Regression Analysis on NBA Claims

Brandon Chan, Brayden Van De Wynckel

2024-11-13

Objective

Many older NBA fans like to make claims about how the league was different back in their day. They often criticize today's game for the high frequency of threes attempted. The aim of this analysis is to see how the game has changed, and why these changes have taken place. # Data Description The data is from basketballreference.com. They offer CSV files that are free to download on their website. This particular dataset includes many common stats (ex. points, offensive rating), taken as the league average of per-game stats of teams for a given season. For example the PTS stat in the 2023-24 season represents that average points per game scored by a team in a single game.

```
## Rows: 78 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): Season, Lg, Ht
## dbl (29): Rk, Age, Wt, G, FG, FGA, 3P, 3PA, FT, FTA, ORB, DRB, TRB, AST, STL...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 74 x 30
##    Season    Age Ht       Wt     G    FG   FGA  `3P` `3PA`    FT   FTA   ORB
##    <chr>   <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 2023-24  26.4 6-7     216  1230  42.6  89.7  13    35.4  17.2  21.9  10.7
##  2 2022-23  26.1 6-6     216  1230  42    88.4  12.4  34.2  18.4  23.6  10.4
##  3 2021-22  26.1 6-6     215  1230  41.1  89.2  12.6  35.6  17.1  22.1  10.5
##  4 2020-21  26.1 6-6     217  1080  41.3  88.6  12.7  34.7  17    21.9   9.9
##  5 2019-20  26.1 6-6     218  1059  40.4  87.9  12.1  33.8  17.7  22.9  10
##  6 2018-19  26.3 6-6     219  1230  40.8  88.6  11.3  31.8  17.6  22.9  10.3
##  7 2017-18  26.4 6-7     220  1230  40.5  87.9  10.7  29.6  17    22.1   9.9
##  8 2016-17  26.6 6-7     221  1230  40.2  88     9.9  27.8  18.4  23.8  10.4
##  9 2015-16  26.7 6-7     222  1230  39.6  87.6   8.8  25    18.3  24.2  10.8
## 10 2014-15  26.7 6-7     223  1230  39.6  88.3   8.3  23.7  18.1  24.1  11.5
## # i 64 more rows
## # i 18 more variables: DRB <dbl>, TRB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>,
## #   TOV <dbl>, PF <dbl>, PTS <dbl>, `FG%` <dbl>, `3P%` <dbl>, `FT%` <dbl>,
## #   Pace <dbl>, `eFG%` <dbl>, `TOV%` <dbl>, `ORB%` <dbl>, `FT/FGA` <dbl>,
## #   ORtg <dbl>, `TS%` <dbl>
```

EDA

We are going to start by evaluating the claim that players shoot more threes in today's game than in past decades. This can be expressed more formally with a hypothesis test.

Hypothesis Test:

Claim: Players shoot more threes in the current decade than in past decades. We will take the mean number of threes attempted in the 2020's and compare it to the mean number of threes in the 1990's as estimates.

```
## # A tibble: 5 x 2
##   Decade Mean3s
##   <chr>   <dbl>
## 1 20's    34.7
## 2 10's    24.1
## 3 00's    17.0
## 4 90's    12.2
## 5 80's     3.39
```

$$H_0$$

: Mean # 3's in the 2020's (

$$M_{2020s}$$

) = Mean # 3s in the 1990's (

$$M_{1990s}$$

)

$$H_A$$

: Mean # 3's in the 2020's > Mean # 3s in the 1990's

$$t_{obs} = \frac{((M_{2020s} - M_{1990s}) - 0)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

$$t_{obs} = \frac{12.56 - 4.25}{\sqrt{\frac{0.3361547^2}{5} + \frac{1.726429^2}{10}}}$$

At 95% significance:

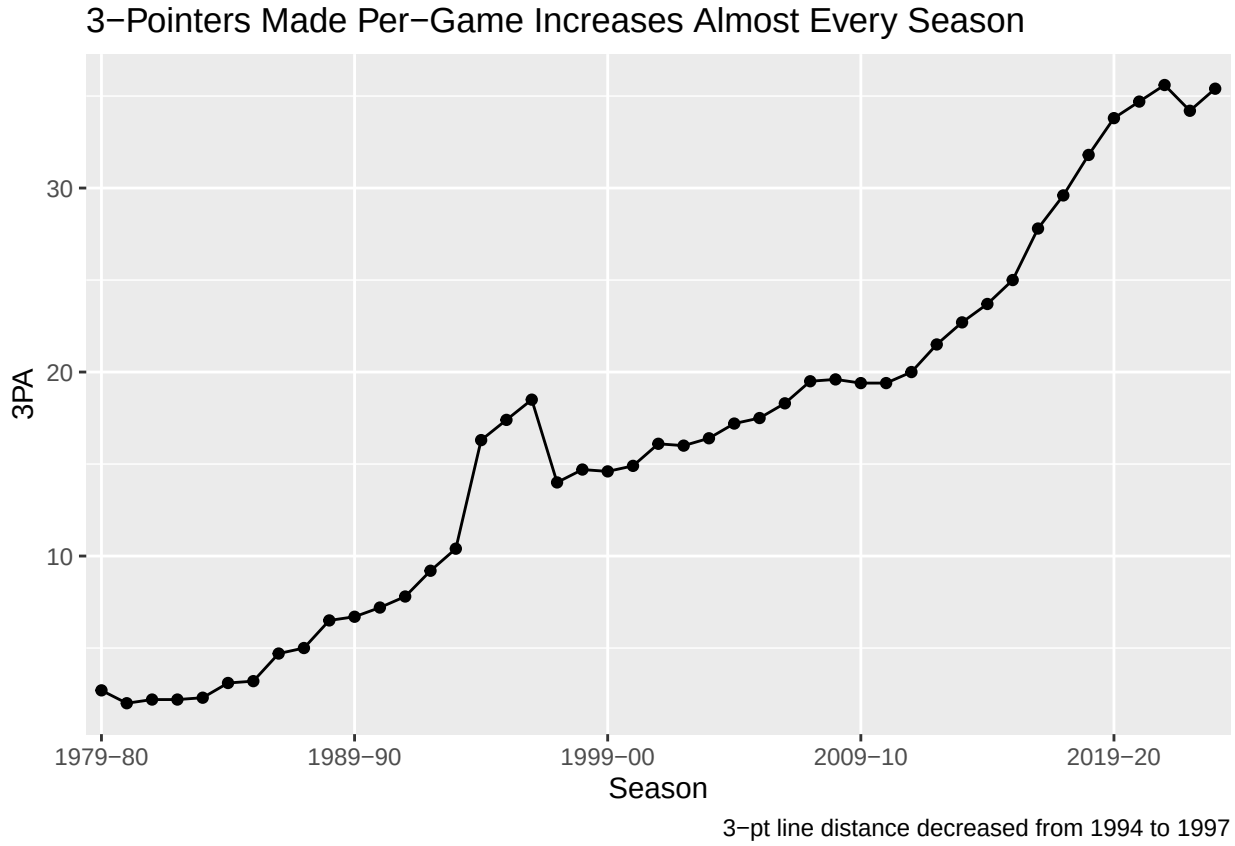
$$t_{0.025,4} = 2.131847$$

$$t_{obs} > t_{\alpha,n-1}$$

So we reject the null hypothesis: Players do shoot more threes in the current decade than in the past.

Plotting Changes in 3-Pointers Made

The evidence strongly suggests that players take more 3s than in the past. We verify this with a visualization:



It is clear that players take more threes in almost every passing season since the addition of the three point line to the NBA game in 1979. This leads to the question; why are players taking more threes?

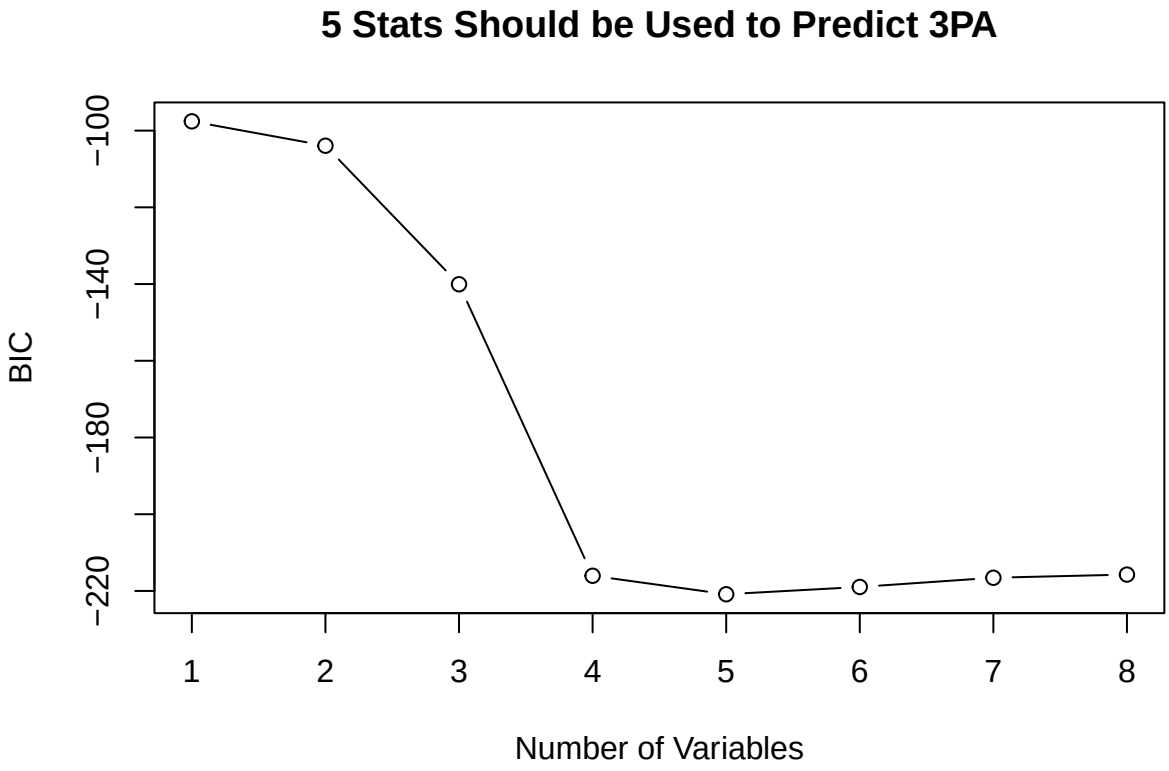
Stat Selection to Predict Threes Attempted

Stats with Direct Linear Dependencies will Be Removed

```
## [1] "The following variables will be removed: Season, Ht, 3P, 3P%, FG, FGA, FG%, FT/FGA, eFG%, TS%"
```

The best stats to predict 3PA will be selected using backwards elmination using BIC (Bayesian Information Criterion).

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found
```



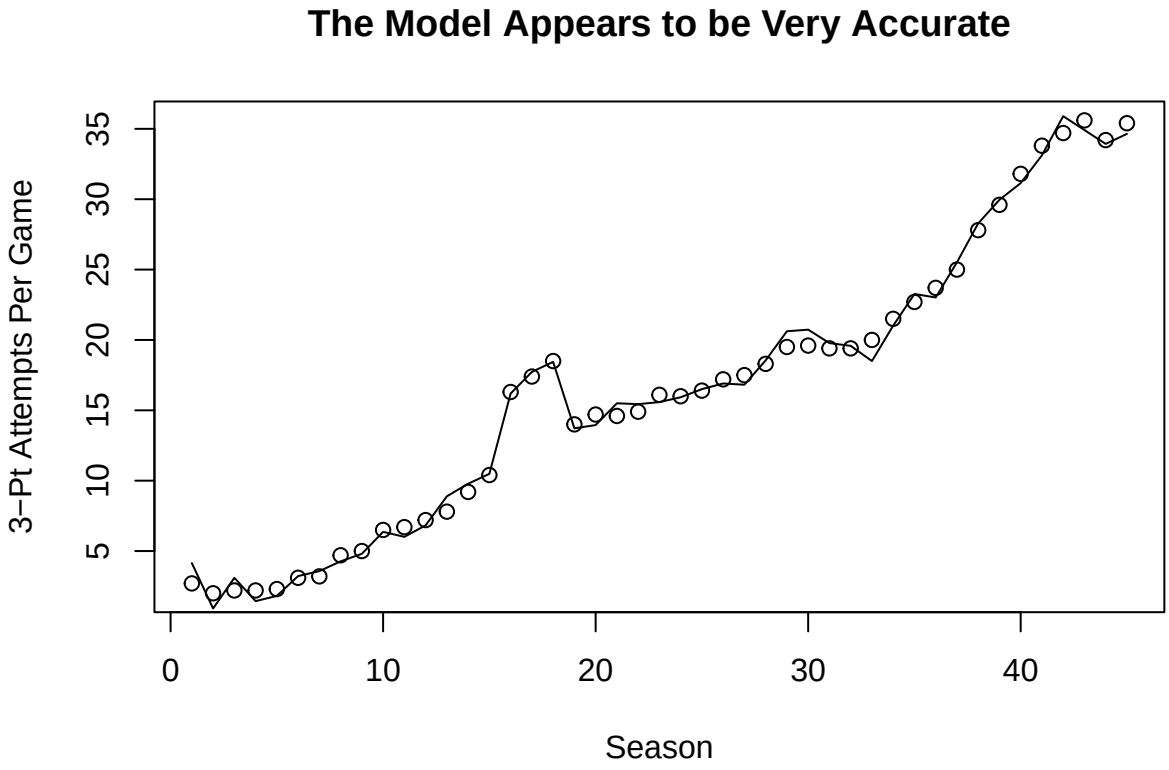
Selecting the Best 5 Variables

```
## [1] "(Intercept)" "TRB"         "STL"         "TOV"         "PTS"
## [6] "`ORB%`"
```

Creating the Model

```
##
## Call:
## lm(formula = `3PA` ~ TRB + STL + TOV + PTS + `ORB%`, data = df_pred3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44218 -0.48094  0.07308  0.51652  1.50071
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -458.13564   22.97046  -19.95  < 2e-16 ***
## TRB            4.21251    0.22953   18.35  < 2e-16 ***
## STL            1.48907    0.51882    2.87   0.0066 **
## TOV            4.96931    0.36845   13.49 2.87e-16 ***
## PTS            2.58828    0.10330   25.06  < 2e-16 ***
## `ORB%`        -2.88542    0.09075  -31.80  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7201 on 39 degrees of freedom
##   (29 observations deleted due to missingness)
## Multiple R-squared:  0.9956, Adjusted R-squared:  0.995
## F-statistic:  1746 on 5 and 39 DF,  p-value: < 2.2e-16
```



ANOVA Table

```
## Analysis of Variance Table
##
## Response: 3PA
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## TRB        1  921.71  921.71 1777.25 < 2.2e-16 ***
## STL        1 2454.16 2454.16 4732.14 < 2.2e-16 ***
## TOV        1  468.44  468.44  903.25 < 2.2e-16 ***
## PTS        1  159.17  159.17  306.92 < 2.2e-16 ***
## `ORB%`     1  524.30  524.30 1010.97 < 2.2e-16 ***
## Residuals 39   20.23    0.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

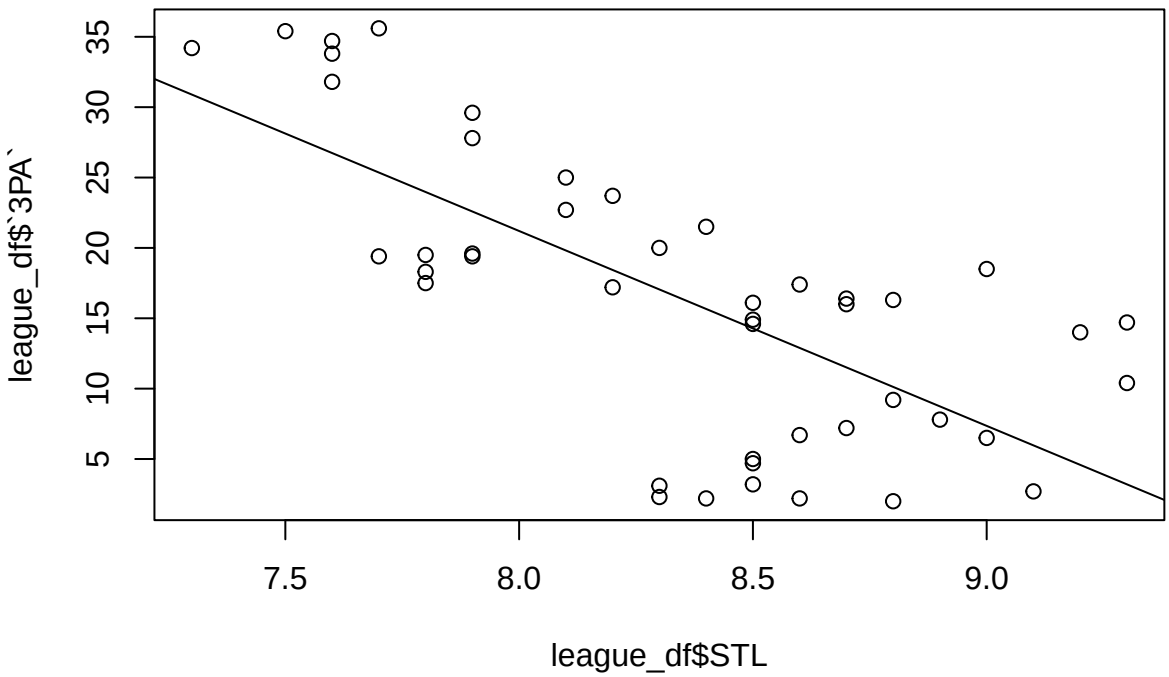
Why are there more 3 point shots attempted?

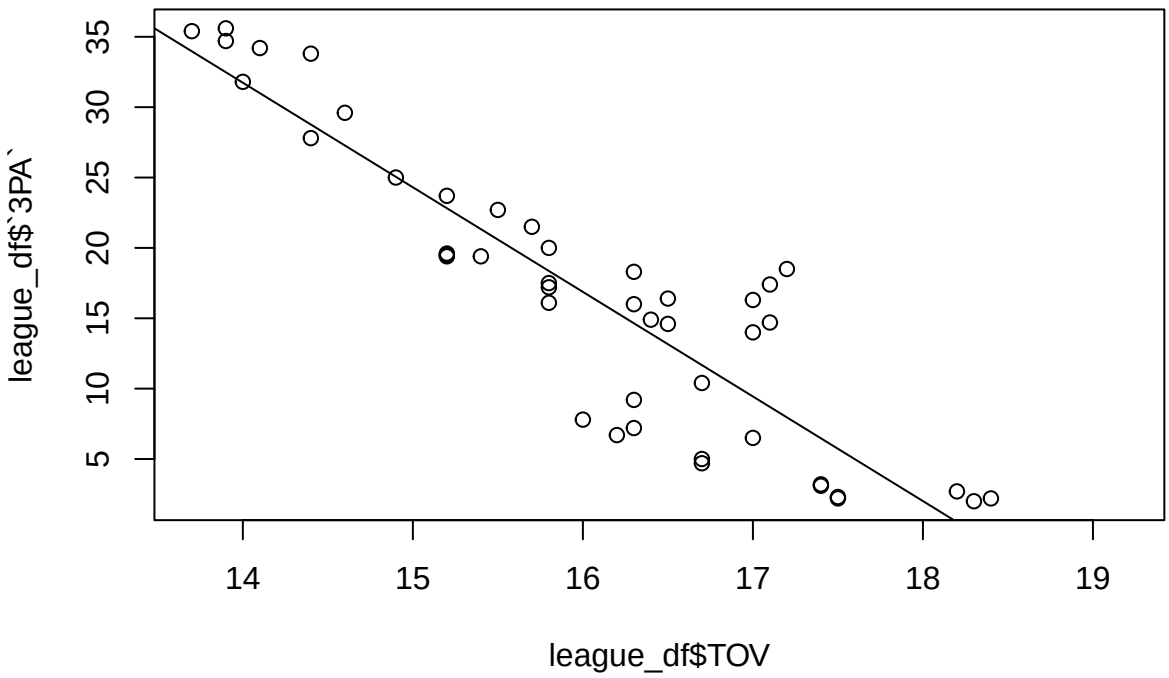
Is the reason teams are taking more 3 point attempts because teams are defending better making driving to the rim and scoring even harder as the league progresses? In order to determine if this is true we'll regress various common defensive stats taken league-wide to see if better defence leads to more 3 point attemots taken.

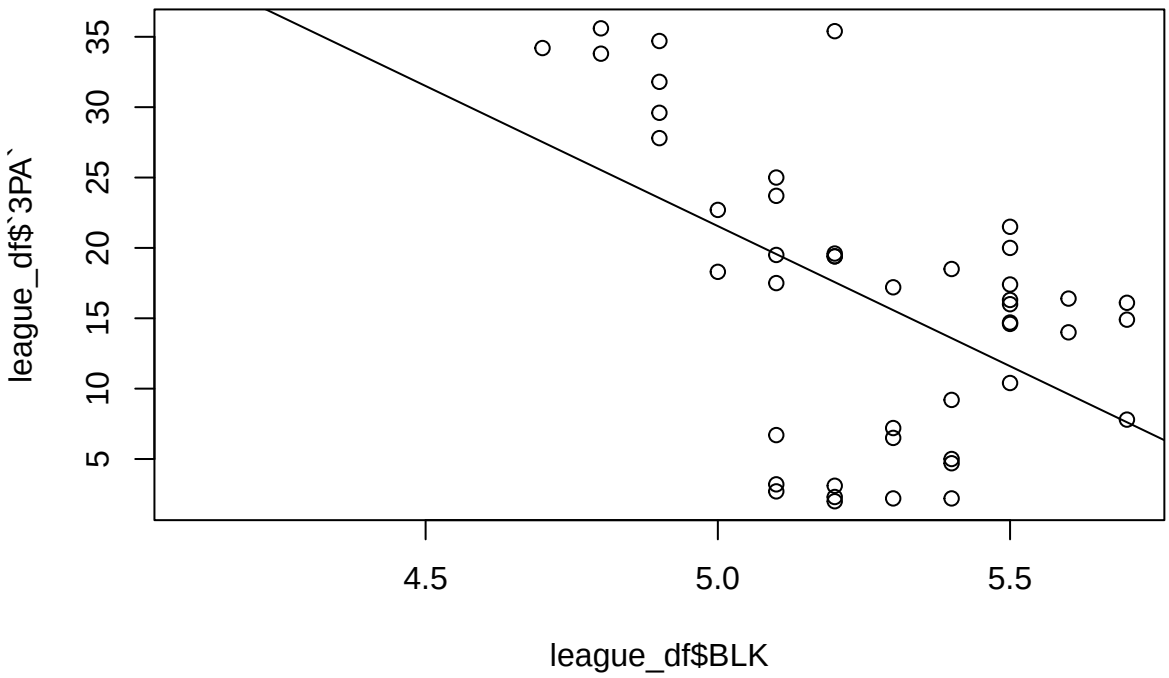
```
##
## Call:
## lm(formula = `3PA` ~ TRB + STL + TOV + PTS + `ORB%`, data = df_pred3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44218 -0.48094  0.07308  0.51652  1.50071
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -458.13564   22.97046  -19.95  < 2e-16 ***
## TRB            4.21251    0.22953   18.35  < 2e-16 ***
## STL            1.48907    0.51882    2.87   0.0066 **
## TOV            4.96931    0.36845   13.49 2.87e-16 ***
## PTS            2.58828    0.10330   25.06  < 2e-16 ***
## `ORB%`        -2.88542    0.09075  -31.80  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7201 on 39 degrees of freedom
##   (29 observations deleted due to missingness)
## Multiple R-squared:  0.9956, Adjusted R-squared:  0.995
## F-statistic:  1746 on 5 and 39 DF,  p-value: < 2.2e-16
```

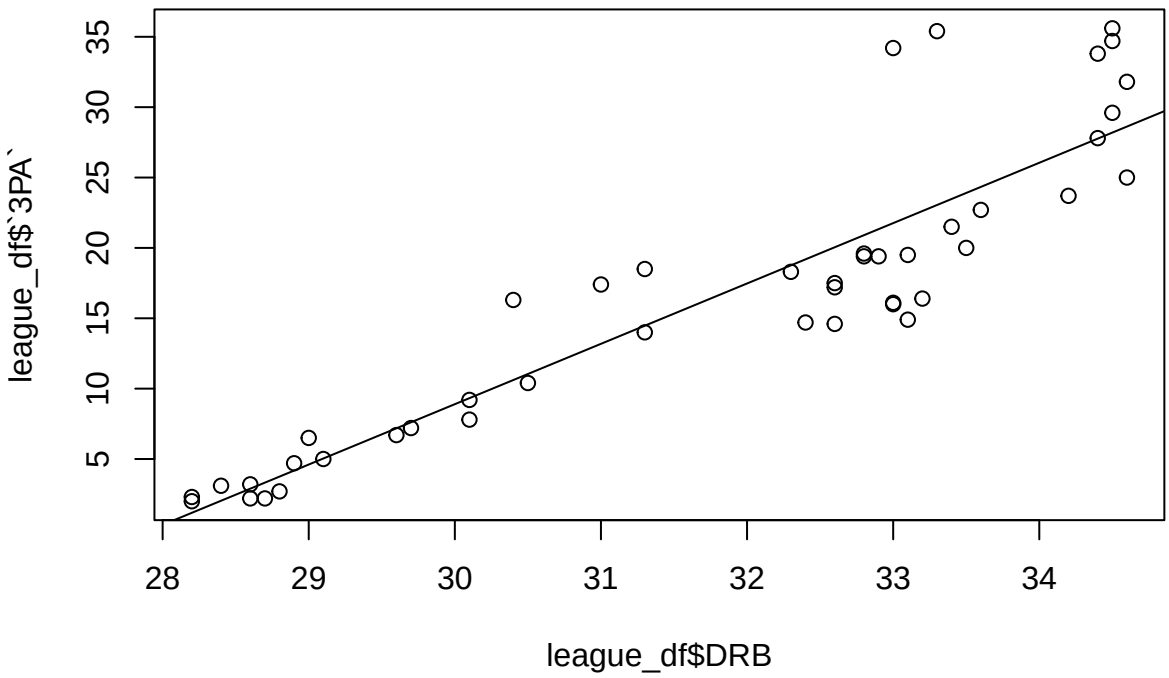
```
## Analysis of Variance Table
##
## Response: 3PA
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## STL        1 2260.71 2260.71 210.5327 < 2.2e-16 ***
## TOV        1 1439.45 1439.45 134.0516 2.396e-14 ***
## BLK        1    4.67    4.67   0.4354    0.5132
## DRB        1  413.65  413.65  38.5219 2.421e-07 ***
## Residuals 40  429.52   10.74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                  2.5 %     97.5 %
## (Intercept) -41.318823 54.3236519
## STL          -3.093618  4.4855666
## TOV          -5.027969 -1.1334385
## BLK         -11.778527 -0.3140392
## DRB           1.812910  3.5637187
```


league_df$`3PA`
5
10
15
20
25
30
35
7.5
8.0
8.5
9.0
league_df$STL


league_df$`3PA`
5
10
15
20
25
30
35
14
15
16
17
18
19
league_df$TOV


league_df$`3PA`
5
10
15
20
25
30
35
4.5
5.0
5.5
league_df$BLK



Based on this output, the opposite appears to be true. As defence gets better 3 point appempts go down. This could also be that due to better defence, teams feel less safe taking deeper shots as well as open shots being much less frequent, fewer 3 point attempts are made. The only discrepancy being that as more 3 point appempts are taken, more defensive rebounds are made which could just be attributed to when teams on offence are all at the 3 point line hoping to shoot the defence is more likely to recover missed shots as theyre closer to the rim meaning they're faster to grab any rebounds

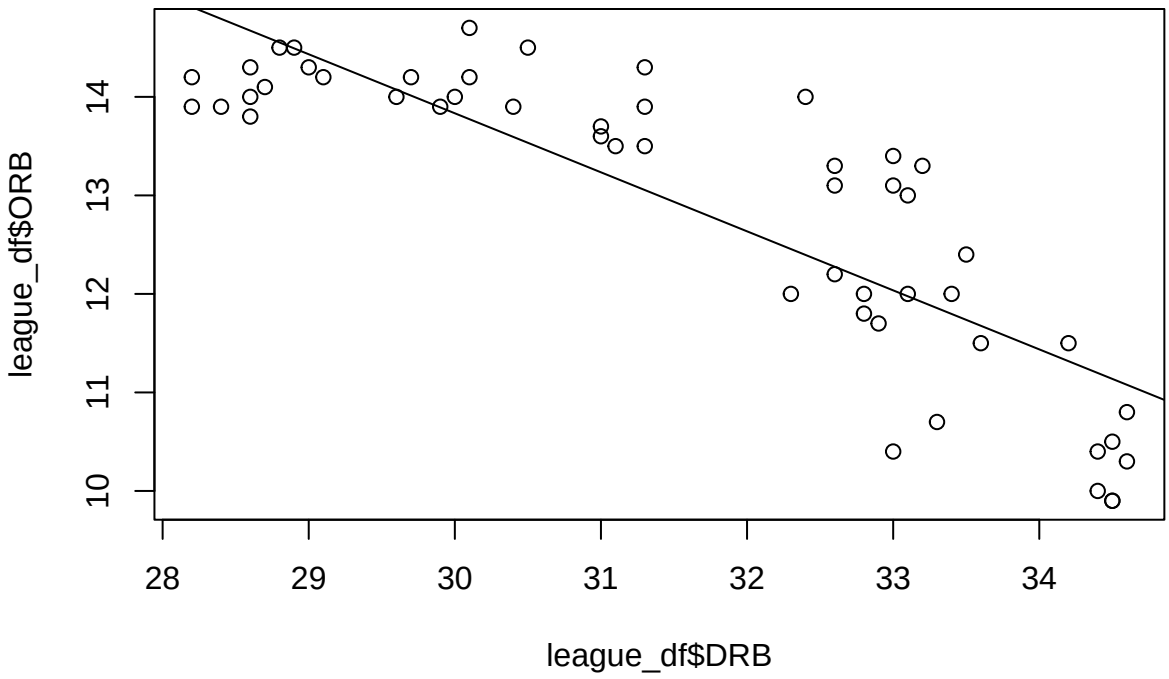
Offensive vs defensive rebounds

Now let's see if a relationship exists between offensive and defensive rebounds to determine if when teams are picking up more rebounds on one side of the court are they picking up fewer on the other side. The implications this has for our experiment are that are when teams are on offence vs defence how do their playstyles change? So for example if when a team is on offence and their playstyle is to just have everyone hover around the 3 point line looking to shoot they won't be able to get as many offensive rebounds. Compared to when theyre on defence if they're still able to get a lot of rebounds showing their capability to rebound then that is good evidence that they aren't as worried about offensive rebounds and are simply trusting their 3 point shooting.

```
##
## Call:
## lm(formula = ORB ~ DRB, data = league_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6361 -0.5749 -0.0525  0.4555  1.6044
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.80776    1.68933   18.83  < 2e-16 ***
## DRB         -0.59914    0.05335  -11.23 3.73e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7903 on 49 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.7202, Adjusted R-squared:  0.7144
## F-statistic: 126.1 on 1 and 49 DF,  p-value: 3.733e-15
```

```
##                  2.5 %     97.5 %
## (Intercept) 28.4129308 35.2025948
## DRB         -0.7063616 -0.4919204
```



Based on these results we can say with reasonable confidence that a relationship exists between the number of defensive rebounds vs the number of offensive rebounds. Where as the number of defensive rebounds increases, the number of offensive rebounds decreases.