Use Season Stats to Predict NBA Players Salary

Group members:

Adela Yang,

Jack Lo,

Yuka Chen

Statement of the Problem

NBA players are on average the highest-paid athletes in the world, according to Statista.com.

Oftentimes sports players would seem to have major contracts with high annual salaries (some people would even think they should not get paid so much).

We want to find out whether the NBA players and their season total performance have a strong correlation.

Purpose

Discover	Discover which predictors variables are critical to the salaries of the NBA players
Regression	Use a multiple regression model to predict NBA players' salaries
Examine	Examine the difference between the predicted salaries and actual salaries

Data Section

➢Data source

Our season total performance and salary data sets were collected from Basketball Reference (https://www.basketball-reference.com/)

➢Data set

Combined NBA player performance and salary data by using **player ID and team**. During the regular season, some of the players will change their team, so they **have two different performance and salaries.**

Processing the Data(Data Cleaning)

Combined three datasets (NBA player salaries summary 1985-2018, season performance for both 2016-2017 and 2017-2018)

Joined the data sets based on the player ID and their team names.

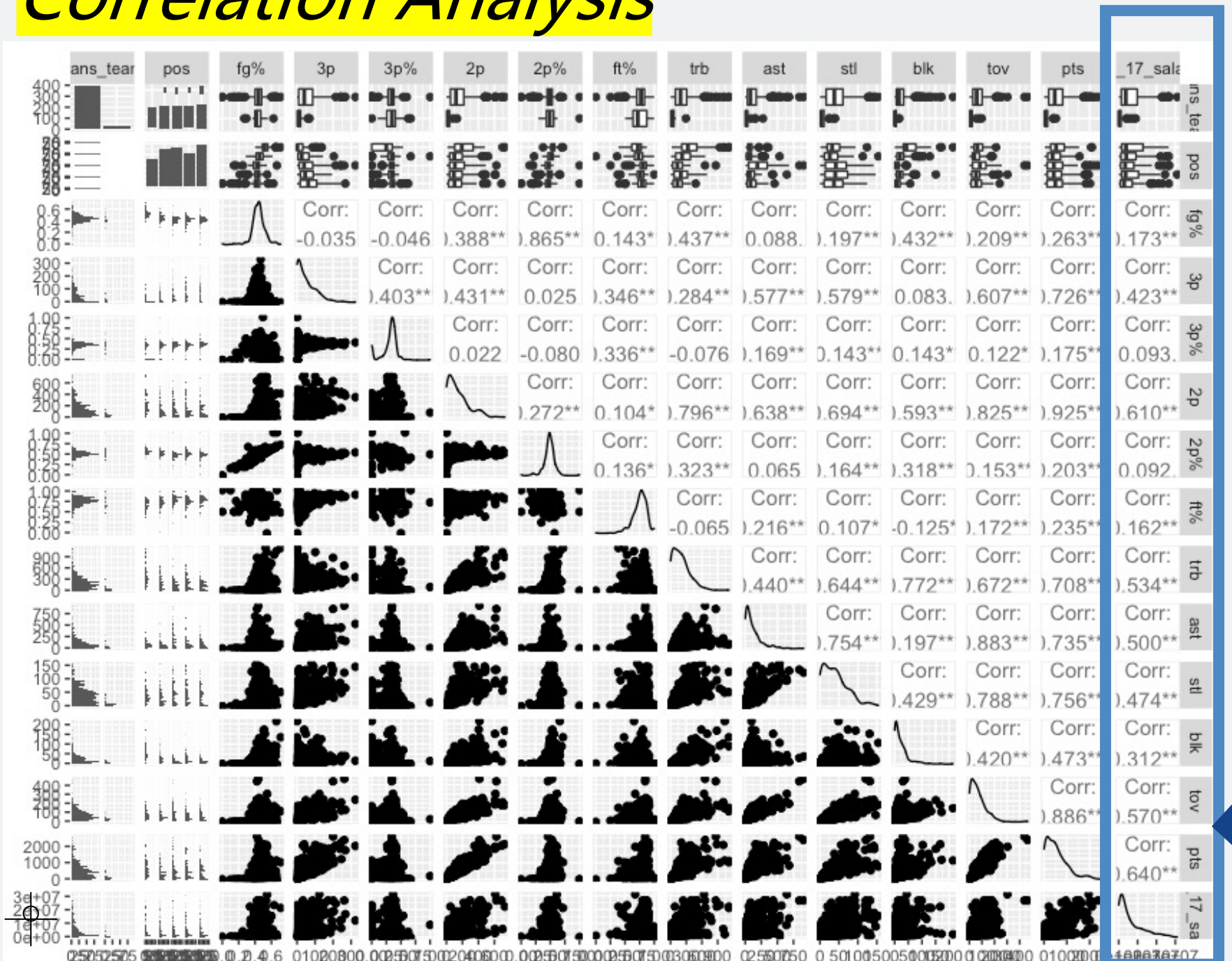
Removed 20 variables that seem to be either duplicated or are a combination of other variables. (i.e. trb = orb + drb).

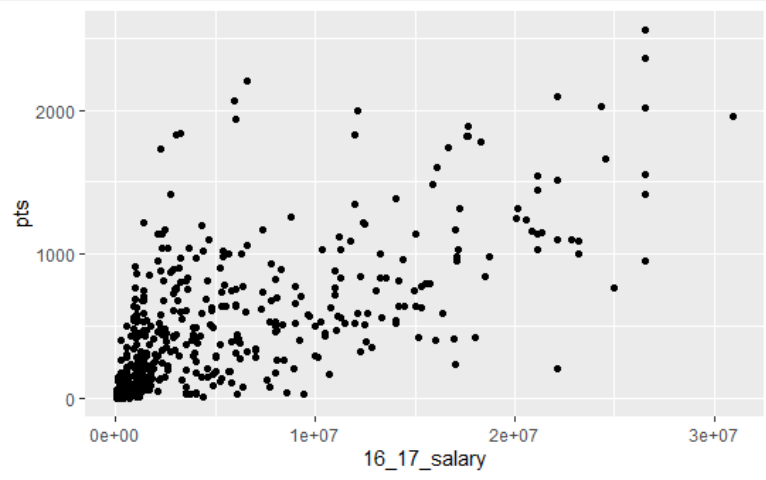
Dropped the person who has the total performances as they transferred the teams during the season in order to focus on their performance.

Model Building Process

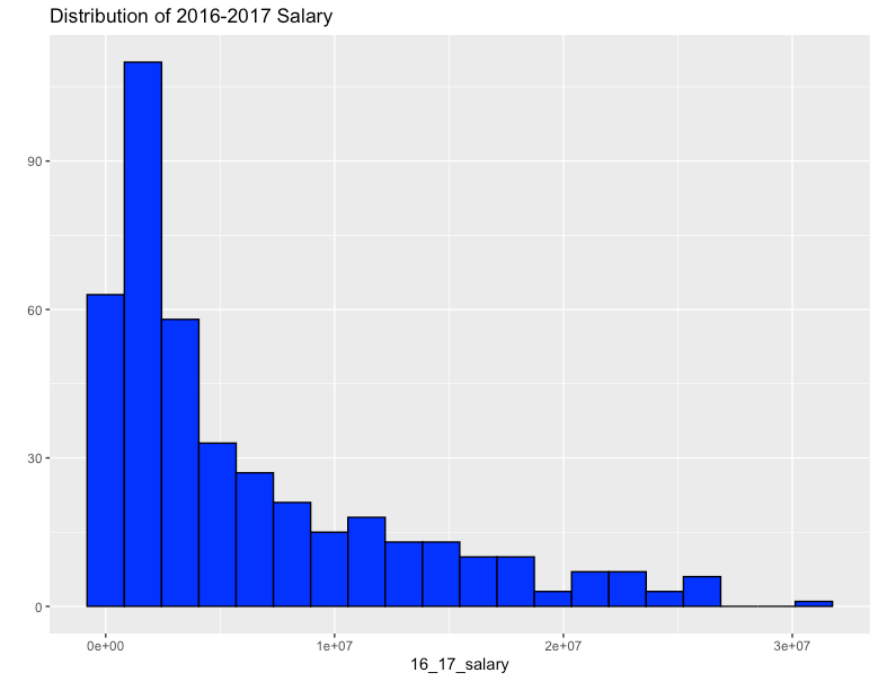
- The **response variable** is salary.
- The **predictor variables** include 2 categorical variables and 12 continuous variables.
- fg%, 3p, 3p%, 2p, 2p%, ft%, trb, ast, stl, blk, tov, pts are our continuous variables.
- pos and trans_team are categorical variables.

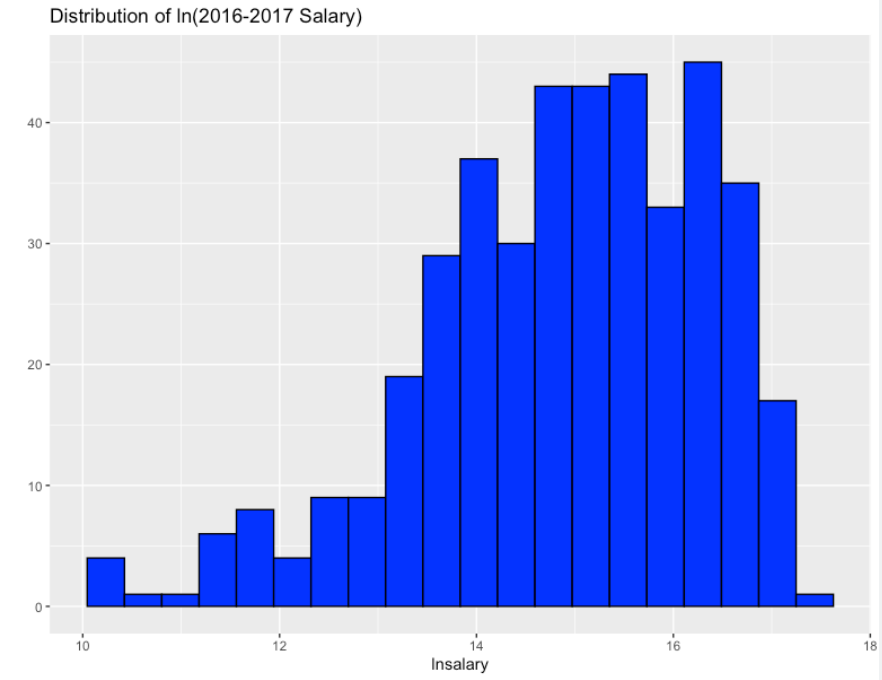
Correlation Analysis





Checking Y Predictable





- According to the histogram plot, we can see that the plot is **right-skewed**. To remedy the issue, we transformed the salary by taking **log**.
- We see possibly **a very slight left skew**, but it is **closer to symmetric** than without transform salary.

Introduce Three Comparing Model

1. Original Model(only continuous variables)

- Fitted model: $\widehat{ln(salary)} = 13.05 + 6.5 \cdot fg + 0.01 \cdot 3p$ - $0.028 \cdot 3p + 0.004 \cdot 2p$ - $5.02 \cdot 2p + 0.39 \cdot ft + 0.0024 \cdot trb + 0.0013 \cdot ast + 0.0019 \cdot stl$

2. Reduce Model Based on All Subset Models Methods

- Reduce model: $\widehat{ln(salary)} = 13.441 + 6.308 \cdot fg + 0.009 \cdot 3p + 0.002 \cdot 2p$ - $5.039 \cdot 2p + 0.002 \cdot trb$

3. Adding 5 Dummy Variables

- Reduce model: $\widehat{ln(salary)} = 12.277 + 6.357 \cdot fg + 0.009 \cdot 3p + 0.002 \cdot 2p$ -$4.949 \cdot 2p + 0.001 \cdot trb + 0.274 \cdot posC + 0.451 \cdot pos_PF + 0.313 \cdot posPG + 0.516 \cdot posSF + 0.955 \cdot transteamnone$

Comparing Model

For the full model without dummy variables (12 variables)

R^2_{adj} = 0.4096448

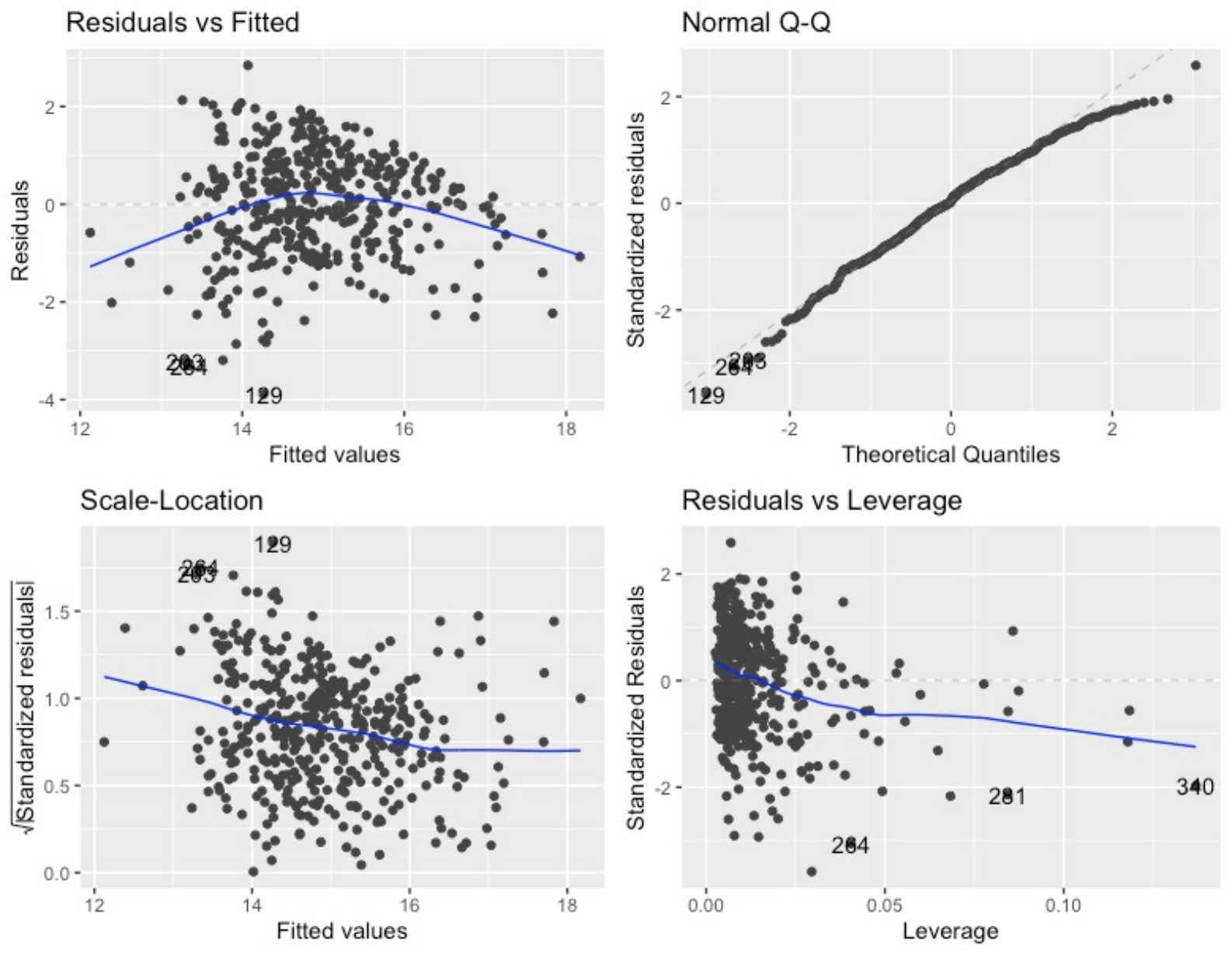
For the reduced model (5 variables) R^2_{adj} = 0.4095163

For the reduce model including dummy variables (10 variables)

R^2_{adj} = 0.4455106

Model Testing

Based on the fitted residual plot, it seems some multi-linear regression **assumptions are violated**.



Model Testing(2)

1. Residual Normality Test

```
        Shapiro-Wilk normality test

data:  reducemodel$residuals
W = 0.9868, p-value = 0.0007614
```

2. Residual Independence Test

```
 lag Autocorrelation D-W Statistic p-value
   1       0.1207212      1.756199   0.002
 Alternative hypothesis: rho != 0
```

3. Residual Variance Homogeneity Test

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 19.76537, Df = 1, p = 8.7555e-06
```

4. Testing for Non-Constant Variance Residual by Using Breusch-Pagan

```
 studentized Breusch-Pagan test

data:  dummy_model
BP = 28.039, df = 10, p-value = 0.001779
```

Heteroskedascity

Fixing Heteroskedasticity by Using WLS

- Since the residual plot show that the error look like uneven distribution, it violates the assumption of homogeneity of variance. As the result, it has heterodasticity issue, so we solve this violation by using WLS.

- **Final Model**

Test: $H_0 : \beta_1 = \beta_2 = ... = \beta_{10} = 0$ $H_1 : \beta_i \neq 0$ for some $i = 1, 2, ...10$

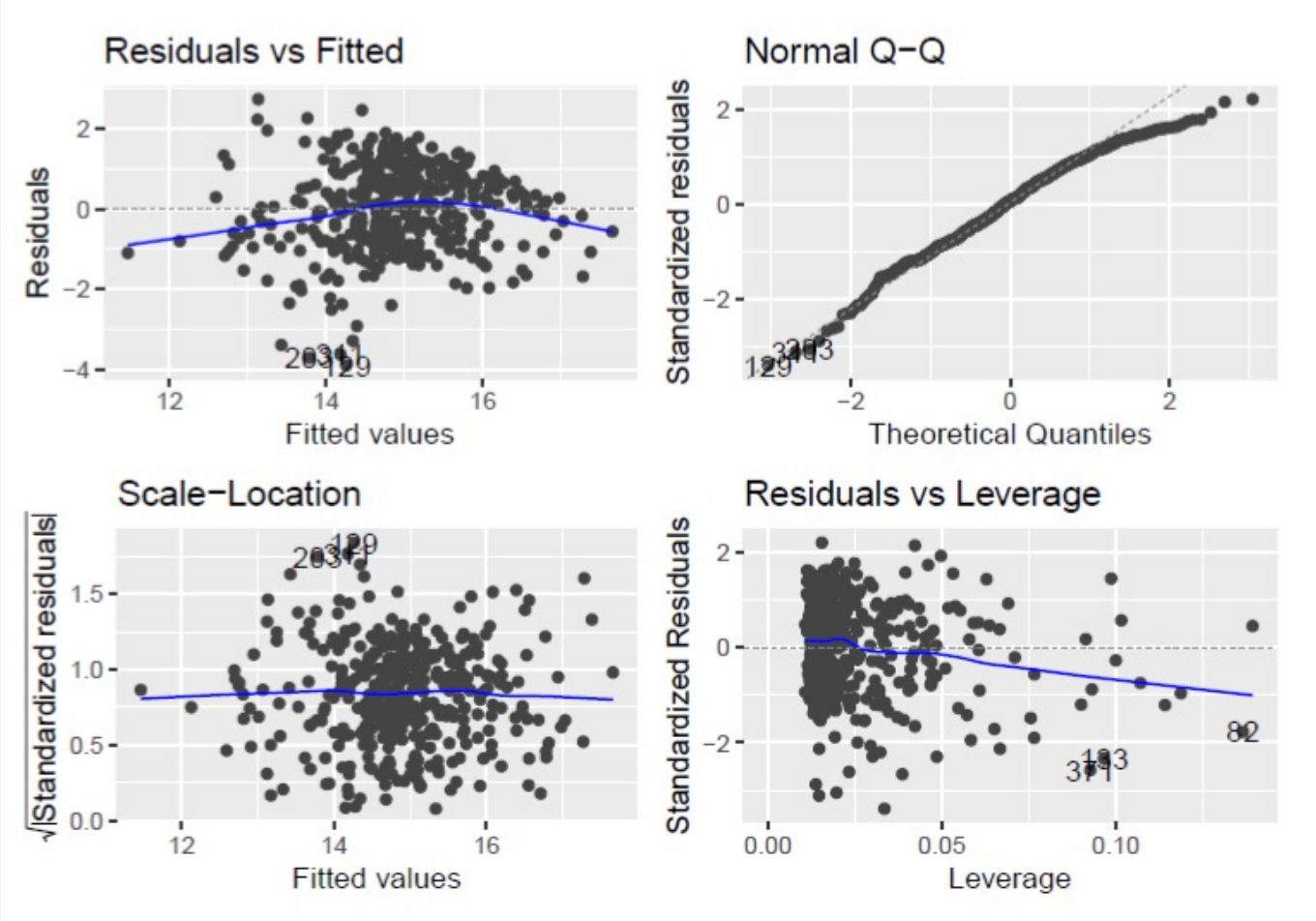
Test Statistic: * $F = \frac{MSR}{MSE} = 33.334$ * p-value < 0.00001

- Reduce model: $\widehat{ln(salary)} = 12.496 + 5.729 \cdot fg + 0.007 \cdot 3p + 0.002 \cdot 2p$ -4.648 $\cdot 2p + 0.001 \cdot trb +$ $0.198 \cdot posC + 0.412 \cdot posPF + 0.241 \cdot posPG + 0.424 \cdot posSF + 1.033 \cdot transteamnone$

For the final model $R^2_{adj} = 0.436743$.

Assumptions

The model has followed all assumptions, except residual normality test.



1. Residual Normality Test

```
 Shapiro-Wilk normality test

data:  lm_wls$residuals
W = 0.98346, p-value = 0.0001044
```

2. Residual Independence Test

```
lag Autocorrelation D-W Statistic p-value
  1       0.09488015       1.80693   0.036
Alternative hypothesis: rho != 0
```

3. Residual Variance Homogeneity Test

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.8191605, Df = 1, p = 0.36543
```

Evaluate Forecast Model

We use a forecasting model to determine how well it does in producing accurate forecasts, not how well it fits the historical model.

Measuring forecast accuracy, MAD=0.864, MSE=1.124.

From the result, both MAD and MSE are small and close to 0, actual values are very close to the predicted values.

It means that the prediction model we done is working well.

Inferences Based on the Model

After we build the multiple regression model, we can predict NBA players' salaries. The model has followed all assumptions except residual normality test.

Reduce model: $\widehat{ln(salary)} = 12.496 + 5.729 \cdot fg + 0.007 \cdot 3p + 0.002 \cdot 2p$ -4.648 $\cdot 2p + 0.001 \cdot trb + 0.198 \cdot posC + 0.412 \cdot posPF + 0.241 \cdot posPG + 0.424 \cdot posSF + 1.033 \cdot transteamnone$

Furthermore, the difference between predicted and actual salaries is small(MAD=0.864, MSE=1.124), which means that our model is great for applying.

Future Directions

- Since the data only offer the data that indicate players trans team during the regular season, isn’ t include off the season data that most of the players trans team time. For the further study, we recommend that add the resign the contrast or not, because the longer time interval model contrast effect maybe improve the results.
- The coefficient of 2p% is -4.949. (weird)
- We would like to collect

  1)seniority

  2)the points per game to help improve your results.
- We only use 16-17 salary to build our model, if we can add different year of salary data in our model, we could consider to use panel data analysis in our future research.

Group Work

- Project Concept Contribution: Jack
- Data collection: Adela
- Data cleaning: Yuka
- Model Building Process: Yuka, Adela
- Analysis result: Yuka, Adela, Jack
- PPT: Jack