Using Linear Regression to Predict NBA Salaries

Ralf Adam, Lucien McNulty, Cameron Scolari

Loyola Marymount University

December 2024

Abstract

The NBA is renowned not only for its global popularity but also for the huge salaries awarded to its hundreds of players who compete across 30 different teams. This study aims to analyze the factors influencing player salaries using linear regression techniques. Using various statistics from the 2023-2024 season and salary data from the 2024-2025 season, we developed predictive models that identified key variables that contribute to salary determination. This is important because players deserve to get paid fair amounts with respect to others in the league and teams must not overpay players. Each team has a "salary cap" that limits the amount that they can spend on their players' salaries. Our research aims to address this issue and provide context as to why players get paid so much, highlight what teams should consider when offering salaries, and help teams find a balance between paying their players just enough, but not too much so that they still have money to pay other players. Stepwise selection methods were used to optimize the regression model, with the backward selection model demonstrating the best fit based on AIC values. Due to a violation of regression assumptions, we had to transform our data using the square-root function and choose to use robust regression techniques. The results showed that metrics like a player's height, draft number, age, points per game, assists per game, and minutes per game have a big impact on salaries. Despite the challenges faced with model building and time constraints, we were able to construct a model that, on average, missed a player's salary by less than a million dollars. This study shows the value of using data to guide team salary decisions and points to future research opportunities, such as looking at historical salary trends.

1 Introduction

The NBA is one of the most popular sports leagues in the world, known for its star athletes and global fanbase. The league provides basketball players an opportunity to showcase their talents on a world stage but also provides astronomical financial rewards with player's salaries often making the news for their staggering prices. LeBron James, a veteran player of the league, has accumulated more than $400 million

in his career from professional contracts. These salaries are not only influenced by on-court performance, but also by factors like marketability, team strategy, and the overall economy of the NBA. All NBA teams spend over $4B on the players that make up the league every year, but do they have to? Our research aims to investigate why players get paid so much. We implement a linear regression analysis on several potential regressor variables to see what affects a player's salary. In this paper, we introduce the methods we used to collect data, construct our model, and perform the various tests we used to extract accurate results. From there, we present our findings and suggestions for future research.

2 Data Summary

2.1 Data Source

We used a public NBA API to collect various statistics for 429 players. This data extraction process was completed using JupyterNotebook. We used the API to accurately filter the data and curate a dataset we could use in R. During this step, we removed various columns that we felt would not contribute to the player's salary or columns that provided redundant information. Some columns, like a player's position, required us to convert categorical variables to numerical data that could be used during model building. The next step was to manually input each player's salary. We used hoopshype.com for the salary information. We found that players' salary follows a right-skewed distribution. It is important to mention that the salary data was taken from the 2024-2025 season while every other statistic was taken from the 2023-2024 season. We chose to use this model because we felt as if a player's performance should influence their future salary, not the one they are currently on. Also, there is not a large enough sample size from this season to build the best dataset. On top of this, we had to manually add three players due to some issues during data collection and preprocessing. After the data was in a format we could work with, we imported it to R and began conducting various data analysis tasks.

2.2 Variables

Table 2 shows a list of the 13 regressor variables and the response variable that were used in this study. As stated before, regressor variable statistics were taken from the 2023-2024 season and the salary is the player's current 2024-2025 season salary. In the table, the descriptions of the explanatory and response variables are displayed.

Table 1: All Variables

variable	description
POSITION	A players position (categorical)
DRAFT_NUMBER	Draft position
WEIGHT	A player's weight (lbs)
HEIGHT	A player's height (in)
SEASON_EXP	Number of seasons in the league
GP	Total games played in the season
MPG	Minutes played per game
RPG	Rebounds per game
APG	Assists per game
SPG	Steals per game
BPG	Blocks per game
PPG	Points per game
Age	A player's age (years)
Salary	A player's salary ($)

For players that went undrafted, there would be no data for them in the 'DRAFT_NUMBER' column. Since the draft consists of 60 picks, we chose to assign the number 61 to players who went undrafted.

2.3 Summary Statistics

Table 2 displays the mean and standard deviation for each variable we are testing. It would not make sense to calculate the mean or standard deviation for the 'POSITION' regressor variable because it is a categorical variable. This is the only categorial variable in the study. We used R to automatically make 6 indicator variables for the 7 levels of position.

Table 2: Variable Summary Statistics

variable	mean	standard deviation
POSITION	N/A	N/A
DRAFT_NUMBER	29.48	20.76
WEIGHT	215.65	23.22
HEIGHT	78.47	3.07
SEASON_EXP	5.38	4.06
GP	53.73	22.75
MPG	21.37	9.46
RPG	3.90	2.49
APG	2.35	1.97
SPG	0.67	0.38
BPG	0.46	0.44
PPG	10.08	6.97
Age	26.44	4.31
Salary	11,546,556.80	12,820,700.56

There were 7 total positions and the following are the position proportions. About 29.1% of players were forwards. About 8.1% were centers. About 4.2% were center-forwards. About 41.4% were guards. About 7% were forward-centers. About 2.8% were forward-guards. About 7.4% were guard-forwards.

3 Methods and Results

3.1 Model Development

We tried forward stepwise selection, backward stepwise selection, and both-direction stepwise selection to see if these different methods produced different results. The summary statistics for the forward selection model can be found in Table 3. See Table 4 for the summary statistics of the backward selection model. The same model as the one found in Table 4 was found for the both-direction stepwise selection model so it's table is omitted.

These results were surprising as a negative y-intercept that large was not something we expected. Additionally, the 'RPG' and 'HEIGHT' p-values are greater than

Table 3: Forward Stepwise Selection Summary Statistics

Variable	Estimate	Std. Error	t value	p-value
(Intercept)	-25282706	12226511	-2.068	0.03926
PPG	1230751	89339	13.776	$2*10^{-16}$
SEASON_EXP	663317	89977	7.372	$8.95*10^{-13}$
GP	-56544	17389	-3.252	0.00124
RPG	217058	226751	0.957	0.33899
APG	844634	297636	2.838	0.00476
HEIGHT	268081	157856	1.698	0.09019

0.05 which may suggest some multicollinearity in this model.

Table 4: Backward Stepwise Selection Summary Statistics

Variable	Estimate	Std. Error	t value	p-value
(Intercept)	-32654245	9495720	-3.439	0.000642
HEIGHT	365246	120882	3.022	0.002668
SEASON_EXP	667957	89837	7.435	$5.85*10^{-13}$
PPG	1262457	82964	15.217	$2*10^{-16}$
APG	886598	294361	3.012	0.002752
GP	-53746	17140	-3.136	0.001834

This model has the same predictor variables with the exception of 'RPG' which is excluded from this model. To determine what model we should continue investigating, we can use the AIC score. The AIC for the forward model was 14740.41 and the AIC for the backward model was 14739.34. This suggests that the backward model is slightly better than the forward model so we will use it to conduct future analysis.

3.2 Assumptions

A linear regression model is constructed on top of numerous assumptions. The first assumptions we have to consider are those relating to the errors. We have to assume that the errors are independent of one another and that they each follow a normal distribution. We know the errors are independent of one another because each player's salary is dependent on their own statistics and not those of other players. To check whether the errors follow a normal distribution, we can look at a QQ plot. A QQ plot of the studentized residuals can be viewed in Figure 1. We used studentized residuals rather than standard residuals because they are more reliable at detecting outliers. The data in the QQ-plot shows evidence of a heavy-tailed distribution.

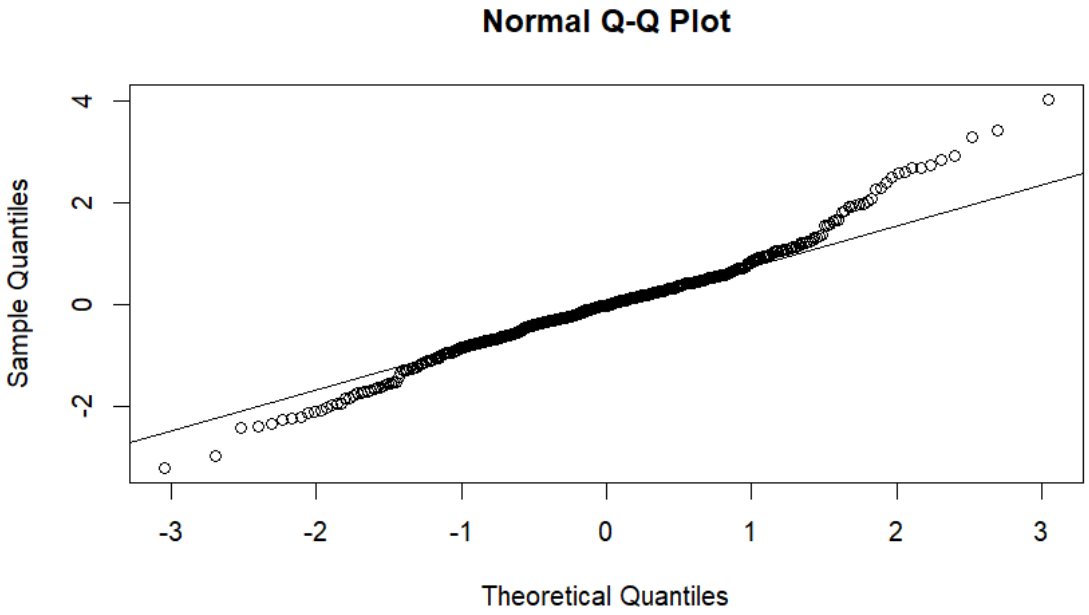

Figure 1: Q-Q Plot of Studentized Residuals.

Since the data is heavy-tailed, we can transform the salary data. We transformed the salary data using the square-root function and ran each model selecting method again. The model that performed the best was the backward selection model with an AIC of 7044.244 to the forward's 7025.278. Table 5 shows the new summary statistics for the model. Additionally, the QQ plot for this model is seen in Figure 2. Since the data follows the line, there is evidence that the errors are normally distributed.

Table 5: Square-Root Transformed Backward Stepwise Selection Summary Statistics

Variable	Estimate	Std. Error	t value	p-value
(Intercept)	-4936.105	1264.946	-3.902	0.000111
HEIGHT	49.488	15.556	3.181	0.001575
DRAFT_NUMBER	-7.856	2.464	-3.188	0.001539
Age	78.118	10.299	7.585	$2.15*10^{-13}$
PPG	125.279	14.583	8.591	$2*10^{-16}$
APG	88.974	37.735	2.358	0.018837
MPG	31.525	9.913	3.180	0.001581

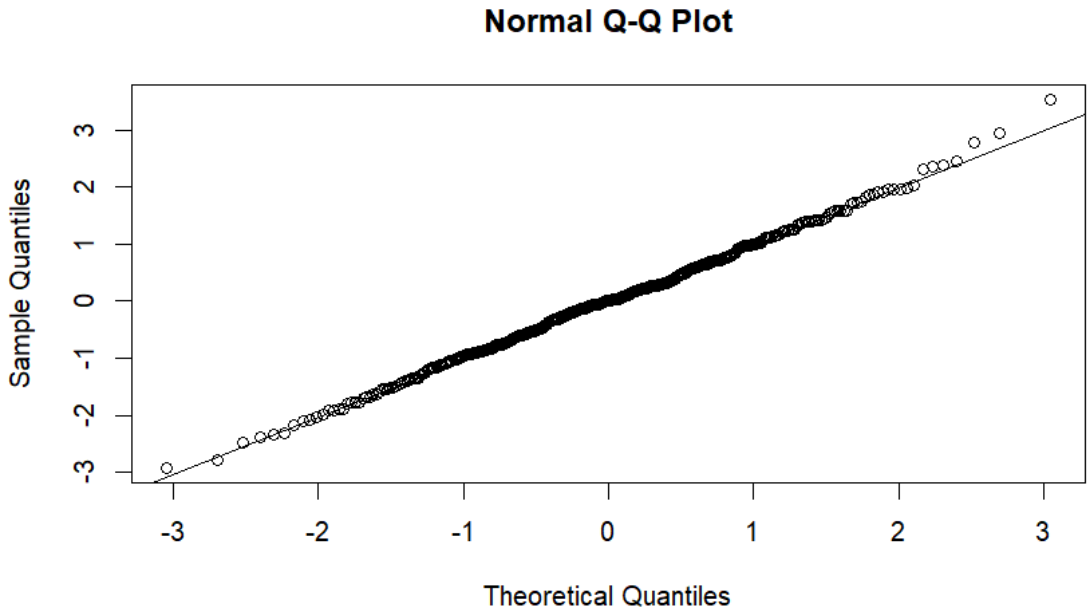

Figure 2: Q-Q Plot of Studentized Residuals.

The next assumption we can test is whether there is equal variance of the error terms by plotting the fitted values against the residuals. That can be seen in Figure 3.

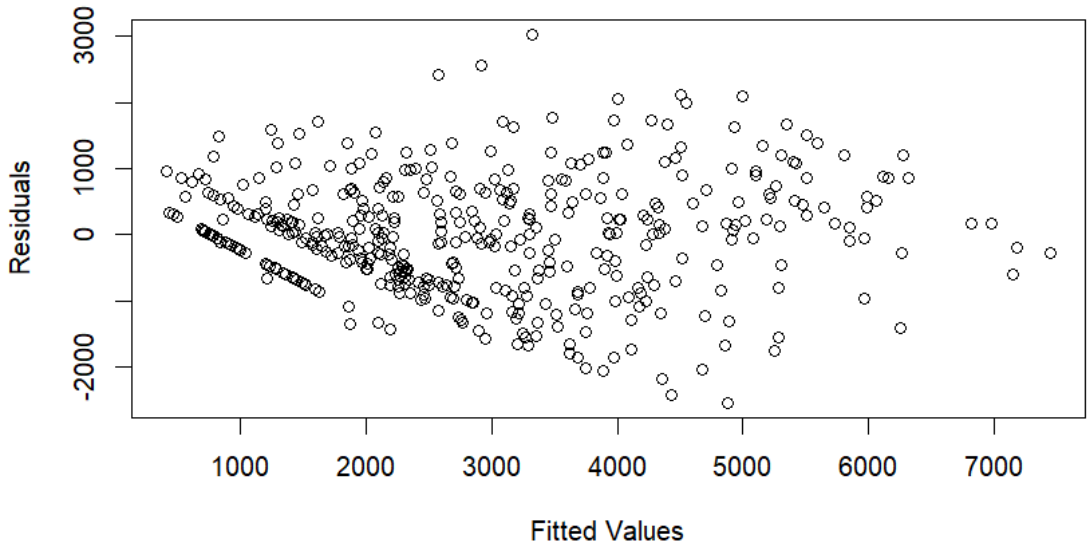

Figure 3: Fitted Values vs. Residuals.

This plot is shaped like this due to the structure of the underlying data. There are hundreds of players in the NBA but only a few are in the upper echelon of salaries. On the other end, there are a ton of players who have low salaries due to being on their rookie contract or due to them not being superstars. This graph may suggest heteroscedasticity but since the QQ plot is fine, we will continue with this model.

The next thing we can check is whether there is multicollinearity in the regressor variables. We can do this by calculating the variance inflation factor (VIF) for each regressor. The VIF score for each regressor was less than ten which indicates that there is no multicollinearity in the model.

Now we need to check for outliers, leverage, and influence points. One player's salary had a studentized residual greater than 3. This player is likely getting underpaid or overpaid which is why the model had a hard time predicting their salary. Additionally, this player was injured for a majority of the season and only played 32 games which could suggest that the data for this player should not be used for constructing the model. We found that there were 26 leverage points in the data, meaning these observations had unusual x-values. To test for influence points, we used Cook's D, difference in fits (DFFits), and difference in betas (DFBetas). Cook's D helps us identify influential data points. An influential data point is one that might influence the fitted regression model. For Cook's D, none of the D values for any data point were greater than 1 which suggests there are no influence points. Next, we conducted the DFFits test, this suggested that 28 points were influence points. DFBetas suggested that each predictor has multiple data points influencing it. Since Cook's D did not detect any influence points, the thresholds for DFFits and DFBetas may be too low. Either way, due to the potential presence of outliers and influence points, it may be better to use robust regression which is less sensitive to extreme or erroneous values.

3.3 Results

The results of doing a robust regression analysis are displayed in Table 6. If you compare these results to those in Table 5, you can see how the estimated coefficients changed. Additionally, the residual standard error for the original backward selection model was 880.6967 while the residual standard error for the robust regression model was 842.512, indicating how the robust regression model performed better during prediction.

Table 6: Square-Root Transformed Robust Regression Summary Statistics

Variable	Estimate	Std. Error	t value	p-value
(Intercept)	-4124.625	1285.730	-3.208	$1.442*10^{-3}$
HEIGHT	39.648	15.812	2.507	$1.255*10^{-2}$
DRAFT_NUMBER	-7.979	2.505	-3.185	$1.557*10^{-3}$
Age	77.237	10.469	7.378	$9.042*10^{-13}$
PPG	137.176	14.823	9.255	0
APG	74.476	38.355	1.942	$5.286*10^{-2}$
MPG	26.838	10.076	2.664	$8.039*10^{-3}$

Here is the equation for the regression line of our robust regression model with square-root transformed data:

$$\hat{\sqrt{y}} = -4124.625 + 39.648x_0 - 7.979x_1 + 77.237x_2 + 137.176x_3 + 74.476x_4 + 26.838x_5 \quad (1)$$

where x_0 is the player's height, x_1 is the position in which the player got drafted (1-61), x_2 is the player's age, x_3 is the player's points per game during the 2023-2024 season, x_4 is the player's assists per game during the 2023-2024 season, and x_5 is the player's minutes per game during the 2023-2024 season. Figure 4 displays the actual vs. predicted values. The model does a poor job predicting players with lower salaries and does a better job at predicting players with higher salaries. The R^2 value of 0.74 indicates that approximately 74% of the variability in the actual salaries is explained by the predicted salaries derived from our model.

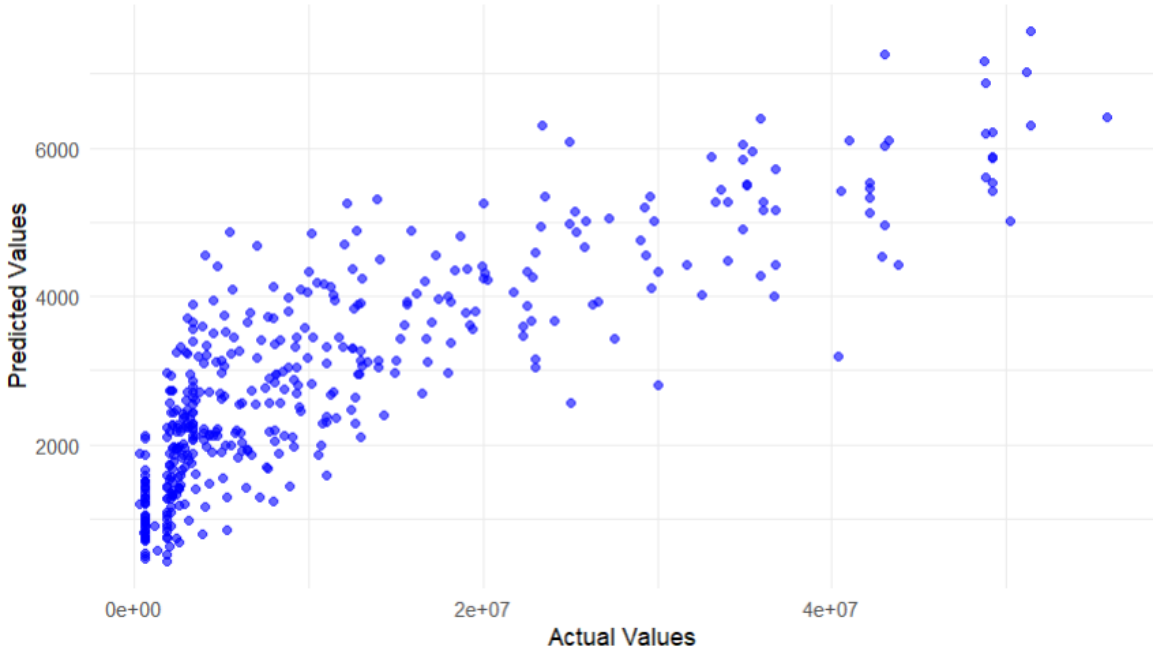

Figure 4: Actual Values vs. Predicted Values.

To accurately interpret the coefficients, we must convert the variables back to their original scale. To predict new values, we can plug in the data for our desired player into the equation and then square both sides. The result should be the player's expected salary. An example interpretation of the model is for every one increase in the players height in inches, their salary is expected to increase by 39.648^2 or about \$1,572 assuming all other variables are held constant. This makes sense as height is favored in the NBA and taller players seem to have an advantage. The only coefficient that is negative is 'DRAFT_NUMBER' which makes sense because players who are drafted early will have low x values and typically high salaries. It is also interesting that the intercept is negative. If we included the salaries for players who were drafted this year but have no statistics from last year, the intercept would likely be positive and serve as a better benchmark.

Now that we have a final model, we can conduct some hypothesis tests, construct confidence intervals (CI), and construct prediction intervals (PI). We could perform a hypothesis test that checks if all the coefficients are equal to 0. This is also called the global test or ANOVA test. We can only run this test if all of the following assumptions hold: linearity, independence of errors, errors have equal variance, errors are normally distributed, and no multicollinearity. Since our robust model does not rely on constant variance, we should not use the ANOVA test on the robust model. We will run the test on our backward model for the sake of example. Our model has 429 observations and 6 regressor coefficients so our F-statistic will be $F_{6,422}$. We already validated the previous assumptions earlier in the paper so no need to prove them again. Here are the null and alternative hypotheses we are testing on:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$
$$H_A : \text{At least one } \beta_j \neq 0 \text{ for j=1, 2,..., k}$$

After running this ANOVA test at an alpha value of 0.05, we get an F-statistic of 202.8. The p-value associated with this F-statistic is $2.2 * 10^{-16}$ which means we have statistically significant evidence to reject the null hypothesis that all the regressor coefficients are equal to 0 and accept the alternative hypothesis.

We also construced a 95% CI and a 95% PI using the robust regression model. The assumptions for the regression model are also assumptions for these tests. The 95% CI for a player whose height is 6ft, was picked 30th in the draft, is 24 years old, averages 12 points, 4 assists, and 25 minutes per game was $[\$2,760.857^2, \$3,157.706^2]$ or $[\$7,622,331, \$9,971,107]$. We are 95% confident that this interval contains the true mean salary when a player has statistics identical to this. We constructed the 95% PI using the same statistics. The interval was $[\$1,296.109^2, \$4,622.454^2]$ or $[\$1,679,898, \$21,367,080]$. We are 95% confident, we predict that the next time we see a player with these exact statistics, their salary will be within this range. As you can see, the prediction interval has higher variability.

4 Conclusion

This study was able to highlight some of the important factors that go into an NBA player's salary. It also showed that salary varies greatly and how numerous players are outliers with respect to the entire league. Hopefully this study can provide useful information to players so that they can better evaluate themselves. It should also give valuable statistical-driven insight to team management as to how much their players truly deserve. It would be interesting to see how this model responds to more historical data as well as to study how salaries have changed throughout the years. We were limited in this study by time constraints and would not have enough time to collect more historical data. The time constraints also limited the number of potential regressors we could test. Future studies may include more regressors as there are hundreds more available. The model we built is only valid for the 2024-2025 NBA season as salaries change. The final backward selection model we built had an R^2 value of 0.7389 which suggests that roughly 74% of the variance in NBA player salary is explained by the statistics that make up the regressor variables in this model and about 26% goes unexplained. Using an R^2 value for robust regression is not as useful, but the root mean squared error was 764,792 which means that on average, the predicted salary was $764,792 away from the actual salary.