehoopsgeek.com/average-nba-salary/.