MULTIPLE LINEAR REGRESSION ANALYSIS OF THE 2021-2022 LAKERS SEASON

Are effective field goal percentage, free throw percentage, games played, personal fouls, steals, and blocks are good indicators of points per game performance during the season?

Azuka Atum

Multiple Linear Regression Analysis of the 2021-2022 Lakers Season

Are effective field goal percentage, free throw percentage, games played, personal fouls, steals, and blocks are good indicators of points per game performance during the season?

Azuka Atum
Department of Statistics and Biostatistics, California State University, East Bay, Hayward, CA 94542

Introduction

As the value of sports teams increases and technical analytics advances, many find that basketball analytics is a burgeoning field of research for application of these avenues. From sports betting to predicting player performance using aggregate data provided by many basketball sources such as NBA.com or Basketball-reference.com, many statisticians and analysts are clamoring to use modeling to their advantage.

The goal of this MLR analysis of the 2021-2022 Lakers season is to assess whether or not effective field goal percentage, free throw percentage, games played, personal fouls, steals, and blocks are good indicators of points per game performance during the season.

Data Description

The data for this analysis originates from an amalgamation of data from basketball-reference and nba.com. At first, I opted to use the full average stat per game from basketball-reference, however the website wasn't updated recently enough for proper analysis, so some players had NA or missing values for important stats. Other discrepancies such as players being traded midseason resulted in some inaccuracies. I went to NBA.com as they are the official source for statistical data. I opted to use basketball-reference because downloading data was easier than with NBA stats, which would require some complex web-scraper. NBA.com also included additional stats I wasn't interested in, and also did not contain an effective field-goal percentage column like basketball-reference.com did. So I had to hand-calculate that data from the data available.

For my analysis, my predictor variables were:

1) **effective field goal percentage**, or eFG%, which is defined as the measured field goal percentage adjusting for made 3-point field goals being 1.5 times more valuable than 2-point made field goals[1], or

$$\frac{FGM + (1.5 * 3PM)}{FGA}$$

Where FGM is field goals made, 3PM is three-points made, and FGA is field goal attempts.

2) **free throw percentage**, or FT%, which is defined as the percentage of free throw attempts that a player or team has made [2], or

$$\frac{FT}{FTA}$$

3) **games played**, G, which is the number of games a player or team played where a specified criteria occurred[3],
4) **personal fouls**, PF, which is the number of personal fouls a player or team committed[4]
5) **steals**, or STL , which is Number of times a defensive player or team takes the ball from a player on offense, causing a turnover[5] , and
6) **blocks**, or BLK, defined as when an offensive player attempts a shot, and the defense player tips the ball, blocking their chance to score[6].

My response variable is points per game, or PTS.G.

[1] https://www.nba.com/stats/help/glossary/#efgpct
[2] https://www.nba.com/stats/help/glossary/#ftpct
[3] https://www.nba.com/stats/help/glossary/#g
[4] https://www.nba.com/stats/help/glossary/#pf
[5] https://www.nba.com/stats/help/glossary/#stl
[6] https://www.nba.com/stats/help/glossary/#blk

The data itself is 21 rows by 22 columns. It's averaged over an 82 game period for the whole team, which is a typical number of games for the regular season.

Each dataset adapted from either website had too many columns to work with in R, specifically there were too many predictors so it was hard to draw a reasonable conclusion. For example running a line of code assessing all predictors gave NaN errors and aliasing results in the summary statistics shown below. This means that multiple predictors are perfectly correlated and cannot be adequately analyzed. This also means stepwise model selection using **step()** and variance inflation factor computation using **vif()** could not be performed because AIC is negative infinity for the model. Also note that more predictors means a higher R-squared value, which impacts accuracy of results.

```
Call:
lm(formula = PTS.G ~ ., data = lakdf)

Residuals:
ALL 19 residuals are 0: no residual degrees of freedom!

Coefficients: (7 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.863288        NaN     NaN      NaN
Age            0.017959        NaN     NaN      NaN
G              0.008906        NaN     NaN      NaN
GS            -0.001952        NaN     NaN      NaN
MP             0.016912        NaN     NaN      NaN
FG             3.802445        NaN     NaN      NaN
FGA           -1.834816        NaN     NaN      NaN
FG_per        11.764389        NaN     NaN      NaN
thrpt         -1.096963        NaN     NaN      NaN
thrptat        1.992149        NaN     NaN      NaN
thrptat_per   -5.534987        NaN     NaN      NaN
twopt         -1.703139        NaN     NaN      NaN
twoptatt       1.558774        NaN     NaN      NaN
twopt_per     -6.300286        NaN     NaN      NaN
effFG_per      1.936462        NaN     NaN      NaN
FT             0.508897        NaN     NaN      NaN
freethrowAtt   0.725891        NaN     NaN      NaN
```

Figure 1.Summary statistic for full model.

So in order to work around this, I was able to code a heat-map correlation matrix of all the predictors to assess how they relate to each other.

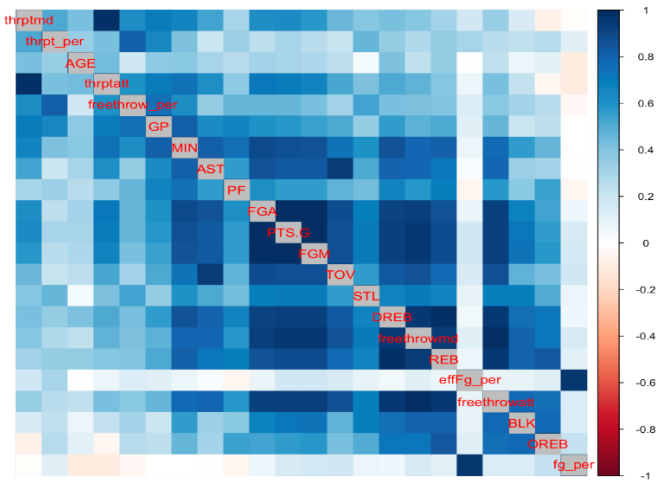


Figure 2. Correlation matrix for all the predictors and response variable. Darker squares indicate a correlation coefficient close to 1 or -1. Lighter squares indicate a correlation coefficient close to 0.

Methods and Results

Because multiple variables are highly correlated with each other, I had to do variable elimination for all predictors that had an R-squared higher than 0.80.

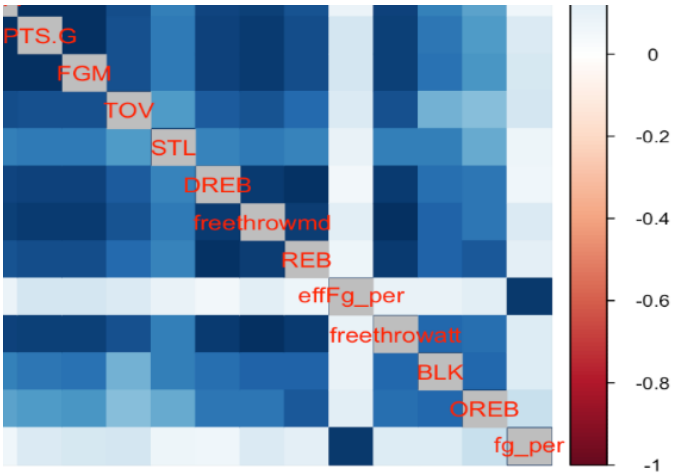


Figure 3.Trimmed-down correlation matrix with relevant predictors.

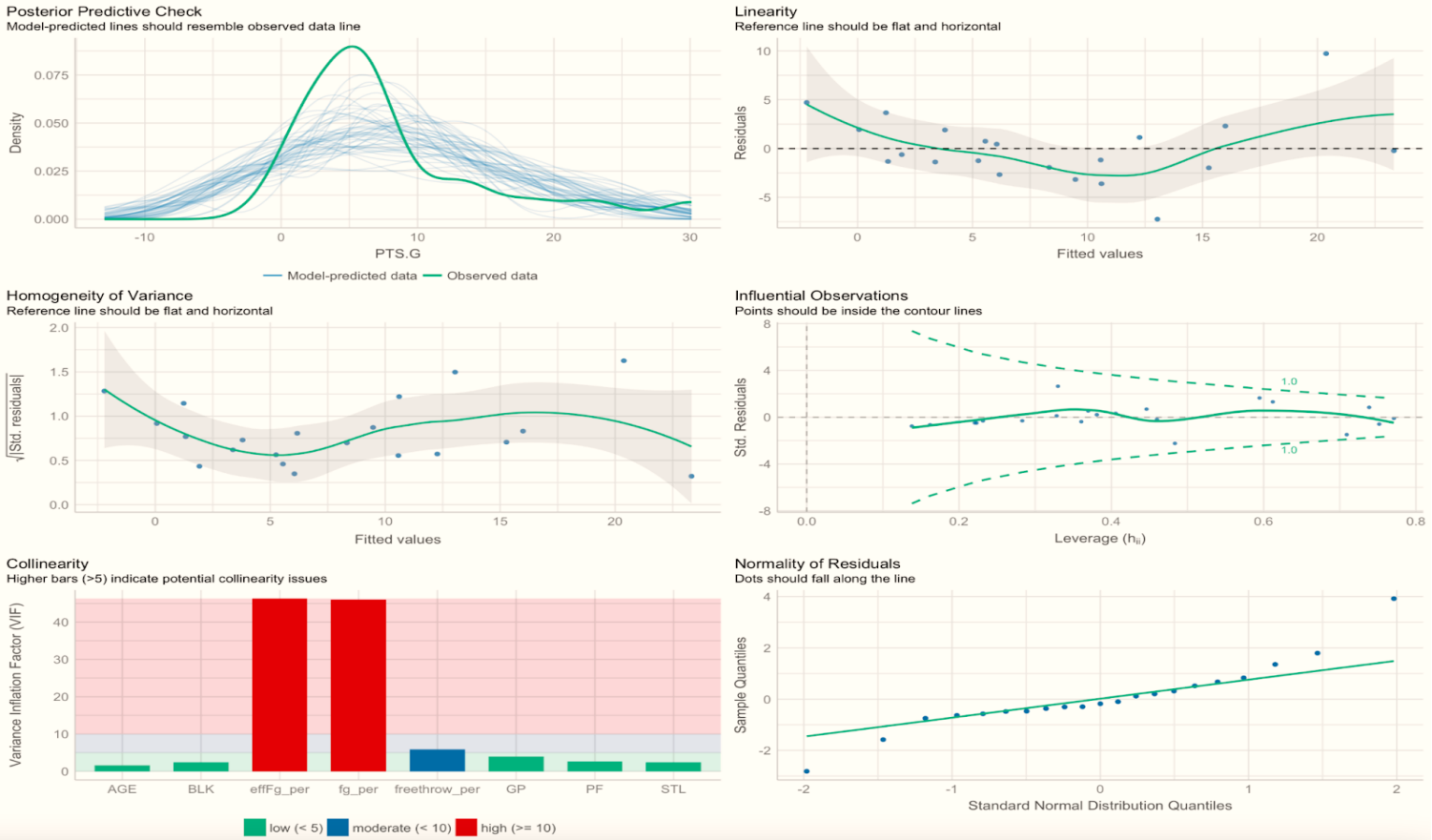


Figure 4. Six graph performance check of remaining predictors, with data untransformed. Variance inflation factor graph shows high degree of collinearity between eFG% and FG%. Other predictors are minimally correlated.

As you can see, preliminary performance checks show there is an issue of collinearity among the final predictors, so a decision was made to remove FG%. It was a reasonable decision to make because eFG% already accounts for FG% in its formula, as in, eFG% is the sum of the ratio of FGM to FGA and the ratio of 1.5*3PM and FGA. Because eFG% contains FG%, the predictors are essentially double counted and highly correlated because they are the same measure.

Also, three-point shots are more impactful to the game as two point shots, and they are worth 1.5 times more than a two-point field goal. Additionally, FGA includes three-point shots attempted and made, thus, eFG% is a simpler predictor combining all the relevant shooting predictors, while also comprising of a scaling component between two-point field goals and three-point field goals.

```
Call:
lm(formula = PTS.G ~ . - MIN - FGM - FGA - thrptmd - thrptatt -
    thrpt_per - freethrowmd - freethrowatt - OREB - DREB - REB -
    AST - TOV, data = lakdf)

Residuals:
   Min     1Q Median     3Q    Max
-7.238 -1.934 -0.621  1.901  9.723

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.295218   7.147021  -1.021   0.3275
AGE            0.163362   0.221964   0.736   0.4759
GP             0.158186   0.077733   2.035   0.0646 .
fg_per         0.025950   0.358873   0.072   0.9435
effFg_per      0.852237  35.961443   0.024   0.9815
freethrow_per -0.069985   0.078450  -0.892   0.3899
STL            7.295829   4.233209   1.723   0.1104
BLK            6.252582   2.911778   2.147   0.0529 .
PF            -0.001557   1.856400  -0.001   0.9993
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.492 on 12 degrees of freedom
Multiple R-squared:  0.7912,    Adjusted R-squared:  0.652
F-statistic: 5.684 on 8 and 12 DF,  p-value: 0.003865
```

Figure 5. Summary statistics after predictors were removed. R-squared is 0.79, indicating decent correlation with points per game.

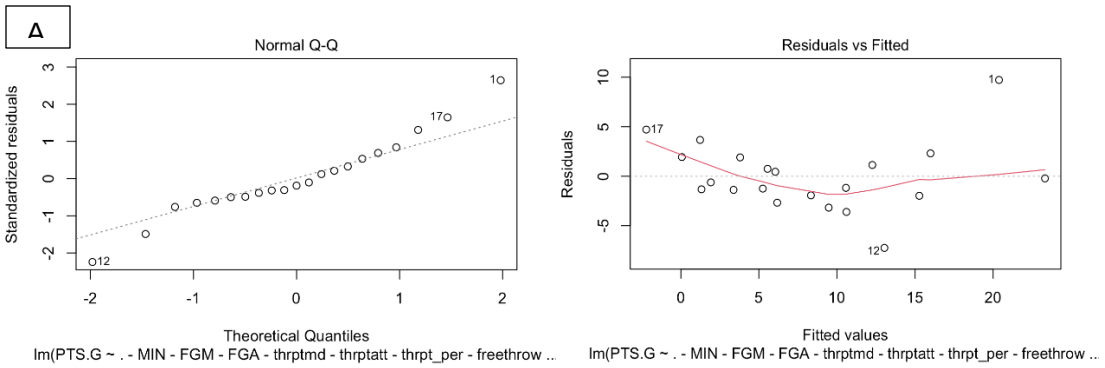
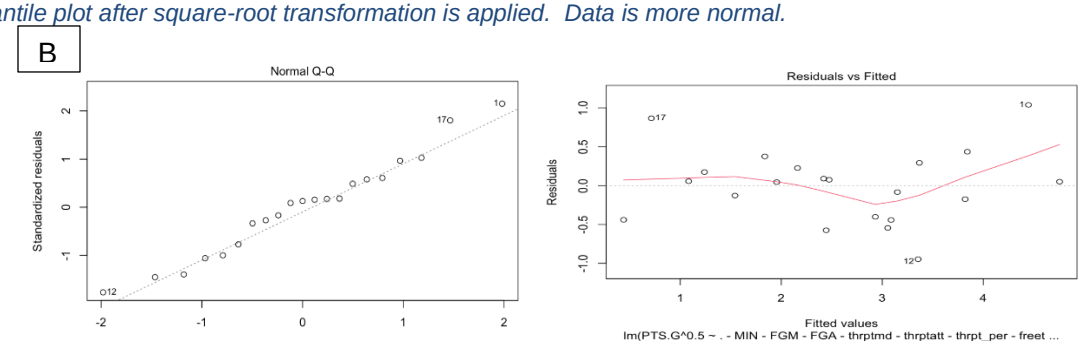


Figure 6[A]. Normal Quantile-Quantile plot for data that did not have FG% removed. Residuals are a little bit not constant. [B]Normal Quantile-Quantile plot after square-root transformation is applied. Data is more normal.



Here the normal quantile-quantile plot shows some data points that lie outside the fit line, indicating there are some outliers that must be resolved. It was decided that outliers would not be removed, because the main outlier indicated in the plot corresponds to LeBron James, the player with the highest playing stats of any player on the team. Basketball teams in general tend to foster outliers in their team set up because they are generally built around outlier players, as they tend to be the best players on the team. So removing LeBron James would not give us a meaningful result.

A 0.5 power transformation was applied after FG% was removed. Additionally an F test was performed to make sure that removing FG% did not impact the model in a significant way. Final R-squared was 0.86, indicating moderate correlation.

```
Call:
lm(formula = PTS.G^0.5 ~ . - MIN - FGM - FGA - thrptmd - thrptatt -
    thrpt_per - freethrowmd - freethrowatt - OREB - DREB - REB -
    AST - TOV - fg_per, data = lakdf)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94654 -0.40128  0.05076  0.22674  1.03918

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7028908  0.8688391  -0.809   0.4331
AGE           0.0197069  0.0251795   0.783   0.4479
GP            0.0195850  0.0101543   1.929   0.0759 .
effFg_per     1.4698015  0.7151276   2.055   0.0605 .
freethrow_per -0.0009501  0.0073011  -0.130   0.8985
STL           1.0321481  0.5526960   1.867   0.0845 .
BLK           0.7605361  0.3479689   2.186   0.0477 *
PF            0.1768826  0.2225320   0.795   0.4410
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5868 on 13 degrees of freedom
Multiple R-squared:  0.8615,    Adjusted R-squared:  0.787
F-statistic: 11.55 on 7 and 13 DF,  p-value: 0.0001113
```

Figure 7. Final model with transformation applied. R-squared is 0.86.

The final model is given by the following:

$$Y^{0.5}(Points\ Per\ Game) = -0.702 + 0.019(\text{Age}) + 0.019(\text{Games Played}) + 1.47\ (\text{eFG\%}) - 0.001(FT\%) + 1.032(Steals) + 0.760(Blocks) + 0.177(Personal\ Fouls)$$

Where every change corresponds to:

- Every one unit increase in AGE is associated with to a 0.019 change in the square root of PPG.
- A one unit increase in Games played is associated with a 0.019 change in square root of PPG.
- A one unit increase in effective Field goal percentage is associated with a 1.47 change in square root of PPG.
- A one unit increase in free throw percentage is associated with a -0.001 change in square root of PPG.
- A one unit increase in steals is associated with a 1.032 change in square root of PPG.
- A one unit change in blocks is associate with a 0.760 change in square root of PPG.
- A one unit change in personal fouls is associated with a 0.177 change in square root of PPG.

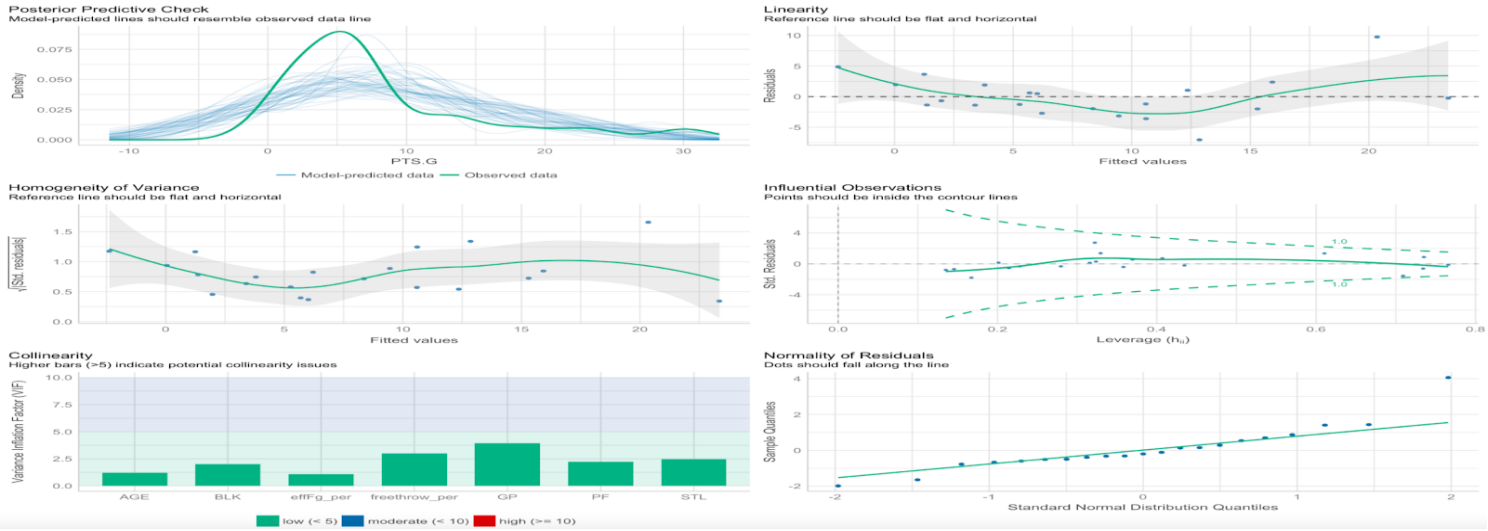


Figure 8. Performance model check for final model. Variance inflation factor plot shows there is no or minimal collinearity issues.

Conclusion

We can conclude that the above predictors are effective field goal percentage, free throw percentage, games played, personal fouls, steals, and blocks are good indicators of points per game performance during the 2021-2022 season for the Lakers.

To put things into context, a player that was 26 years old, that player 82 games this season, who had an eFG% of 0.50, a FT% of 0.60, averaged 1.7 steals a game, averaged 1 block per game, and had 2 personal fouls per game would average 2.22 points per game. The player closest to this statline is Jemerrio Jones. However, Jemerrio only played 2% of games this season. No one on the Lakers roster in 21-22 played all 82 games, only Russell Westbrook who played 75 games or 94% with an average PPG of 18.3 came close. On average, Lakers players played 36 games in the 21-22 season. So this indicates to us that if a player played more than the whole Lakers on average, they would outperform the model.

One way the model can be improve is by obtaining more data points, and also maybe normalizing the Lakers average stats against the whole league, or perhaps against previous Lakers seasons. Another important potential limiter is the fact that there was a pandemic which caused a shift in season playing times. A lot of players between 2020 and 2022 suffered more injuries than normal because they had to play an extended schedule to make up for lost time, which led to less time spent conditioning and/or resting between games.

Links: GitHub Repository: See link:.
https://github.com/lebronable/MLRLakers2