MGTA 453 Business Analytics Case Study #3

Team 7: Zhengyu Jiang, Karina Mark, Yamini Nekkanti, Shihan Wang

10/17/2019

NBA Salary Regression

Problem

This case is to explore and understand how NBA player performance statistics from 2012 and 2013 can build an optimal model to predict a player's salary. We will discover which player performance metrics are better predictors of player salary.

Known Attributes

The dataset from Draft Express and Basketball Reference provides NBA player names, salaries, and performance predictors from 2012 and 2013 NBA seasons. The variables are provided below.

Player = Name of the basketball player Salary = Annual salary in $1000 log.Salary = log of Salary Age = Age per player FG = Field goals per game (includes 2-pointers and 3-pointers) RB = Rebounds per game AST = Assists per game STL = Steals per game BLK = Blocks per game PTS = Points per game (includes field goals and free throws)

Uncertainty of Data

We assumed the below conditions for our analysis:

Linearity: Dependent variable Salary/log.Salary depends linearly on the values of the independent variables, Age, FG, RB, AST, STL, BLK, PTS. Normality: Noise/Unaccounted differences obey a Normal distribution. Heteroscedasticity: Error terms are drawn from distributions with the same standard deviation. Any two independent predictors are not correlated. The predictors are calculated consistently between all players The predictors are independent among each other

Analysis

Based on the predictors given, we ran a linear regression comparing salary on the first six predictors (Age, FG, RB, AST, STL, and BLK). In this model, age and FG impact salary positively. The adjusted R-squared of the regression is 0.4841, an indicator of how the model explains the salary. Since it is in the middle of the range, the data isn't overfitting or underfitting. However, according to the residual plot, the residuals aren't randomly distributed around zero, which violates the assumption that the relationship is linear.

To fix the problem above, we then ran a linear regression of log-salary on the first six predictors (Age, FG, RB, AST, STL, and BLK). By executing the log-salary, the distribution of salary will be normally distributed. In this model, Age, FG, and RB influenced log salary positively. The residuals are also scattered against the fitted value and the salary distribution is close to normal distribution. However, this regression is still not a good model because the values are not as fitted as that of the previous regression (The adjusted R-squared is $0.446 < 0.484$).

To improve the model further, we included the PTS predictor to the log-salary regression. The adjusted R-squared is 0.455 resulting in a better fit. However, PTS causes FG to be insignificant with a high p-value and negative coefficient. Therefore, we infer that PTS explains the variation in player salary better than FG. FG, AST, STL, and BLK predictors are not significant in this regression. For the final salary prediction model, we remove the three unnecessary predictors.

Final Regression

Our optimal linear regression model comparing log-salaries on the Age, RB, BLK, and PTS predictors: Log Salary = 5.267 + 0.059*Age + 0.067*RB + 0.212*BLK + 0.089*PTS + errors

According to the residual scatter plot and histogram plot, the residuals are randomly distributed around 0, confirming the validity of this linear regression model. The adjusted R-squared is 0.456, slightly higher than the previous model. After standardizing coefficients, the standardized coefficient of PTS is the highest, indicating that points per game impacts player salary the most. It makes sense because the ultimate goal of basketball is to score higher points than your opponents and win the game. The ability to score more points in a game is most crucial in predicting the player's salary.

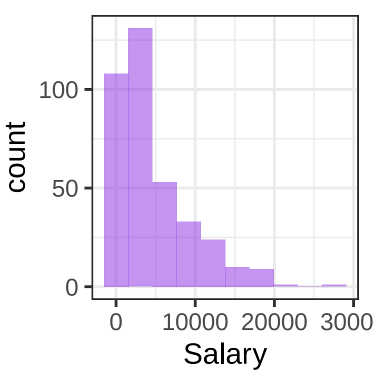
Conclusion

We concluded age, rebound ability, block ability, and points per game produced the best optimal model for NBA salaries. All predictors have positive coefficients, and an increase in any can affect player salaries positively. Assuming players skill level to be synonymous with the player's salary, we, therefore, can recommend our model to the teams/franchises and suggest that players who are around 25 years old and score high points should be selected for better results.

Appendix

Please refer to RMD file, Hot Hand Regression, for further coding details.

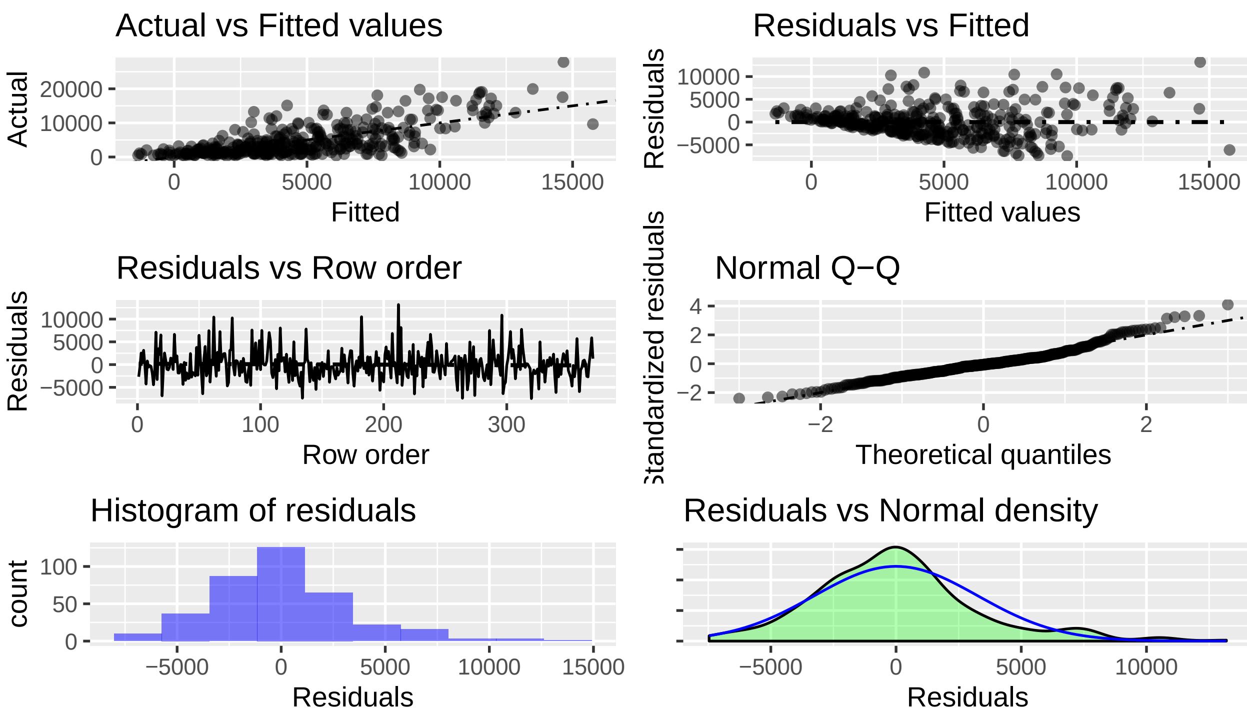
Histogram of Salary



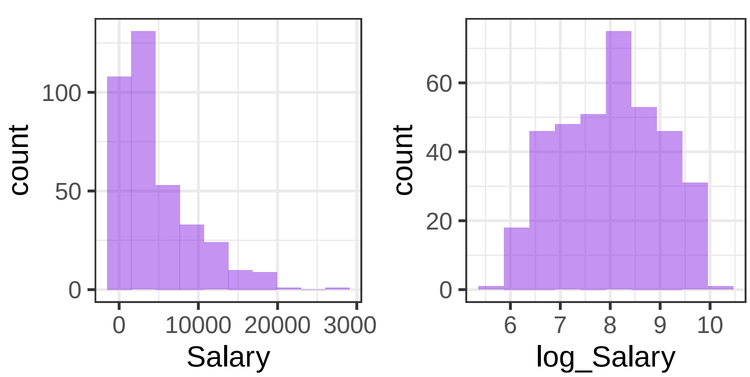
A Regression of Salary on the Predictors: Age, FG, RB, AST, STL, and BLK.

```
Linear regression (OLS)
Data     : nba_pgdata
Response variable    : Salary
Explanatory variables: Age, FG, RB, AST, STL, BLK
Null hyp.: the effect of x on Salary is zero
Alt. hyp.: the effect of x on Salary is not zero
```

```
                coefficient std.error t.value p.value
 (Intercept)      -8724.667  1131.833  -7.708  < .001 ***
 Age                312.092    39.520   7.897  < .001 ***
 FG                1156.982   153.881   7.519  < .001 ***
 RB                 223.312   117.004   1.909   0.057 .
 AST                280.498   146.504   1.915   0.056 .
 STL              -1064.070   613.248  -1.735   0.084 .
 BLK               1071.100   517.030   2.072   0.039 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.492,  Adjusted R-squared: 0.484
F-statistic: 58.632 df(6,363), p.value < .001
Nr obs: 370

Prediction error (RMSE):  3257.947
Residual st.dev   (RSD):  3289.21
```

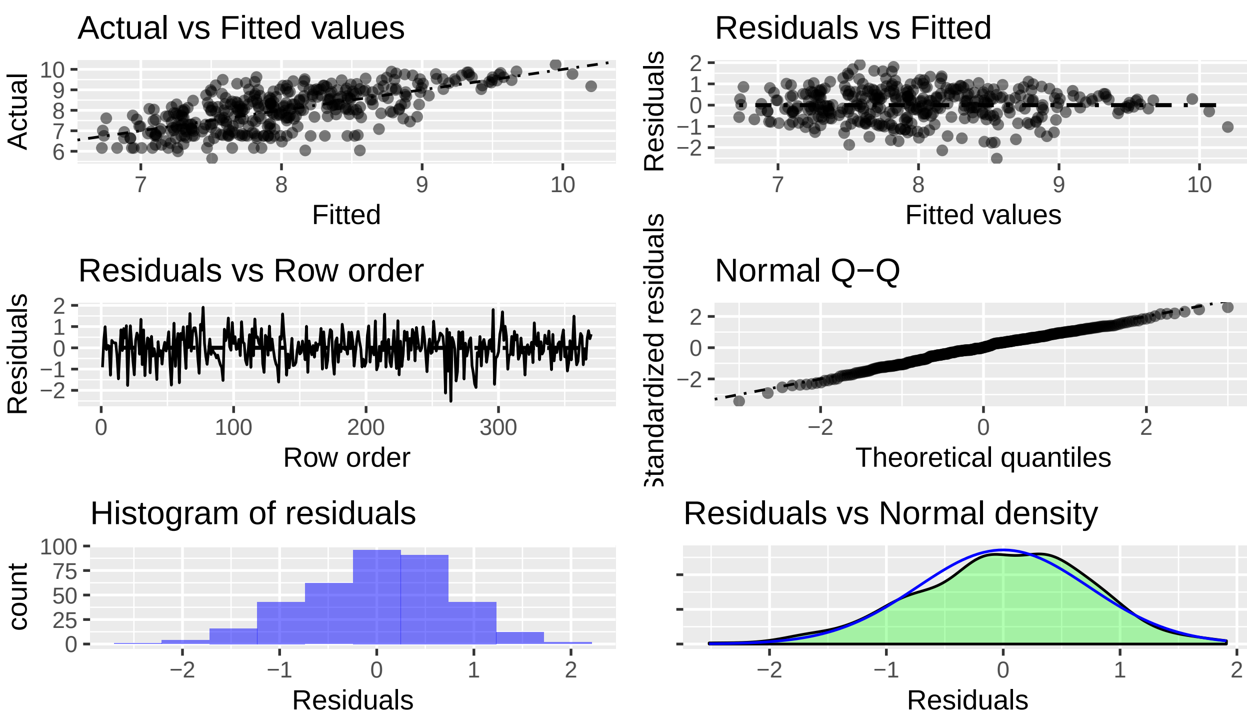


Histogram of log.Salary vs Salary



A Regression of log.Salary on the Predictors: Age, FG, RB, AST, STL, and BLK.

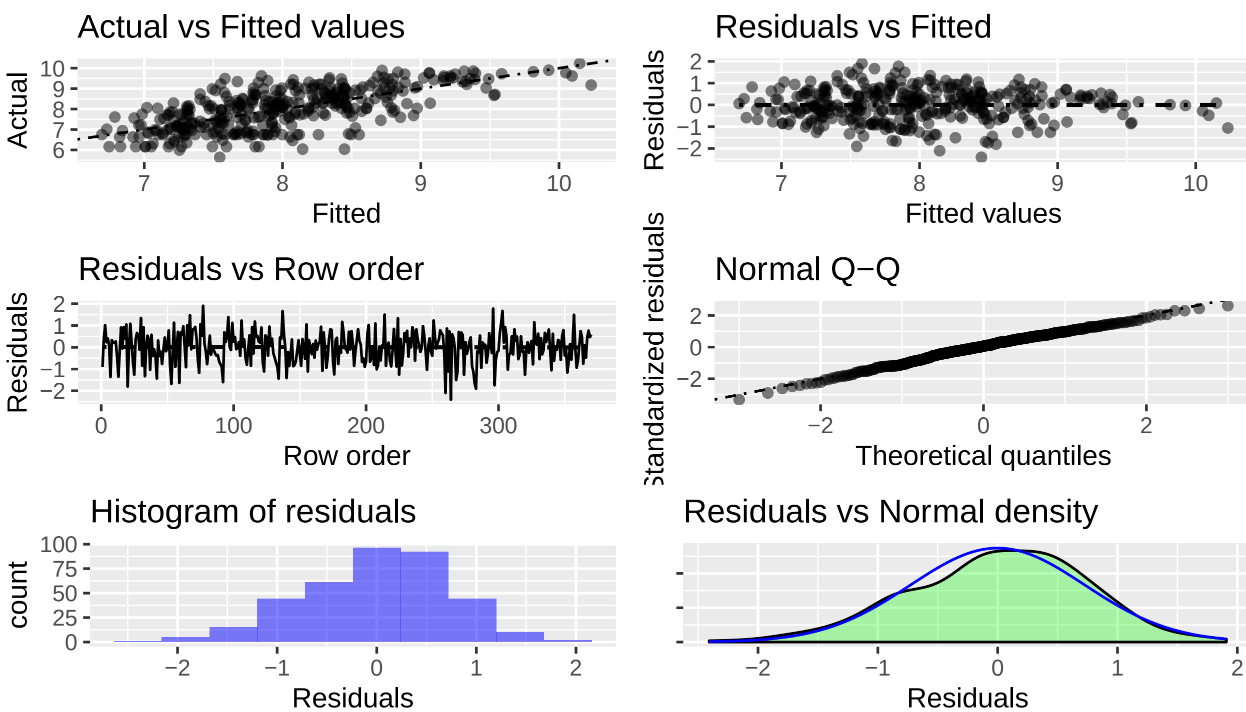
```
Linear regression (OLS)
Data      : nba_pgdata
Response variable    : log_Salary
Explanatory variables: Age, FG, RB, AST, STL, BLK
Null hyp.: the effect of x on log_Salary is zero
Alt. hyp.: the effect of x on log_Salary is not zero

             coefficient std.error t.value p.value
 (Intercept)       5.302     0.255  20.803  < .001 ***
 Age               0.058     0.009   6.487  < .001 ***
 FG                0.208     0.035   5.994  < .001 ***
 RB                0.067     0.026   2.533   0.012 *
 AST               0.046     0.033   1.403   0.162
 STL               0.002     0.138   0.013   0.990
 BLK               0.210     0.116   1.807   0.072 .

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.455,  Adjusted R-squared: 0.446
F-statistic: 50.599 df(6,363), p.value < .001
Nr obs: 370

Prediction error (RMSE):  0.734
Residual st.dev   (RSD):  0.741
```



A Regression of log.Salary on the Predictors: Age, FG, RB, AST, STL, BLK, and PTS.

```
Linear regression (OLS)
Data      : nba_pgdata
Response variable    : log_Salary
```

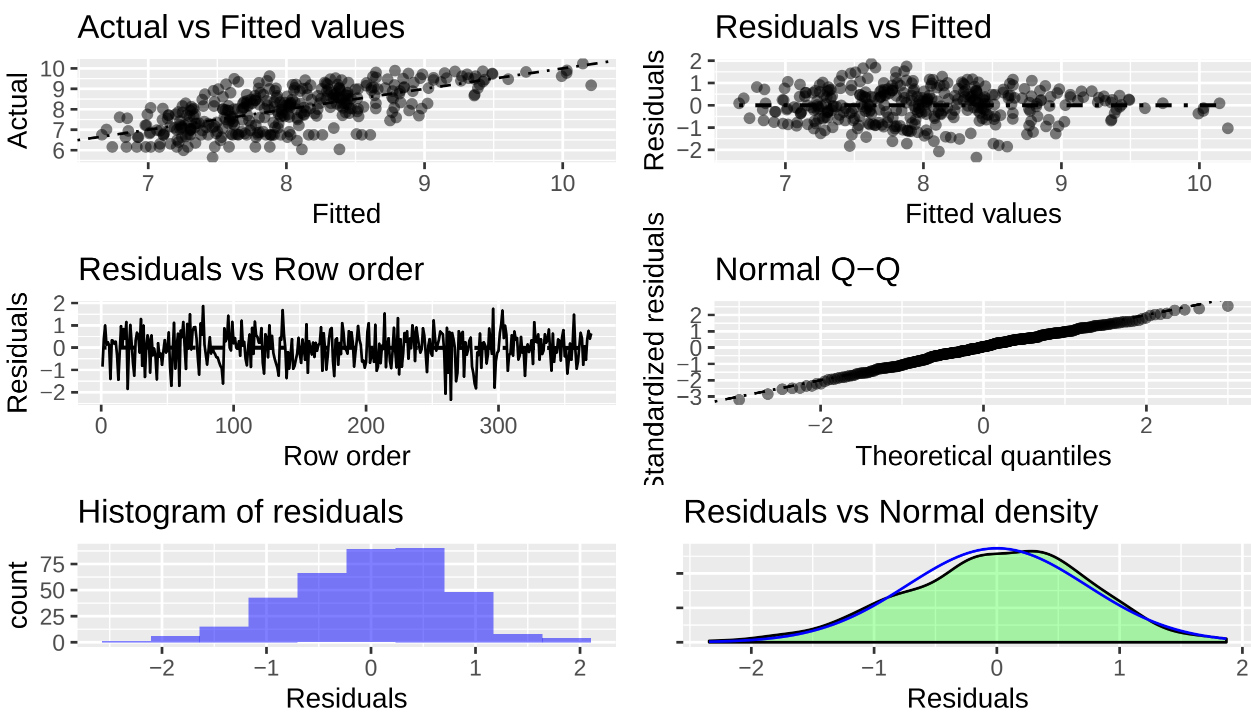
```
Explanatory variables: Age, FG, RB, AST, STL, BLK, PTS
Null hyp.: the effect of x on log_Salary is zero
Alt. hyp.: the effect of x on log_Salary is not zero

             coefficient std.error t.value p.value
 (Intercept)       5.297     0.253  20.944  < .001 ***
 Age               0.058     0.009   6.524  < .001 ***
 FG               -0.117     0.130  -0.901   0.368
 RB                0.083     0.027   3.081   0.002 **
 AST               0.044     0.033   1.347   0.179
 STL              -0.052     0.139  -0.378   0.706
 BLK               0.251     0.117   2.153   0.032 *
 PTS               0.120     0.046   2.592   0.010 **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.465,  Adjusted R-squared: 0.455
F-statistic: 45.013 df(7,362), p.value < .001
Nr obs: 370

Prediction error (RMSE):  0.727
Residual st.dev   (RSD):  0.735
```



A Regression of log.Salary on the Predictors: Age, RB, BLK, and PTS.

```
Linear regression (OLS)
Data     : nba_pgdata
Response variable    : log_Salary
Explanatory variables: Age, RB, BLK, PTS
Null hyp.: the effect of x on log_Salary is zero
Alt. hyp.: the effect of x on log_Salary is not zero
```

```
               coefficient std.error t.value p.value
 (Intercept)         5.267     0.249  21.117  < .001 ***
 Age                 0.059     0.009   6.734  < .001 ***
 RB                  0.067     0.025   2.754   0.006 **
 BLK                 0.212     0.114   1.868   0.063 .
 PTS                 0.089     0.009  10.298  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.462,  Adjusted R-squared: 0.456
F-statistic: 78.235 df(4,365), p.value < .001
Nr obs: 370

Prediction error (RMSE):  0.729
Residual st.dev   (RSD):  0.734
```



A Standardized Regression of Salary on the Predictors: Age, RB, BLK, and PTS.

```
Linear regression (OLS)
Data     : nba_pgdata
Response variable    : log_Salary
Explanatory variables: Age, RB, BLK, PTS
Null hyp.: the effect of x on log_Salary is zero
Alt. hyp.: the effect of x on log_Salary is not zero
**Standardized coefficients shown (2 X SD)**

               coefficient std.error t.value p.value
 (Intercept)        -0.000     0.019  -0.000   1.000
 Age                 0.259     0.039   6.734  < .001 ***
 RB                  0.168     0.061   2.754   0.006 **
```

```
 BLK                 0.100      0.053    1.868    0.063 .
 PTS                 0.478      0.046   10.298   < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.462,  Adjusted R-squared: 0.456
F-statistic: 78.235 df(4,365), p.value < .001
Nr obs: 370

Prediction error (RMSE):  0.366
Residual st.dev   (RSD):  0.369
```