Predicting Points per game of NBA Players

Het Patel (7972424)

1 Dec. 2023

Question

In this project I would like to predict number of points per game from for an NBA player. This prediction might be helpful for team management to decide their players, a player's overall performance, player's valuation and many more.

The response variable in my project is the number of points scored by an NBA player per game (PTS). And I will use number of field throws per game (FTA) and number of free throws per game (FTA) by that player per game as my explanatory variable to predict the number of points per game.

The reason why I choose these two variables as my explanatory is because the more number of field throws and number of free throws attempted by a player per game, the more chances that a player will score more points per game.

Data Set

I have used data from the NBA's 2022-23 regular season. This data is available on the official stat centre website of NBA. The original data consist of every player who played in 2022-23 regular season, but I have short it down to top 50 players according to average points per game in that season.

The dataset contains total of 3 variables and the details for each is as follows:

1) The **PTS** variable is the average number of points scored per game by a player in 2022-23 regular season.
2) The **FGA** variable is the average number of field throws attempted per game by a player in 2022-23 regular season.
3) The **FTA** variable is the average number of free throws attempted per game by a player in 2022-23 regular season.

Here is the actual data set containing 50 observations:

```r
nba <- read.csv("nbaplayers.csv")
library(kableExtra) # for kable()
kable(nba, format = "markdown")
```

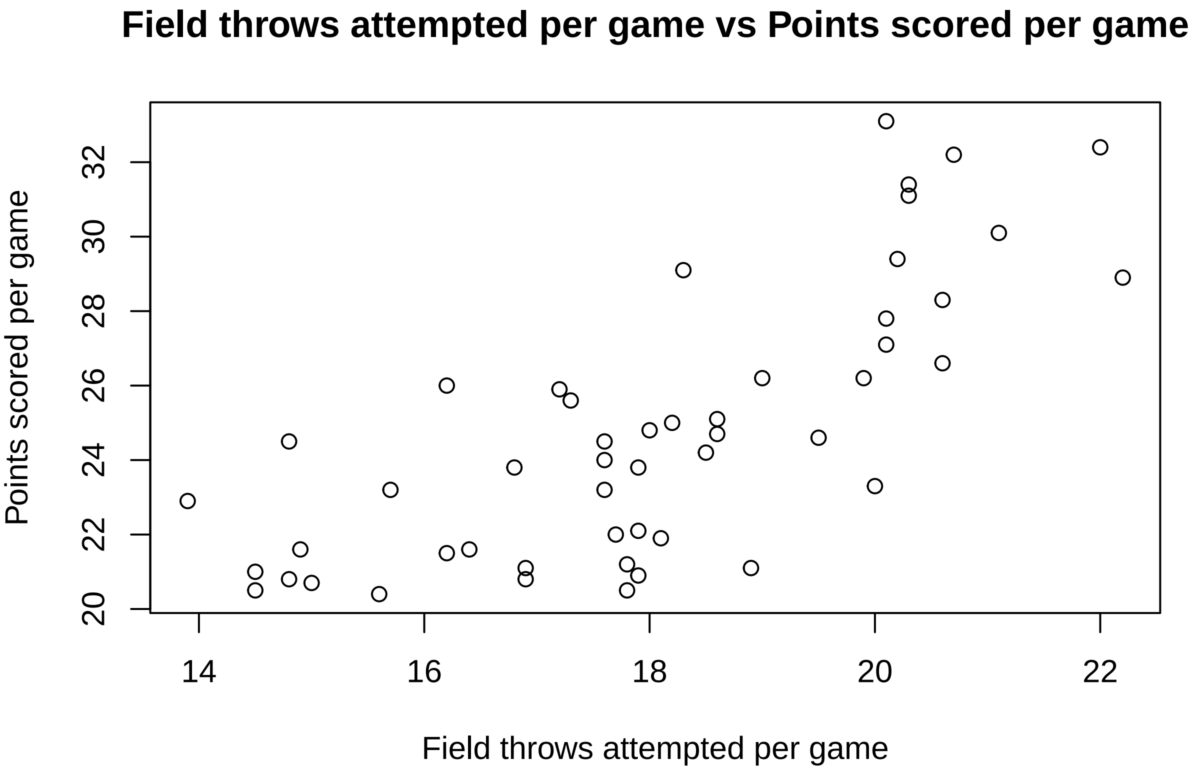
PTS	FGA	FTA
33.1	20.1	11.7
32.4	22.0	10.5
32.2	20.7	9.6
31.4	20.3	10.9
31.1	20.3	12.3
30.1	21.1	8.4
29.4	20.2	5.0
29.1	18.3	7.1
28.9	22.2	5.9
28.3	20.6	5.4
27.8	20.1	6.8
27.1	20.1	4.6
26.6	20.6	5.1
26.2	19.0	8.8
26.2	19.9	8.1
26.0	16.2	8.6
25.9	17.2	7.9
25.6	17.3	6.0
25.1	18.6	6.9
25.0	18.2	6.0
24.8	18.0	5.6
24.7	18.6	6.0
24.6	19.5	5.3
24.5	17.6	7.1
24.5	14.8	6.0
24.2	18.5	6.7
24.0	17.6	5.8
23.8	16.8	5.4
23.8	17.9	5.3
23.3	20.0	3.4
23.2	17.6	4.6
23.2	15.7	6.4
22.9	13.9	8.7
22.1	17.9	6.1
22.0	17.7	5.2
21.9	18.1	1.9
21.6	16.4	4.7
21.6	14.9	5.1
21.5	16.2	3.5
21.2	17.8	3.7
21.1	18.9	3.5
21.1	16.9	2.9
21.0	14.5	6.2
20.9	17.9	3.2
20.8	16.9	4.0

PTS	FGA	FTA
20.8	14.8	4.7
20.7	15.0	3.6
20.5	14.5	5.4
20.5	17.8	2.6
20.4	15.6	5.1

National Basketball Association.(2023). *NBA Advance Player Stats(2022-23 Season)*[Data set]. NBA Media Ventures. https://www.nba.com/stats/players/traditional?PerMode=Totals&sort=PTS&dir=-1&Season=2022-23

Scatterplot of field throws attempted per game and points scored per game:

```r
plot(nba$FGA, nba$PTS, main = "Field throws attempted per game vs Points scored per game",
     xlab = "Field throws attempted per game", ylab = "Points scored per game")
```



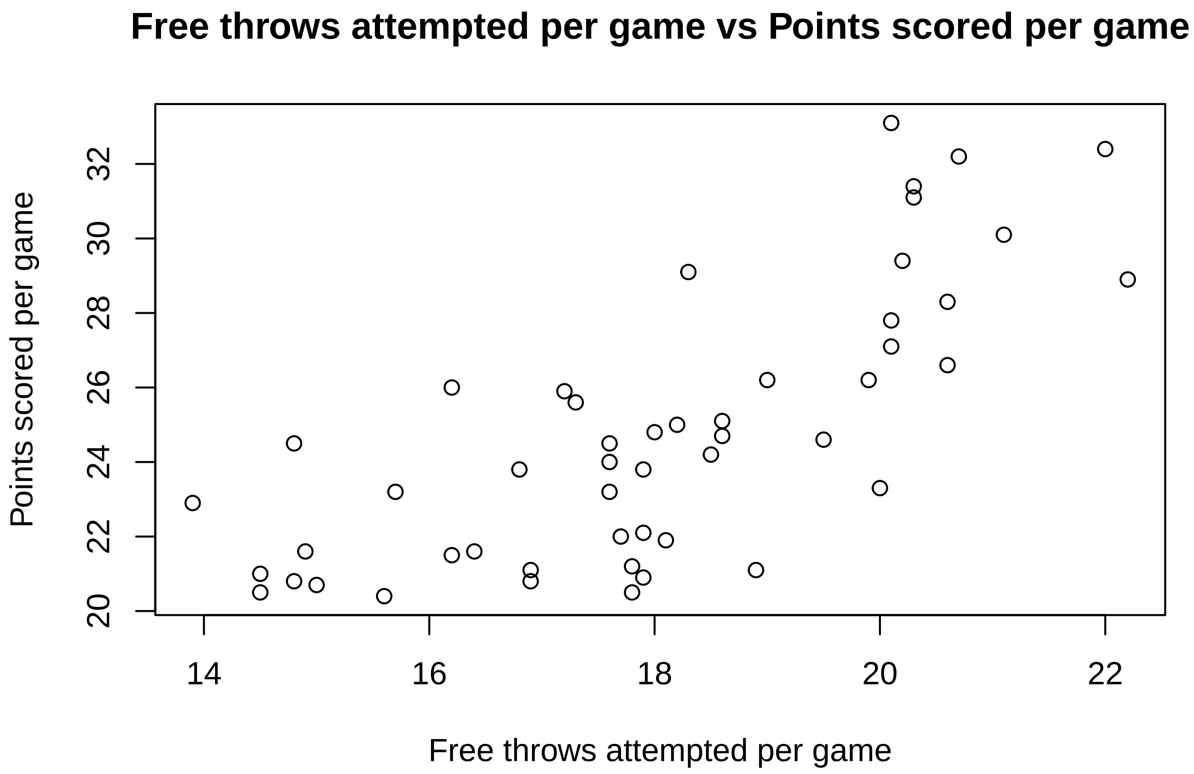
```r
rsqr1 <- cor(nba$FGA, nba$PTS)^2
rsqr1
```

```
## [1] 0.555407
```

The r^2 for field throws attempted and points scored per game is 0.555407, which shows that they appear to be in relationship.

Scatterplot of field throws attempted per game and points scored per game:

```
plot(nba$FGA, nba$PTS, main = "Free throws attempted per game vs Points scored per game",
     xlab = "Free throws attempted per game", ylab = "Points scored per game")
```



```
rsqr2 <- cor(nba$FTA, nba$PTS)^2
rsqr2
```

```
## [1] 0.5949966
```

Since the r^2 for free throws attempted and points scored per game is 0.5949966, they are correlated with each other.

Preliminary Model

Simple linear model for predicting points scored per game from number of field throws attempted per game by a NBA player:

```
nbax1.lm <- lm(PTS ~ FGA, data = nba)
nbax1.lm
```

```
##
## Call:
## lm(formula = PTS ~ FGA, data = nba)
##
## Coefficients:
## (Intercept)          FGA
##       1.584        1.286
```

$\hat{y} = 1.584 + 1.286x_1$

Simple linear model for predicting points scored per game from number of free throws attempted per game by a NBA player:

```r
nbax2.lm <- lm(PTS ~ FTA, data = nba)
nbax2.lm
```

```
## 
## Call:
## lm(formula = PTS ~ FTA, data = nba)
## 
## Coefficients:
## (Intercept)          FTA  
##      17.532        1.194
```

$\hat{y} = 17.532 + 1.194x_2$

Additive model for predicting points scored per game from number of field throws attempted per game and number of free throws attempted per game by a NBA player:

```r
nba.lm <- lm(PTS ~ FGA + FTA, data = nba)
nba.lm
```

```
## 
## Call:
## lm(formula = PTS ~ FGA + FTA, data = nba)
## 
## Coefficients:
## (Intercept)          FGA          FTA  
##      2.1481       0.9486       0.9111
```

$\hat{y} = 2.1481 + 0.9486x_1 + 0.9111x_2$

The R^2_{adj} for all three model can be found in their respective summary output.

```r
summary(nbax1.lm)
```

```
## 
## Call:
## lm(formula = PTS ~ FGA, data = nba)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.7984 -1.2507 -0.0895  1.4066  5.6579 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   1.5844     3.0141   0.526    0.602    
## FGA           1.2865     0.1661   7.744 5.38e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 2.417 on 48 degrees of freedom
## Multiple R-squared:  0.5554, Adjusted R-squared:  0.5461 
## F-statistic: 59.96 on 1 and 48 DF,  p-value: 5.385e-10
```

From the summary we have R^2_{adj} for the model predicting points per game from number of field throw attempted per game for a NBA player is 0.5461.

```
summary(nbax2.lm)
```

```
##
## Call:
## lm(formula = PTS ~ FTA, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0188 -1.5082 -0.2028  1.4257  5.8987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.5316     0.9221  19.013  < 2e-16 ***
## FTA           1.1939     0.1422   8.397 5.56e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.307 on 48 degrees of freedom
## Multiple R-squared:  0.595,  Adjusted R-squared:  0.5866
## F-statistic: 70.52 on 1 and 48 DF,  p-value: 5.56e-11
```

From the summary we have R^2_{adj} for the model predicting points per game from number of free throw attempted per game for a NBA player is 0.5866.

```
summary(nba.lm)
```

```
##
## Call:
## lm(formula = PTS ~ FGA + FTA, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5855 -0.9135  0.0929  0.6659  3.5350
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.14808    1.68813   1.272    0.209
## FGA          0.94858    0.09861   9.620 1.10e-12 ***
## FTA          0.91111    0.08842  10.304 1.21e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.353 on 47 degrees of freedom
## Multiple R-squared:  0.8636, Adjusted R-squared:  0.8578
## F-statistic: 148.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

From the summary we have R^2_{adj} for the additive model predicting points per game from number of field throw attempted and number of free throw per game for a NBA player is 0.8578.

We can clearly see that the R^2_{adj} for additive model is greater than the R^2_{adj} for the model for predicting points per game from field throws attempted and for the model for predicting points per game from free throws attempted per game.

Full second-order model for predicting points scored per game from number of field throws attempted per game and number of free throws attempted per game by a NBA player:

```
nba.full <- lm(PTS ~ FGA + FTA + I(FGA^2) + I(FTA^2) + FGA*FTA, data = nba)
nba.full
```

```
##
## Call:
## lm(formula = PTS ~ FGA + FTA + I(FGA^2) + I(FTA^2) + FGA * FTA,
##     data = nba)
##
## Coefficients:
## (Intercept)          FGA          FTA     I(FGA^2)     I(FTA^2)      FGA:FTA
##   22.488764    -1.384117     0.987313     0.064920    -0.015331     0.004703
```

$\hat{y} = 22.488764 - 1.384117x_1 + 0.987313x_2 + 0.064920x_1^2 - 0.015331x_2^2 + 0.004703x_1x_2$

ANOVA Table for the full second-order model:

```
anova(nba.full)
```

```
## Analysis of Variance Table
##
## Response: PTS
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## FGA        1 350.22  350.22 190.0525 < 2.2e-16 ***
## FTA        1 194.32  194.32 105.4542 2.927e-13 ***
## I(FGA^2)   1   4.48    4.48   2.4294    0.1262
## I(FTA^2)   1   0.45    0.45   0.2426    0.6248
## FGA:FTA    1   0.01    0.01   0.0072    0.9326
## Residuals 44  81.08    1.84
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Source	df	Sum of Squares	Mean Square	F
Model	5	549.48	109.896	59.726
Error	44	81.08	1.84	
Total	49	630.56		

P-value for the ANOVA test:

```
pf(59.726, 5, 44, lower.tail = FALSE)
```

```
## [1] 1.699667e-18
```

ANOVA test on the full second-order model to identify if one term is significant or not in model:

1) LEVEL OF SIGNIFICANCE: $\alpha = 0.05$
2) HYPOTHESIS: $H_0 : \beta_i = 0$, i = 1, 2, 3, 4, 5 vs H_a : at least one of $\beta_i \neq 0$, i = 1, 2, 3, 4, 5
3) DECISION RULE: Reject H_0 if p-value $\leq \alpha = 0.05$
4) TEST STATISTICS: F = 59.726
5) P-VALUE: p-value = 1.7×10^{-18}
6) CONCLUSION: As p-value $< \alpha = 0.05$, we reject H_0. At 5% level of significance we have enough evidence that at least one of the term is significant in the full second-order model for predicting points from field throw attempted and free throw attempted per game by a NBA player in 2022-23 season.

Model Refinement

In this section I will check if the terms present in the full second-order model for predicting points scored per game from field throws attempted and free throws attempted per game, are significant or not.

To check if terms are significant or not I have to perform t-test on each term present in full model, this can be achieved by passing the full second-order model in the summary() function.

```
summary(nba.full)
```

```
##
## Call:
## lm(formula = PTS ~ FGA + FTA + I(FGA^2) + I(FTA^2) + FGA * FTA,
##     data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3796 -0.8824 -0.0542  0.7347  3.3524
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.488764  14.283937   1.574    0.123
## FGA         -1.384117   1.523174  -0.909    0.368
## FTA          0.987313   0.894672   1.104    0.276
## I(FGA^2)     0.064920   0.042696   1.521    0.136
## I(FTA^2)    -0.015331   0.033185  -0.462    0.646
## FGA:FTA      0.004703   0.055306   0.085    0.933
##
## Residual standard error: 1.357 on 44 degrees of freedom
## Multiple R-squared:  0.8714, Adjusted R-squared:  0.8568
## F-statistic: 59.64 on 5 and 44 DF,  p-value: < 2.2e-16
```

I can see from the summary() function that there is no significant term in the full second-order model for predicting points per game by field throws attempted and free throws attempted per game by a NBA player.

I will check the VIF values for each term in the full second-order to assess multicollinearity in our model. Running the *vif()* on the full second-order model.

```r
library(car) # for vif()
vif(nba.full)
```

```
##       FGA       FTA  I(FGA^2)  I(FTA^2)   FGA:FTA
## 266.43186 114.32454 270.74792  31.99203 204.58385
```

Since all of the VIF values for coefficient in full model except FTA^2 is more than 5, multicollinearity is present in the model. To resolve this I will remove the FGA^2, the term with the largest VIF value of 270.74792 among all other terms in the full-second model. Now, our reduced model consists of FGA, FTA, FTA^2 and FGA*FTA terms.

```r
nba.reduced <- lm(PTS ~ FGA + FTA + I(FTA^2) + FGA*FTA, data = nba)
summary(nba.reduced)
```

```
##
## Call:
## lm(formula = PTS ~ FGA + FTA + I(FTA^2) + FGA * FTA, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6668 -0.9151  0.0878  0.7016  3.5109
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.84076    6.17540   0.460   0.6477
## FGA          0.86789    0.36076   2.406   0.0203 *
## FTA          0.90611    0.90600   1.000   0.3226
## I(FTA^2)    -0.01993    0.03353  -0.594   0.5552
## FGA:FTA      0.01470    0.05571   0.264   0.7931
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.377 on 45 degrees of freedom
## Multiple R-squared:  0.8647, Adjusted R-squared:  0.8526
## F-statistic: 71.87 on 4 and 45 DF,  p-value: < 2.2e-16
```

I can observe again that not all of the term in reduced model are significant, so I will check the VIF scores for each term in reduced model and check if multicollinearity is still present in the model.

```r
vif(nba.reduced)
```

```
##       FGA       FTA  I(FTA^2)   FGA:FTA
##  14.52256 113.91722  31.72628 201.69293
```

It is clear from the VIF scores that multicollinearity is still present in the reduced model, so I will again remove the term with the largest VIF score and make a new reduced model. As FGA*FTA has the largest VIF score of 201.69293, I will remove it from the reduced and now our new reduced model contains FGA, FTA and FTA^2 terms.

```r
nba.reduced <- lm(PTS ~ FGA + FTA + I(FTA^2), data = nba)
summary(nba.reduced)
```

```
##
## Call:
## lm(formula = PTS ~ FGA + FTA + I(FTA^2), data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6652 -0.9311  0.0995  0.6526  3.4750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.32857    2.27678   0.584  0.56239
## FGA          0.95917    0.10127   9.472 2.22e-12 ***
## FTA          1.12050    0.39681   2.824  0.00699 **
## I(FTA^2)    -0.01525    0.02817  -0.541  0.59078
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.363 on 46 degrees of freedom
## Multiple R-squared:  0.8644, Adjusted R-squared:  0.8556
## F-statistic: 97.79 on 3 and 46 DF,  p-value: < 2.2e-16
```

From the summary of our reduced model, it is clear that, for FTA^2 variable, p-value $= 0.59078 > \alpha = 0.05$, so we conclude that FTA^2 is not significant and I will remove that term from the model, resulting a new reduced model in which only FGA and FTA term is there.

The list of the co-efficient that are significant in the reduced model is:

Co-efficient	P-value
FGA	2.22×10^{-12}
FTA	0.00699

After analyzing the full second-order model with series of VIF scores check and t-test on the terms in the model I propose the final reduced model as follows:

```r
nba.reduced <- lm(PTS ~ FGA + FTA, data = nba)
nba.reduced
```

```
##
## Call:
## lm(formula = PTS ~ FGA + FTA, data = nba)
##
## Coefficients:
## (Intercept)          FGA          FTA
##      2.1481       0.9486       0.9111
```

$y = 2.1481 + 0.9486x_1 + 0.9111x_2$

The above is the reduced model which is trying to predict number of points scored per game by a NBA player from number of field throws and number of free throws attempted per game.

Now let us run a nested F-test to test that the terms I removed were in fact zero comparing full model to reduced model.

```
anova(nba.reduced, nba.full)
```

```
## Analysis of Variance Table
##
## Model 1: PTS ~ FGA + FTA
## Model 2: PTS ~ FGA + FTA + I(FGA^2) + I(FTA^2) + FGA * FTA
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     47 86.017
## 2     44 81.080  3    4.9372 0.8931 0.4523
```

Nested F-test comparing full model to reduced model:

1) LEVEL OF SIGNIFICANCE: $\alpha = 0.05$
2) HYPOTHESIS: $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ vs $H_a :$ atleast one of $\beta_i \neq 0, \;$ i = 3, 4, 5
3) DESICION RULE: Reject H_0 if p-value $\leq \alpha$
4) TEST STATISTICS: F = 0.8931
5) P-VALUE: p-value = 0.4523
6) CONCLUSION: As p-value $> \alpha = 0.05$, we fail to reject H_0. At 5% level of significance we have insufficient evidence that FGA^2, FTA^2, and $FGA*FTA$ are significant when FGA and FTA are present in the model.

Final Model and Assessment

Now let us perform an ANOVA test on the reduced model to test if the reduced model adequately explains the relationship with number of points scored per game by a NBA player. To perform an ANOVA test we need F statistics and it respective p-value, which we can get from the output by running summary() on the reduced model.

```
summary(nba.reduced)
```

```
##
## Call:
## lm(formula = PTS ~ FGA + FTA, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5855 -0.9135  0.0929  0.6659  3.5350
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.14808    1.68813   1.272    0.209
## FGA          0.94858    0.09861   9.620 1.10e-12 ***
## FTA          0.91111    0.08842  10.304 1.21e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.353 on 47 degrees of freedom
## Multiple R-squared:  0.8636, Adjusted R-squared:  0.8578
## F-statistic: 148.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

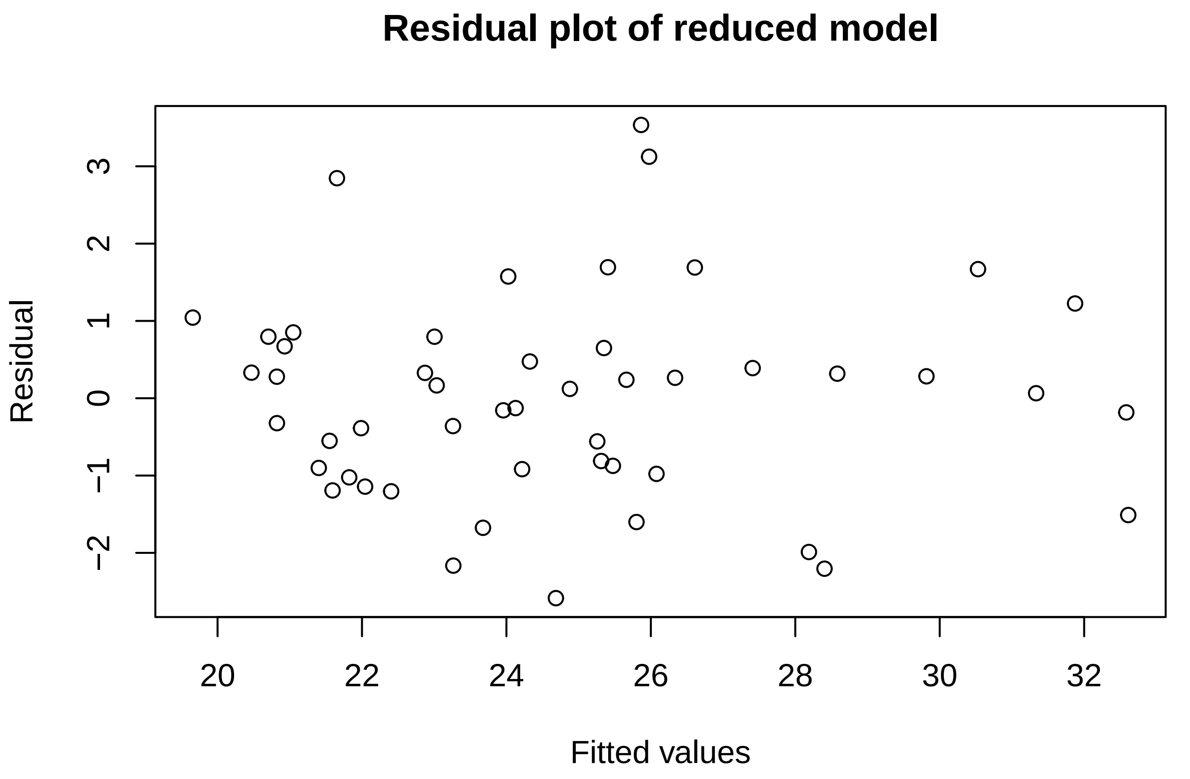
ANOVA test on reduced model to identify if one term is significant or not in model:

1) LEVEL OF SIGNIFICANCE: $\alpha = 0.05$
2) HYPOTHESIS: $H_0 : \beta_1 = \beta_2 = 0$ vs H_a : at least one of $\beta_i \neq 0$, i = 1, 2
3) DECISION RULE: Reject H_0 if p-value $\leq \alpha = 0.05$
4) TEST STATISTICS: F = 148.8
5) P-VALUE: p-value $< 2.2 \times 10^{-16}$
6) CONCLUSION: As p-value $< \alpha = 0.05$, we reject H_0. At 5% level of significance we have enough evidence that at least one of the term is significant in the reduced model for predicting points from field throw attempted and free throw attempted per game by a NBA player in 2022-23 season.

Now let us check if the our reduced model violates any of the assumption that a linear model should met. I will check three assumption, first one being that error terms have constant variance, second assumption is linearity of relationship and last assumption to check is error terms follow normal distribution.

To check first 2 assumption I have to make a residual plot for the reduced model.

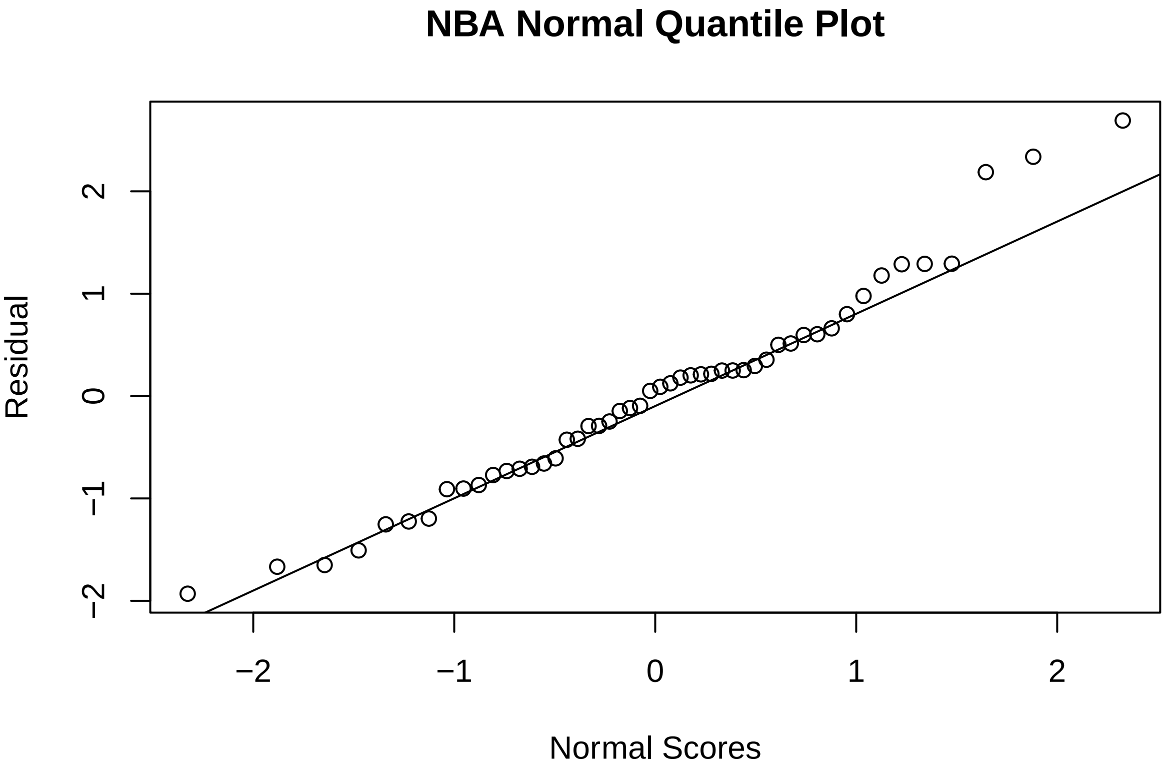
```r
nba.fitted <- fitted.values(nba.reduced)
nba.residual <- resid(nba.reduced)
plot(nba.fitted, nba.residual, main = "Residual plot of reduced model",
     xlab = "Fitted values", ylab = "Residual")
```



As we can see that the width of the plot seems to be narrower at the beginning and increasing at the center of the residual plot, the assumption of error terms have constant variance might be violated. Moreover, there is no discernible pattern in the points so it appears that linearity assumption would be correct.

Now to check the normality assumption we have to make a normal quantile plot for the reduced model

```
nba.stdres <- rstandard(nba.reduced)
qqnorm(nba.stdres, ylab = "Residual", xlab = "Normal Scores", main = "NBA Normal Quantile Plot")
qqline(nba.stdres)
```



Upon inspecting the normal quantile plot we can see that the normality assumption is fairly reasonable though there may be skewness in right tail as the points start to deviate from line at the end of the plot.

Conclusion

Since the assumptions are fairly reasonable and from the conclusion of the ANOVA test on the reduced model I conclude that the number of field throws attempted and number of free throws attempted per game by a NBA player is adequately able to predict the number of points scored per game by that NBA player.

The final regression equation as my best estimate of relationship between points scored per game, number of field throws attempted per game and number of free throws attempted per game for a NBA player is as follows:

$$y = 2.1481 + 0.9486x_1 + 0.9111x_2$$