LINEAR REGRESSION MODEL FOR BASKETBALL ANALYSIS

HongJui Shen &WenYi Shi†

†University of Denison

Introduction

Our data is from the.basketball-reference website, And the data we selected is the 2019-20 Milwaukee Bucks', Los Angeles Lakers', Houston Rockets' Roster and Stats. This set of data basically covers the names of the players of the three teams in 2019-2020, the number of games played, and various game information. We chose this data because we want to use these data to build a model, that is, if you give us information about a new team with relevant information, we can use our modeling to predict the next scoring situation of this team . And the mainly analysis is through PCA and linear regression model.

Data

1. Possible players included in this study are limited to those who are in the team between 2019-2020
2. The main data covers the specific information of the players and their scores on the field and the performance of the field
3. According to specific information, the data contains players who are not in the game. In order to prevent this information from affecting the model, we have processed the data to remove NA to exclude this information.

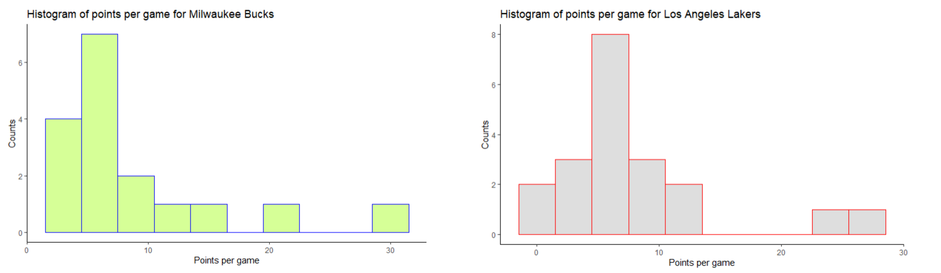

Fig. 1: Points per game for each teams' members and teams.

Principal Component Analysis

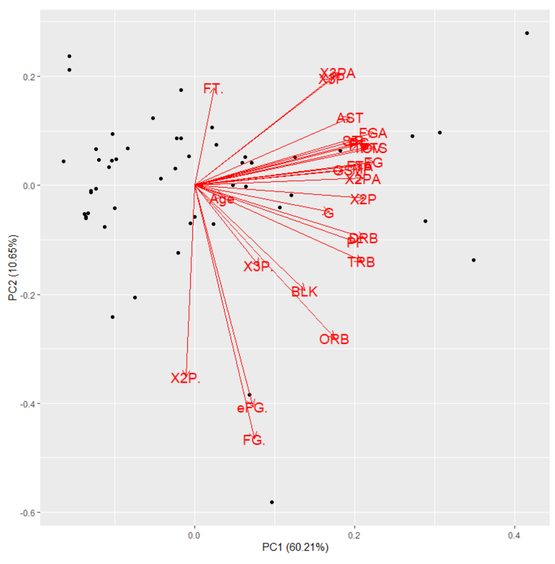

Fig. 2: PCA

Principal component analysis is a technique which can help us dealing with a large data set, in this graph. Each points means the basketball players from Milwaukee Bucks, Los Angeles Lakers, Houston Rockets. PC1 accounts for the variability in most of the per-game statistics. These are the main numbers which determine how good a player is and they do not depend on games played as much. In contrast, PC2 accounts for the variability in statistics which are aggregate measures. Measures like Free Throws, 3-Point Field Goal Attempts, Assists etc. all depend on the number of games played. Since the number of games played widely varies (due to the fact that some people only play 1 games and some play 67 games) the aggregate measures will have the most variance.

linear regression model

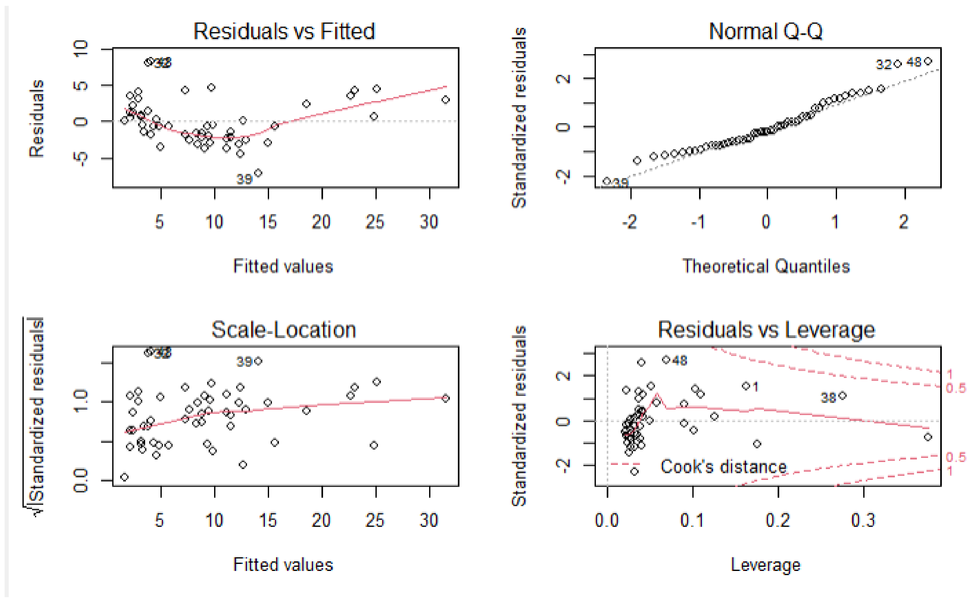

Fig. 3: Linear regression residual graphs

The regression formula is PG=9.2510+1.6770(PC1)+0.6367(PC2), and the data we used is PCATestNA. Roughly 82% of the variance found in the response variable can be explained by the predictor variable. The the Residual Standard Error is 3.167, and the degree of freedom is 48. Also the F-statistic is 116.8 which is relatively larger than 2 given the size of our data. Thus, there may be a relationship between predictor and response variables. And the p-value of this model is 2.2e-16, which is help to reject the Ho. As the p-value for PC1 and PC2 is less than 0.05, which also including the Signif. codes for the response variable, these two will contribute a huge influence to the response variable.

```
Call:
lm(formula = PG ~ PC1 + PC2, data = PCATestNA)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1093 -2.1792 -0.6127  1.8375  8.1896

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2510     0.4435   20.86   <2e-16 ***
PC1           1.6770     0.1111   15.10   <2e-16 ***
PC2           0.6367     0.2642    2.41   0.0199 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.167 on 48 degrees of freedom
Multiple R-squared:  0.8296,    Adjusted R-squared:  0.8225
F-statistic: 116.8 on 2 and 48 DF,  p-value: < 2.2e-16

PC1 PC2
  1   1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -4.724  -3.168  -0.607   0.000   1.833  11.959
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-7.0455 -0.6705  0.3775  0.0000  0.8173  3.3754
```

Fig. 4: Linear regression model

Conclusion and Future topic

As the linear regression we get for basketball, we determined that the multiple R-squared value of 0.8296 was good, even though that there is some small problem in the linear regression model graph. And we concluded that our model still lacks certain data, and at the same time we do not take into account the differences between players of different games played. But if there are some possible future team data information, we can add prediction to predict the average points per game of this new team.