NBA Player Salary and Data Analysis

Yu Yi, Chen Boying, Chen Youbai, Chen Dingyu
{109208064, 109208089, 110208017, 110208018}@nccu.edu.tw

Abstract

This paper examines the determinants of NBA player salaries through the application of a multiple linear regression model with heteroskedasticity-robust standard errors. Leveraging a comprehensive dataset comprising player statistics, lagged performance metrics, and salary data, we initially specified a model incorporating variables such as games played, minutes played, points scored, field goal attempts, and field goal percentages. Subsequently, employing rigorous statistical procedures for model refinement, including significance testing and evaluation of model fit indices, we identified significant predictors of player salaries. Our final model, accounting for approximately 52% of the variance in player salaries, underscores the pivotal role of performance metrics in salary determination. Additionally, this paper offers a comprehensive analysis of NBA player salaries, detailing data preparation techniques and model selection processes, and provides insights into the factors shaping player compensation dynamics in professional basketball. Open source code and dataset available at: [https://github.com/yvonne90190/NBA_Salary_Advanced_Stats_Analysis]

1 Introduction

The National Basketball Association (NBA) stands at the apex of professional sports, boasting a multi-billion-dollar industry and featuring top-tier athletes who command substantial salaries. Understanding the intricate factors that shape these salaries holds paramount importance for a diverse range of stakeholders, including team management, players, and analysts. Player salaries are widely recognized to be influenced by a multitude of performance metrics, both current and historical. A comprehensive understanding of these determinants not only facilitates player evaluation but also guides contract negotiations and informs strategic team-building efforts.

While prior research in sports economics has explored the factors affecting player compensation, much of it has focused on single-season performance metrics. However, given the dynamic nature of player performance, it becomes imperative to incorporate lagged performance variables to effectively capture consistency and long-term potential. Thus, the essence of this study lies in constructing a robust regression model that integrates both current and lagged performance metrics to identify the pivotal determinants of NBA player salaries.

Spanning from 2001 to 2023, our dataset encompasses player performance metrics alongside corresponding salaries. This paper embarks on a comprehensive exploration of NBA player salaries, leveraging data-driven methodologies to unveil underlying patterns and drivers of salary disparities. Our objectives are twofold: firstly, to ascertain the significant predictors of player salaries, and secondly, to address pertinent issues such as stationarity and multicollinearity. Methodologically, we utilize heteroskedasticity-robust standard errors to ensure robust statistical inference, effectively mitigating the common issue of heteroskedasticity in cross-sectional data. By adopting this approach, we aim to provide a more precise estimation of the impact of performance metrics on salaries. Additionally, the paper evaluates the statistical significance of various predictors and compares model performance using AIC and BIC criteria, culminating in a well-specified model that elucidates the primary determinants of NBA player salaries.

2 Data Preparation

The first step in our methodology involved data preparation. We obtained the dataset from NBA records, spanning from 2001 to 2023, encompassing various player statistics such as games played (GP), minutes played (MIN), points (PTS), field goals attempted (FGA), among others. The dataset also included salary information for each player, totaling 9130 entries.

2.1 Data Collection

First, we scraped the data from the ESPN website and saved it as a CSV file. We then read the dataset from the CSV file using the following R code for further analysis:

```
# Read the dataset
data <- read.csv('C:/Users/Downloads/NBA_Data.csv')
```

This command imports the data from the specified file path into the `data` variable, creating a dataframe that contains all the relevant information about NBA players and their salaries over time.

2.2 Stationarity Test

To ensure the reliability of our regression models, we first tested the stationarity of the salary variable using the Augmented Dickey-Fuller (ADF) test to detect unit roots. The stationarity test is essential because non-stationary data can lead to spurious regression results.

Next, we extracted a list of unique player names from the dataset since the ADF test needs to be performed individually for each player to determine the stationarity of their salary data. We implemented the ADF test using a loop across individual player salary data to accommodate potential non-stationarity:

```
# Unique player names
unique_players <- unique(data$player)
adf_test_results <- list()

for (player_name in unique_players) {
  # 提取每個球員的資料
  player_data <- subset(data, player == player_name)

  # 提取球員的薪資數據
  player_salaries <- player_data$salary

  # 檢查資料點數量
  if (length(player_salaries) > 1) {
    # 使用 tryCatch 捕獲可能的錯誤
    adf_test_result <- tryCatch({
      adf.test(player_salaries, alternative = "stationary")
    }, error = function(e) {
      return(NA)  # 如果發生錯誤，返回 NA
    })

    adf_test_results[[player_name]] <- adf_test_result
  } else {
     adf_test_results[[player_name]] <- "Not enough data points for ADF
test"
  }
}
```

This loop goes through each player, extracts their salary data, and checks if there are enough data points to perform the ADF test. If there are sufficient data points, it performs the ADF test; otherwise, it notes that there are not enough data points for the test.

Finally, we displayed the results of the ADF test for all players:

```
# Display ADF test results
cat("\nADF檢定結果:\n")
for (player_name in names(adf_test_results)) {
  cat("\nPlayer:", player_name, "\n")
  print(adf_test_results[[player_name]])
}
```

This code iterates over the results and prints them. Due to the nature of NBA careers, which typically span less than 20 years, and the dataset covering the years 2001 to 2023, most players did not have enough data points to perform the ADF test successfully.

Only one player, Vin Baker, had a sufficient number of data points for the ADF test. Here are the results for Vin Baker:

```
Player: Vin Baker
Augmented Dickey-Fuller
Test data: player_salaries
Dickey-Fuller = -1.7321,
Lag order = 1,
p-value = 0.6745
alternative hypothesis: stationary
```

The ADF test for Vin Baker's salary data resulted in a Dickey-Fuller statistic of -1.7321 with a p-value of 0.6745, indicating that we cannot reject the null hypothesis of non-stationarity. This suggests that Vin Baker's salary data is not stationary over the period analyzed.

In summary, the preparation of the data involved reading the dataset, extracting unique player names, and performing the ADF test on each player's salary data. Due to the limited length of NBA careers and the timeframe of the dataset, most players did not have enough data points for the ADF test. Only one player, Vin Baker, had a successful ADF test, which indicated non-stationarity in his salary data.

3 Model Building

In this section, we describe our approach to building a predictive model for NBA player salaries based on historical performance data. Our goal is to create a model that can accurately predict player salaries using a set of performance metrics from previous seasons. To achieve this, we employ a step-by-step methodology that includes data preparation, initial model creation, stepwise regression for variable selection, and assessment of multicollinearity.

3.1 Initial Model Creation

Step 1: Data Preparation

A crucial aspect of our methodology was to create lagged variables to incorporate historical player performance into the analysis. We implemented a process to create lagged variables for various player statistics such as games played (GP), minutes played (MIN), points (PTS), field goals attempted (FGA), and others. These lagged variables allowed us to analyze the impact of past performance on current salary levels.

The data was first arranged and grouped by player, then lagged variables for up to five previous periods were created for various performance metrics. This step ensures that the historical performance of players is considered when modeling their salaries. Lagging the variables up to five periods is based on the typical length of NBA player contracts, which rarely exceed five years. Hence, performance data older than five years is considered less relevant to current salary predictions.

```
# Create lagged variables
data <- data %>%
  arrange(player, year) %>%
  group_by(player) %>%
  mutate(
    GP_lag1 = lag(GP, 1),      GP_lag2 = lag(GP, 2),      ...   GP_lag5 = lag(GP, 5),
    MIN_lag1 = lag(MIN, 1),    MIN_lag2 = lag(MIN, 2),    ...   MIN_lag5 = lag(MIN, 5),
    PTS_lag1 = lag(PTS, 1),    PTS_lag2 = lag(PTS, 2),    ...   PTS_lag5 = lag(PTS, 5),
    FGA_lag1 = lag(FGA, 1),    FGA_lag2 = lag(FGA, 2),    ...   FGA_lag5 = lag(FGA, 5),
    FG_lag1 = lag(FG., 1),     FG_lag2 = lag(FG., 2),     ...   FG_lag5 = lag(FG., 5),
     X3PA_lag1 = lag(X3PA, 1),   X3PA_lag2 = lag(X3PA, 2), ...    X3PA_lag5 = lag(X3PA,
5),
    X3P_lag1 = lag(X3P., 1),   X3P_lag2 = lag(X3P., 2),   ...   X3P_lag5 = lag(X3P., 5),
    FTA_lag1 = lag(FTA, 1),    FTA_lag2 = lag(FTA, 2),    ...   FTA_lag5 = lag(FTA, 5),
    FT_lag1 = lag(FT., 1),     FT_lag2 = lag(FT., 2),     ...   FT_lag5 = lag(FT., 5),
    REB_lag1 = lag(REB, 1),    REB_lag2 = lag(REB, 2),    ...   REB_lag5 = lag(REB, 5),
    AST_lag1 = lag(AST, 1),    AST_lag2 = lag(AST, 2),    ...   AST_lag5 = lag(AST, 5),
    STL_lag1 = lag(STL, 1),    STL_lag2 = lag(STL, 2),    ...   STL_lag5 = lag(STL, 5),
    BLK_lag1 = lag(BLK, 1),    BLK_lag2 = lag(BLK, 2),    ...   BLK_lag5 = lag(BLK, 5),
    TO_lag1 = lag(TO, 1),      TO_lag2 = lag(TO, 2),      ...   TO_lag5 = lag(TO, 5),
    log_salary = log(salary)
  ) %>%
  ungroup()
```

Step 2: Removing Missing Values

After creating the lagged variables, rows with any missing values were removed to ensure the completeness of the dataset for regression analysis.

```
# Remove rows with NA values
data <- na.omit(data)
```

Result

An initial linear regression model is created using all the variables, including the original and lagged performance metrics, to predict the log of player salaries.

```
# Build initial model with all variables and lagged terms
initial_model <- lm(log_salary ~
GP + GP_lag1 + GP_lag2 + GP_lag3 + GP_lag4 + GP_lag5
+ MIN + MIN_lag1 + MIN_lag2 + MIN_lag3 + MIN_lag4 + MIN_lag5
+ PTS + PTS_lag1 + PTS_lag2 + PTS_lag3 + PTS_lag4 + PTS_lag5
+ FGA + FGA_lag1 + FGA_lag2 + FGA_lag3 + FGA_lag4 + FGA_lag5
```

```
+ FG. + FG_lag1 + FG_lag2 + FG_lag3 + FG_lag4 + FG_lag5
+ X3PA + X3PA_lag1 + X3PA_lag2 + X3PA_lag3 + X3PA_lag4 + X3PA_lag5
+ X3P. + X3P_lag1 + X3P_lag2 + X3P_lag3 + X3P_lag4 + X3P_lag5
+ FTA + FTA_lag1 + FTA_lag2 + FTA_lag3 + FTA_lag4 + FTA_lag5
+ FT. + FT_lag1 + FT_lag2 + FT_lag3 + FT_lag4 + FT_lag5
+ REB + REB_lag1 + REB_lag2 + REB_lag3 + REB_lag4 + REB_lag5
+ AST + AST_lag1 + AST_lag2 + AST_lag3 + AST_lag4 + AST_lag5
+ STL + STL_lag1 + STL_lag2 + STL_lag3 + STL_lag4 + STL_lag5
+ BLK + BLK_lag1 + BLK_lag2 + BLK_lag3 + BLK_lag4 + BLK_lag5
+ TO + TO_lag1 + TO_lag2 + TO_lag3 + TO_lag4 + TO_lag5,
data = data)
```

3.2 Stepwise Regression

Step 1: Build Stepwise regression model

Stepwise regression is used to refine the initial model by selecting the most significant predictors. This method iteratively adds and removes variables to find the best fitting model based on AIC (Akaike Information Criterion).

```
# Use stepwise regression to select the best model
stepwise_model <- stepAIC(initial_model, direction = "both")

# View summary of the stepwise regression model
summary(stepwise_model)
```

```
Call:
lm(formula = log_salary ~ GP + GP_lag2 + GP_lag5 + MIN + MIN_lag1 +
    MIN_lag3 + PTS_lag1 + PTS_lag3 + PTS_lag4 + FGA + FGA_lag1 +
    FGA_lag2 + FGA_lag3 + FGA_lag4 + FG. + FG_lag1 + FG_lag2 +
    X3PA_lag1 + X3P. + FTA + FTA_lag5 + REB_lag1 + AST + AST_lag1 +
    STL_lag1 + STL_lag5 + BLK_lag4 + TO_lag1 + TO_lag5, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6122 -0.3914  0.0810  0.5278  2.3558

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.0590824  0.2214381  54.458  < 2e-16 ***
GP           0.0028736  0.0009702   2.962 0.003082 **
GP_lag2      0.0020042  0.0010395   1.928 0.053948 .
GP_lag5      0.0017130  0.0010039   1.706 0.088045 .
MIN          0.0260728  0.0060130   4.336 1.50e-05 ***
MIN_lag1     0.0094859  0.0066415   1.428 0.153324
```

```
MIN_lag3     0.0150448  0.0051506   2.921 0.003517 **
PTS_lag1     0.0302940  0.0209471   1.446 0.148224
PTS_lag3     0.0856012  0.0193034   4.435 9.57e-06 ***
PTS_lag4     0.0413044  0.0184633   2.237 0.025355 *
FGA          0.0622633  0.0135131   4.608 4.25e-06 ***
FGA_lag1    -0.0421147  0.0282946  -1.488 0.136745
FGA_lag2     0.0385469  0.0091715   4.203 2.72e-05 ***
FGA_lag3    -0.1215264  0.0260570  -4.664 3.24e-06 ***
FGA_lag4    -0.0485462  0.0239464  -2.027 0.042725 *
FG.         -0.0055044  0.0028054  -1.962 0.049850 *
FG_lag1      0.0138132  0.0040104   3.444 0.000581 ***
FG_lag2      0.0126403  0.0036513   3.462 0.000544 ***
X3PA_lag1    0.0736512  0.0125422   5.872 4.79e-09 ***
X3P.        -0.0024961  0.0012931  -1.930 0.053655 .
FTA         -0.0525369  0.0214108  -2.454 0.014196 *
FTA_lag5    -0.0600245  0.0178739  -3.358 0.000795 ***
REB_lag1     0.0666793  0.0106032   6.289 3.69e-10 ***
AST          0.0637633  0.0225690   2.825 0.004757 **
AST_lag1     0.0353931  0.0237219   1.492 0.135808
STL_lag1    -0.1425740  0.0671415  -2.123 0.033798 *
STL_lag5     0.1869054  0.0606537   3.082 0.002079 **
BLK_lag4     0.1553822  0.0389587   3.988 6.82e-05 ***
TO_lag1     -0.1792214  0.0511087  -3.507 0.000461 ***
TO_lag5     -0.0958382  0.0430450  -2.226 0.026060 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8443 on 2886 degrees of freedom
Multiple R-squared:  0.5341,     Adjusted R-squared:  0.5294
F-statistic: 114.1 on 29 and 2886 DF,  p-value: < 2.2e-16
```

Result

The final model, selected through stepwise regression, includes a subset of the original and lagged variables that best predict player salaries. Here is the summary of the stepwise regression model:

```
stepwise_model <-lm(formula = log_salary ~
GP + GP_lag2 + GP_lag5
+ MIN + MIN_lag1 + MIN_lag3
+ PTS_lag1 + PTS_lag3 + PTS_lag4
+ FGA + FGA_lag1 + FGA_lag2 + FGA_lag3 + FGA_lag4
+ FG. + FG_lag1 + FG_lag2
+ X3P. + X3PA_lag1
+ FTA + FTA_lag5
+ REB_lag1
+ AST + AST_lag1
```

```
+ STL_lag1 + STL_lag5
+ BLK_lag4
+ TO
```

3.3 Multicollinearity Assessment

To ensure the reliability of our regression model, we assessed multicollinearity among predictor variables. Multicollinearity occurs when independent variables are highly correlated, leading to unstable estimates and inflated standard errors. We calculated the Variance Inflation Factor (VIF) for each predictor variable to identify multicollinearity.

In this section, I describe the steps taken to address multicollinearity in the model by removing variables with a Variance Inflation Factor (VIF) greater than 10.

Step 1: Calculate VIF for the Stepwise Model

The VIF is calculated for each variable in the stepwise regression model.

```
# Load necessary library
library(car)

# Calculate VIF for the stepwise regression model
vif_values <- vif(stepwise_model)
print(vif_values)
```

```
GP       MIN       FGA        FG.       X3P.       FTA       AST
1.631803 11.085414 18.159910  1.967956  1.627465  7.166272  8.841524

GP_lag2   GP_lag5  MIN_lag1  MIN_lag3  PTS_lag1  PTS_lag3  PTS_lag4  FGA_lag1
1.214568  1.214637 11.454754  6.294485 74.102838 57.872129 52.923071 75.126062

FGA_lag2  FGA_lag3  FGA_lag4   FG_lag1   FG_lag2 X3PA_lag1  FTA_lag5  REB_lag1
7.518208 59.559819 50.920591  3.123324  2.339094  3.425129  5.422601  3.174880

AST_lag1  STL_lag1  STL_lag5  BLK_lag4   TO_lag1   TO_lag5
10.040943  3.366498  3.242168  1.964191  7.413733  5.267
```

Step 2: Remove Variables with VIF Greater than 10

Based on the VIF values, variables with VIF greater than 10 are identified and removed to reduce multicollinearity. This threshold is chosen because a VIF above 10 often indicates severe multicollinearity.

```
# Identify variables with VIF > 10
variables_to_remove <- c("MIN", "FGA", "PTS_lag1", "PTS_lag3",
"PTS_lag4", "FGA_lag1", "FGA_lag2", "FGA_lag3", "FGA_lag4")
```

Step 3: Build the New Model

A new linear regression model is created using the revised formula that excludes the variables with high VIF values.

```
# Create new formula by removing identified variables
formula <- as.formula(paste("log_salary ~
GP + GP_lag2 + GP_lag5
+ MIN_lag1 + MIN_lag3
+ FG. + FG_lag1 + FG_lag2
+ X3P. + X3PA_lag1
+ FTA + FTA_lag5
+ REB_lag1
+ AST + AST_lag1
+ STL_lag1 + STL_lag5
+ BLK_lag4
+ TO_lag1 + TO_lag5"
, collapse = " + "))

# Build new model excluding variables with high VIF
new_model <- lm(formula, data = data)

# View the summary of the new model
summary(new_model)
```

```
Call:
lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4157 -0.3996  0.1081  0.5396  2.4111

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.129e+01  1.937e-01  58.291  < 2e-16 ***
GP           5.492e-03  9.187e-04   5.977 2.54e-09 ***
FG.         -3.094e-03  2.853e-03  -1.084  0.27831
X3P.        -7.463e-05  1.308e-03  -0.057  0.95450
FTA          8.710e-02  1.590e-02   5.477 4.69e-08 ***
AST          1.607e-01  2.087e-02   7.702 1.83e-14 ***
GP_lag2      2.123e-03  1.041e-03   2.040  0.04149 *
GP_lag5      1.359e-03  1.020e-03   1.332  0.18303
```

```
MIN_lag1     2.086e-02  5.284e-03   3.949 8.05e-05 ***
MIN_lag3     2.125e-02  3.263e-03   6.512 8.71e-11 ***
FGA_lag1     3.744e-02  9.006e-03   4.157 3.32e-05 ***
FG_lag1      1.917e-02  3.486e-03   5.498 4.18e-08 ***
FG_lag2      1.934e-02  3.648e-03   5.300 1.24e-07 ***
X3PA_lag1    1.118e-01  1.146e-02   9.755  < 2e-16 ***
FTA_lag5    -3.400e-02  1.596e-02  -2.130  0.03330 *
REB_lag1     6.985e-02  1.085e-02   6.437 1.42e-10 ***
AST_lag1    -2.353e-02  2.320e-02  -1.014  0.31055
STL_lag1    -1.993e-01  6.835e-02  -2.916  0.00358 **
STL_lag5     1.776e-01  6.182e-02   2.873  0.00410 **
BLK_lag4     1.685e-01  3.959e-02   4.257 2.14e-05 ***
TO_lag1     -2.133e-01  5.132e-02  -4.155 3.34e-05 ***
TO_lag5     -1.272e-01  4.207e-02  -3.025  0.00251 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8672 on 2894 degrees of freedom
Multiple R-squared:  0.507,      Adjusted R-squared:  0.5035
F-statistic: 141.7 on 21 and 2894 DF,  p-value: < 2.2e-16
```

Step 4: Calculate VIF for the New Model

The VIF is recalculated for the new model to ensure that the multicollinearity has been sufficiently reduced.

```r
# Calculate VIF for the new model
new_vif <- vif(new_model)
print(new_vif)
```

```
GP        FG.       X3P.        FTA        AST
1.386880  1.929585  1.578093  3.746494  7.163495

GP_lag2   GP_lag5  MIN_lag1  MIN_lag3  FGA_lag1
1.154565  1.188809  6.871675  2.393739  7.213690

FG_lag1   FG_lag2 X3PA_lag1  FTA_lag5  REB_lag1
2.236995  2.213482  2.709673  4.099792  3.151033

AST_lag1  STL_lag1  STL_lag5  BLK_lag4   TO_lag1   TO_lag5
9.104212  3.306156  3.191996  1.922683  7.085761  4.768923
```

Results

After removing the variables with high VIF values, the new model includes the remaining predictors with acceptable levels of multicollinearity. Here is the summary of the new model and its VIF values:

```
# Final formula after removing variables with high VIF
new_model <- "log_salary ~
GP + GP_lag2 + GP_lag5
+ MIN_lag1 + MIN_lag3 +
+ FG. + FG_lag1 + FG_lag2
+ X3P. + X3PA_lag1
+ FTA + FTA_lag5
+ REB_lag1
+ AST + AST_lag1
+ STL_lag1 + STL_lag5
+ BLK_lag4
+ TO_lag1 + TO_lag5"
```

The revised model is expected to have lower multicollinearity, improving the robustness and interpretability of the regression results.

4 Experiments

In this section, we interpret the model results and assess its practical relevance for predicting NBA player salaries.

4.1 Heteroskedasticity

We begin by checking the statistical significance of the independent variables using heteroskedasticity-robust standard errors and determining whether statistically insignificant variables should be removed from the model.

Step 1: Calculate Heteroskedasticity-Robust Standard Errors

To account for potential heteroscedasticity in the regression errors, we calculate heteroskedasticity-robust standard errors for the variables in the new model.

```
# Load necessary libraries
library(sandwich)
library(lmtest)

# Calculate heteroskedasticity-robust standard errors
hc_se <- vcovHC(new_model, type = "HC1")

# Conduct t-tests using robust standard errors
robust_test <- coeftest(new_model, vcov = hc_se)
```

```
# Display the results of the regression with robust standard errors
print(robust_test)
```

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  1.1290e+01  2.2049e-01 51.2036 < 2.2e-16 ***
GP           5.4918e-03  1.1473e-03  4.7866 1.781e-06 ***
FG.         -3.0941e-03  4.1809e-03 -0.7401  0.459329
X3P.        -7.4634e-05  1.3693e-03 -0.0545  0.956536
FTA          8.7100e-02  1.5905e-02  5.4763 4.716e-08 ***
AST          1.6072e-01  2.1203e-02  7.5801 4.619e-14 ***
GP_lag2      2.1233e-03  1.1059e-03  1.9199  0.054966 .
GP_lag5      1.3587e-03  1.0608e-03  1.2808  0.200373
MIN_lag1     2.0864e-02  5.2541e-03  3.9710 7.333e-05 ***
MIN_lag3     2.1246e-02  3.4428e-03  6.1712 7.724e-10 ***
FGA_lag1     3.7436e-02  7.9550e-03  4.7060 2.645e-06 ***
FG_lag1      1.9168e-02  4.1528e-03  4.6156 4.090e-06 ***
FG_lag2      1.9337e-02  3.9498e-03  4.8958 1.033e-06 ***
X3PA_lag1    1.1178e-01  1.0095e-02 11.0730 < 2.2e-16 ***
FTA_lag5    -3.3996e-02  1.5868e-02 -2.1424  0.032247 *
REB_lag1     6.9847e-02  1.0454e-02  6.6813 2.830e-11 ***
AST_lag1    -2.3533e-02  2.4477e-02 -0.9614  0.336420
STL_lag1    -1.9928e-01  6.5481e-02 -3.0434  0.002361 **
STL_lag5     1.7759e-01  6.1949e-02  2.8667  0.004178 **
BLK_lag4     1.6854e-01  4.0140e-02  4.1987 2.766e-05 ***
TO_lag1     -2.1326e-01  5.0880e-02 -4.1915 2.855e-05 ***
TO_lag5     -1.2725e-01  4.3589e-02 -2.9193  0.003536 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 2: Evaluate the Original Model

Before removing any variables, we re-evaluate the original model to compare its performance metrics (AIC and BIC) against the simplified model later.

```
# Build the original model
original_model <- lm(log_salary ~ GP + MIN + PTS + FGA + FG. + X3P. +
FTA + AST + GP_lag2 + GP_lag5 + MIN_lag1 + MIN_lag3 + FG_lag1 + FG_lag2
+ X3PA_lag1 + FTA_lag5 + REB_lag1 + STL_lag1 + STL_lag5 + BLK_lag4 +
TO_lag1 + TO_lag5, data = data)
summary(original_model)

# Calculate AIC and BIC for the original model
original_aic <- AIC(original_model)
```

```
original_bic <- BIC(original_model)
```

Step 3: Remove Statistically Insignificant Variables

Variables with a p-value greater than 0.05 are considered statistically insignificant and are removed to simplify the model. The simplified model is then created with the remaining significant variables.

```
# Identify variables with p-value > 0.05
variables_to_remove <- c("FG.", "X3P.", "GP_lag5", "AST_lag1")

# Create new formula excluding insignificant variables
formula <- as.formula("log_salary ~ GP + FTA + AST + GP_lag2 + MIN_lag1 + MIN_lag3 + FGA_lag1 + FG_lag1 + FG_lag2 + X3PA_lag1 + FTA_lag5 + REB_lag1 + STL_lag1 + STL_lag5 + BLK_lag4 + TO_lag1 + TO_lag5")

# Build the simplified model
simplified_model <- lm(formula, data = data)
summary(simplified_model)
```

Step 4: Compare Models Using AIC and BIC

The simplified model's performance is evaluated by comparing its AIC and BIC values against the original model. Lower AIC and BIC values indicate a better model.

```
# Print the AIC and BIC values
print(paste("Original Model AIC:", original_aic))
print(paste("Original Model BIC:", original_bic))

print(paste("Simplified Model AIC:", simplified_aic))
print(paste("Simplified Model BIC:", simplified_bic))
```

```
[1] "Original Model AIC: 7371.84901850112"
[1] "Original Model BIC: 7515.3202527362"

[1] "Simplified Model AIC: 7464.33374949384"
[1] "Simplified Model BIC: 7577.91514326328"
```

The lower AIC and BIC of the original model suggest that it provides a better balance between goodness of fit and complexity compared to the simplified model. This indicates that the original model may be a more appropriate choice for predicting NBA player salaries, as it explains the data better without being overly complex.

Step 5: Conduct ANOVA F-test

An ANOVA F-test is conducted to statistically compare the simplified model with the original model.

```
# Conduct ANOVA F-test to compare models
anova_result <- anova(simplified_model, original_model)

# Print the ANOVA test result
print(anova_result)
```

```
Analysis of Variance Table

Model 1: log_salary ~ GP + FTA + AST + GP_lag2 + MIN_lag1 + MIN_lag3 +
    FGA_lag1 + FG_lag1 + FG_lag2 + X3PA_lag1 + FTA_lag5 + REB_lag1 +
    STL_lag1 + STL_lag5 + BLK_lag4 + TO_lag1 + TO_lag5
Model 2: log_salary ~ GP + MIN + PTS + FGA + FG. + X3P. + FTA + AST +
    GP_lag2 + GP_lag5 + MIN_lag1 + MIN_lag3 + FG_lag1 + FG_lag2 +
    X3PA_lag1 + FTA_lag5 + REB_lag1 + STL_lag1 + STL_lag5 + BLK_lag4 +
    TO_lag1 + TO_lag5
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2898 2179.5
2   2893 2104.2  5    75.269 20.697 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic is 20.697 with a p-value less than 2.2e-16, which is extremely small. This indicates that the difference in the model fits is statistically significant. The p-value suggests strong evidence against the null hypothesis, indicating that the original model with more variables provides a significantly better fit to the data compared to the simplified model.

Results

Despite removing statistically insignificant variables, the comparison of AIC, BIC, and the ANOVA F-test results indicate that the original model is preferable over the simplified model.

The final model is therefore:

```
final_model <- lm(log_salary ~ GP + MIN + PTS + FGA + FG. + X3P. + FTA +
AST + GP_lag2 + GP_lag5 + MIN_lag1 + MIN_lag3 + FG_lag1 + FG_lag2 +
X3PA_lag1 + FTA_lag5 + REB_lag1 + STL_lag1 + STL_lag5 + BLK_lag4 +
TO_lag1 + TO_lag5, data = data)

summary(final_model)
```

```
Call:
lm(formula = log_salary ~ GP + MIN + PTS + FGA + FG. + X3P. +
    FTA + AST + GP_lag2 + GP_lag5 + MIN_lag1 + MIN_lag3 + FG_lag1 +
```

```
    FG_lag2 + X3PA_lag1 + FTA_lag5 + REB_lag1 + STL_lag1 + STL_lag5 +
    BLK_lag4 + TO_lag1 + TO_lag5, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4776 -0.3837  0.0951  0.5255  2.3750

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.7448258  0.2204853  53.268  < 2e-16 ***
GP           0.0022740  0.0009825   2.314 0.020711 *
MIN          0.0224930  0.0051807   4.342 1.46e-05 ***
PTS          0.0836983  0.0239455   3.495 0.000481 ***
FGA         -0.0330223  0.0259975  -1.270 0.204112
FG.         -0.0099731  0.0032935  -3.028 0.002482 **
X3P.        -0.0023667  0.0013302  -1.779 0.075302 .
FTA         -0.0626528  0.0265892  -2.356 0.018523 *
AST          0.0943790  0.0149115   6.329 2.85e-10 ***
GP_lag2      0.0022328  0.0010230   2.183 0.029151 *
GP_lag5      0.0014438  0.0010032   1.439 0.150214
MIN_lag1     0.0135879  0.0051726   2.627 0.008662 **
MIN_lag3     0.0174747  0.0032368   5.399 7.26e-08 ***
FG_lag1      0.0187688  0.0034546   5.433 6.00e-08 ***
FG_lag2      0.0182097  0.0035909   5.071 4.21e-07 ***
X3PA_lag1    0.0864640  0.0123014   7.029 2.59e-12 ***
FTA_lag5    -0.0095706  0.0157274  -0.609 0.542884
REB_lag1     0.0643334  0.0106039   6.067 1.47e-09 ***
STL_lag1    -0.1713474  0.0672160  -2.549 0.010848 *
STL_lag5     0.1793885  0.0608203   2.949 0.003209 **
BLK_lag4     0.1520837  0.0388904   3.911 9.42e-05 ***
TO_lag1     -0.0973238  0.0448687  -2.169 0.030158 *
TO_lag5     -0.1571121  0.0415191  -3.784 0.000157 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8528 on 2893 degrees of freedom
Multiple R-squared:  0.5234,    Adjusted R-squared:  0.5198
F-statistic: 144.4 on 22 and 2893 DF,  p-value: < 2.2e-16
```

The final model, accounting for heteroskedasticity using robust standard errors, explains approximately 52% of the variance in player salaries, with statistically significant predictors including minutes played (MIN), points scored (PTS), assists (AST), rebounds (REB), and turnovers (TO).

5 Evaluation

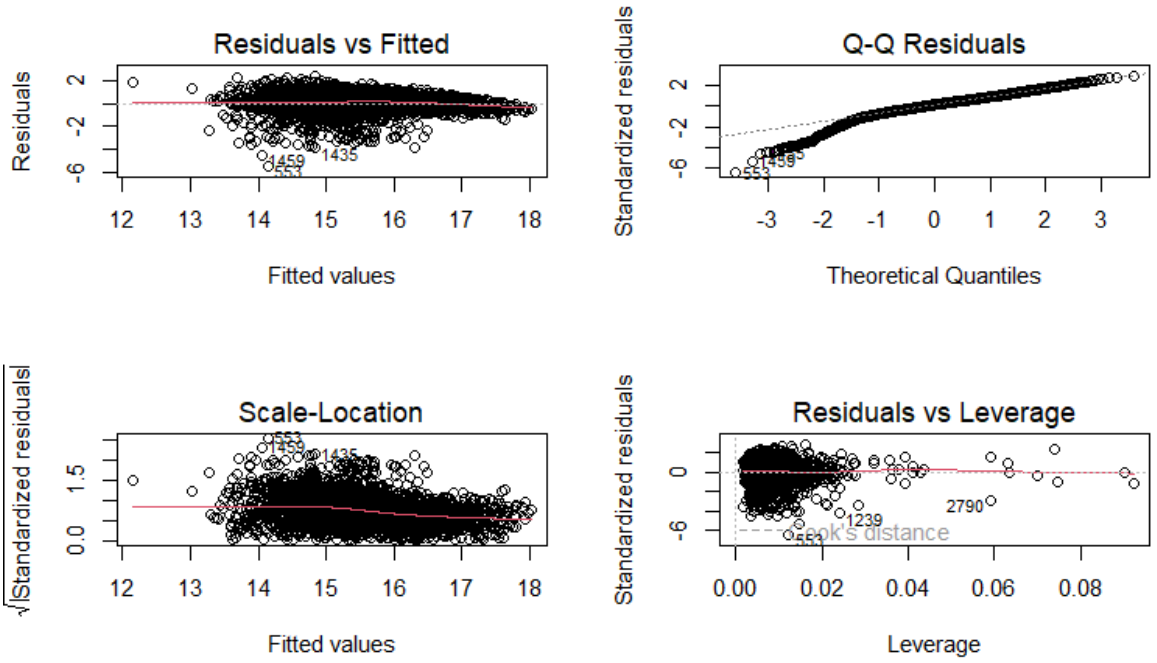
Finally, model diagnostics were performed on the chosen final model, including residual analysis and normality tests. This process ensures that the chosen model is robust, and the

variables included are statistically significant, enhancing the model's reliability and interpretability.

5.1 Residual Analysis

Diagnostic plots were generated to assess the residuals and overall fit of the final model.

```
# Diagnostic plots
par(mfrow = c(2, 2))
plot(final_model)
```



Residuals vs Fitted

This plot shows the residuals (errors) versus the fitted values (predicted values). Ideally, the points should be randomly scattered around the horizontal line (y = 0) with no clear pattern. In the plot, the residuals appear to be relatively random, but there may be a slight pattern suggesting some non-linearity. This indicates that the model's assumptions might not be fully met.

Normal Q-Q

This plot compares the standardized residuals to a normal distribution. If the residuals are normally distributed, the points should lie approximately along the reference line. In the plot, there is some deviation from the line, especially at the tails, suggesting that the residuals are not perfectly normally distributed. This could be due to outliers or heavy tails.

Scale-Location (or Spread-Location)

This plot shows the square root of the standardized residuals against the fitted values. It helps check the homoscedasticity assumption (constant variance of residuals). The plot should show a horizontal line with equally spread points. The plot shows a slight funnel shape, indicating potential heteroscedasticity (non-constant variance of residuals).

Residuals vs Leverage

This plot helps identify influential data points. Points with high leverage have a high potential to influence the regression model. Points outside the Cook's distance lines (marked by red dashed lines) are considered influential. In the plot, there are some points with high leverage, suggesting they might have a significant impact on the model.

5.2 Normality Tests

To formally test for normality of residuals, the Shapiro-Wilk test was used:

```
# Shapiro-Wilk test for normality of residuals
shapiro.test(residuals(final_model))
```

```
      Shapiro-Wilk normality test

data:  residuals(original_model)
W = 0.93678, p-value < 2.2e-16
```

The p-value obtained from the Shapiro-Wilk test is extremely small (p-value < 2.2e-16), indicating that the residuals from the final model are not normally distributed. This suggests a need for further model refinement or alternative approaches to address non-normality.

5.3 Results

Both the diagnostic plots and the Shapiro-Wilk test indicate that the residuals are not normally distributed, suggesting a need for further model refinement or alternative approaches. The non-normality of residuals suggests that the assumptions underlying the linear regression model may not be fully met. This could affect the validity of hypothesis tests on the coefficients and the accuracy of confidence intervals and predictions.

To address this issue, several next steps can be taken. First, consider transforming the dependent variable or predictors to achieve normality. Second, adding polynomial or interaction terms to the model can help capture non-linear relationships. Third, using robust regression techniques can mitigate the influence of outliers. Additionally, identifying and addressing influential data points can improve model performance.

Exploring alternative models, such as generalized linear models (GLMs) or non-parametric methods, may also be beneficial for better capturing the data distribution. By implementing

these steps, the model's performance and reliability can be enhanced, leading to more accurate and valid results.

6 Conclusion

In conclusion, our study adopted a rigorous methodology encompassing extensive data preparation, robust statistical analysis, and iterative model refinement to investigate the determinants of NBA player salaries. By employing advanced econometric techniques and robust statistical methods, we aimed to reveal insights into the factors driving salary disparities among NBA players.

Our study's approach addressed heteroskedasticity and integrated both current and historical performance metrics, enhancing our understanding of the economic dynamics within the NBA. The analysis, utilizing multiple linear regression with heteroskedasticity-robust standard errors, identified significant determinants of player compensation. Key performance indicators such as minutes played, points scored, assists, rebounds, and turnovers were found to be crucial predictors of salaries, underscoring their importance in evaluating player value.

The final model, which included a comprehensive set of current and lagged performance variables, explained approximately 52% of the variance in player salaries. These findings have practical implications for stakeholders in the NBA, aiding in strategic decision-making and promoting a more equitable compensation structure within the league. The insights gained can assist NBA team management in navigating player contracts and salary cap decisions, while also highlighting for players the importance of consistent performance over multiple seasons to achieve higher salaries.

Moreover, while performance metrics are critical, our study suggests that other factors, such as marketability, team budget constraints, and individual negotiations, also significantly influence salaries. Future research could extend our analysis by incorporating additional variables like player marketability and team financial health to provide a more comprehensive understanding of salary determination in professional basketball.

References

ESPN. (n.d.). NBA. https://www.espn.com/NBA/. Accessed 11 June 2024.