Regression Model for Four Factors in the NBA

Ben Chapman

May 15th

To include mathematical equations in your document, write in between two dollar signs. Make sure there is no space after the first dollar sign or before the last one.

E.g. $this is good$ but $ this is bad $.

If you want the equation to be centered and on it’s own line, then put it in between four dollar signs

If you are using an offline version of RStudio, you need to make sure you have LateX installed on your computer as well. Without this, you will not be able to produce neat mathematical equations in your document.

## Introduction

Basketball is a sport in which analytics have become increasingly more used to discover what the most impactful elements of the game are. Basketball has been one of the most analyzed sports and there have been many methods of analysis created to measure different game elements. There has been a lot of research on what game statistics influence wins. (García et al. [2013](#ref-garcia); Mandić et al. [2019](#ref-mandic)) A former scout, Dean Oliver, came up with an influential model called the four factors, which he noted as the four most important elements of a basketball game. The four factors are shooting, turnovers, rebounding, and free throws. ([2004](#ref-oliver)) The four factors have been studied extensively, including studies on the trends over the years (Mandić et al. [2019](#ref-mandic)), regression models for the four factors. Over a decade ago, Winston built a regression model for the four factors using data from 2006-2007 NBA season and he found that four factor model was an effective to look at team’s strengths and weaknesses. ([2009](#ref-winston)) In this paper we’ll be looking at more current NBA data from the 2018-2019 season and exploring the four factors on offense and defense.

## Methods

The data is consists of team data from all 30 NBA teams from the 2018-2019 season. All the data was obtained from basketball-reference. The data consists of team names, win totals, offensive four factors, defensive four factors. For shooting, the statistic used is effective field goal percentage. Effective field goal percentage is calculated by (all field goals made + .5\*(3 pt field goals made))/(field goals attempts). This is used because it weights 3 pointers by 50% since they are worth 50% more points, so it is a better metric than field goal percentage. The variable off\_shooting for each team is their own effective field goal percentage and def\_shooting is the effective field goal percentage of what their opponents against them. Turnovers is the estimated turnovers per possession. A possession is begins when a team has control of the basketball and it ends when the team gives up control of the basketball. The variable off\_turnovers measures the turnovers a team committed per possession and def\_turnovers measures the turnovers a team caused there opponents to commit per possession. The variable off\_rebounding is a percentage of rebounds a team gets from their own missed shots and def\_rebounding is a percentage of rebounds a team gets from their opponents missed shots. The variable off\_freethrows is the (foul shots made)/(field goal attempts) and def\_freethrows is the (opponents foul shots made)/(opponents field goal attempts). (WINSTON [2009](#ref-winston))

I built a multiple linear regression model with the offensive and defensive four factors as the predictors and wins as the response.

Looking at a correlation matrix, none of the variables have any correlations greater than .6, so we can assume that there is no multicollineairty. In the residual plot there is no pattern and the variance is constant, so we can assume that there is a linear relationship and homoskedascity. In the normal probability plot the data has a slight skew, so the normality assumption doesn’t hold.

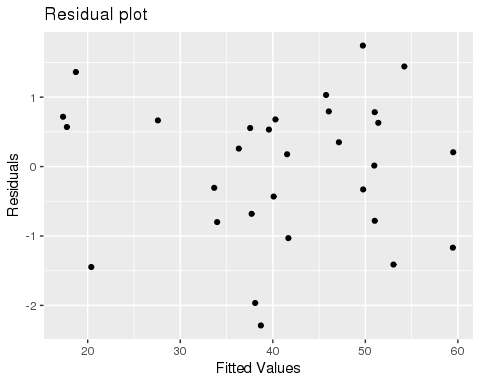
NBA\_data <- read\_csv("NBA\_data.csv")  
mod <- lm(wins ~ off\_shooting + off\_turnovers + off\_rebounding + off\_freethrows + def\_shooting + def\_turnovers + def\_rebounding + def\_freethrows, NBA\_data)  
#correlation matrix  
cor(NBA\_data %>% dplyr::select(-c(team\_name, wins)))

## off\_shooting off\_turnovers off\_rebounding off\_freethrows  
## off\_shooting 1.00000000 -0.02796114 -0.05217861 0.04541527  
## off\_turnovers -0.02796114 1.00000000 -0.02468111 0.16195630  
## off\_rebounding -0.05217861 -0.02468111 1.00000000 -0.02198649  
## off\_freethrows 0.04541527 0.16195630 -0.02198649 1.00000000  
## def\_shooting -0.56981036 -0.05805839 -0.07391434 -0.08560241  
## def\_turnovers -0.08716977 0.01525680 -0.07509507 -0.25808941  
## def\_rebounding 0.16673361 -0.17945826 0.14101416 -0.12485967  
## def\_freethrows -0.18466018 0.54483474 0.07986448 0.25945557  
## def\_shooting def\_turnovers def\_rebounding def\_freethrows  
## off\_shooting -0.56981036 -0.08716977 0.1667336 -0.18466018  
## off\_turnovers -0.05805839 0.01525680 -0.1794583 0.54483474  
## off\_rebounding -0.07391434 -0.07509507 0.1410142 0.07986448  
## off\_freethrows -0.08560241 -0.25808941 -0.1248597 0.25945557  
## def\_shooting 1.00000000 0.11566351 -0.5075197 0.14900951  
## def\_turnovers 0.11566351 1.00000000 -0.4108186 0.40401973  
## def\_rebounding -0.50751966 -0.41081859 1.0000000 -0.48061851  
## def\_freethrows 0.14900951 0.40401973 -0.4806185 1.00000000

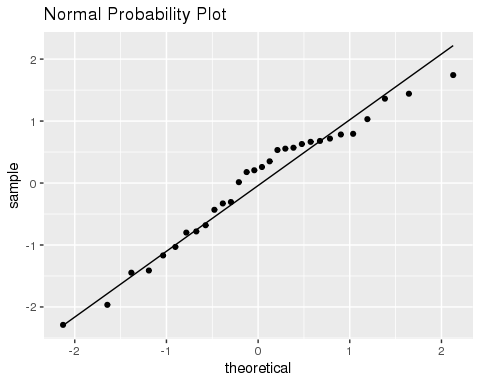
library("MASS")  
mod %>% studres

## 1 2 3 4 5 6 7   
## 0.2061861 -1.1684693 1.4413058 -1.4117119 0.6297018 0.7832887 1.7422396   
## 8 9 10 11 12 13 14   
## -0.7811712 -0.3297106 0.7940934 0.0139358 0.3503511 1.0308880 0.1768630   
## 15 16 17 18 19 20 21   
## 0.6788524 -0.4317585 0.5327828 -1.0309827 -1.9654807 -0.6810092 -2.2890342   
## 22 23 24 25 26 27 28   
## 0.5548042 0.2586642 -0.3063217 -0.7997225 0.6647800 1.3613293 0.7171598   
## 29 30   
## -1.4467662 0.5696723

data\_res <- NBA\_data %>% mutate(stu\_res = studres(mod),  
 fit\_y = mod$fitted.values)  
ggplot(data\_res, aes(x=fit\_y, y=stu\_res)) +   
 geom\_point() +   
 labs(title="Residual plot",  
 y='Residuals',  
 x='Fitted Values')



ggplot(data\_res, aes(sample=stu\_res)) +  
 stat\_qq() +   
 stat\_qq\_line() +   
 labs(title="Normal Probability Plot")



## Results

mod <- lm(wins ~ off\_shooting + off\_turnovers + off\_rebounding + off\_freethrows + def\_shooting + def\_turnovers + def\_rebounding + def\_freethrows, NBA\_data)  
summary(mod)

##   
## Call:  
## lm(formula = wins ~ off\_shooting + off\_turnovers + off\_rebounding +   
## off\_freethrows + def\_shooting + def\_turnovers + def\_rebounding +   
## def\_freethrows, data = NBA\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.7105 -1.9144 0.5913 1.6513 4.2681   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -111.841 79.577 -1.405 0.174512   
## off\_shooting 427.195 45.441 9.401 5.65e-09 \*\*\*  
## off\_turnovers -484.463 89.145 -5.435 2.16e-05 \*\*\*  
## off\_rebounding 141.590 30.352 4.665 0.000133 \*\*\*  
## off\_freethrows 87.696 31.817 2.756 0.011833 \*   
## def\_shooting -333.111 57.260 -5.818 8.95e-06 \*\*\*  
## def\_turnovers 216.137 74.319 2.908 0.008407 \*\*   
## def\_rebounding 111.820 45.309 2.468 0.022265 \*   
## def\_freethrows 4.271 49.148 0.087 0.931573   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.923 on 21 degrees of freedom  
## Multiple R-squared: 0.9572, Adjusted R-squared: 0.941   
## F-statistic: 58.77 on 8 and 21 DF, p-value: 1.211e-12

From our model we can gather that average games won = -111.84 + 427.19(off\_shooting) - 484.46(off\_turnovers) + 141.59(off\_rebounding) + 87.70(off\_freethrows) - 333.11(def\_shooting) + 216(def\_turnovers) + 111.82(def\_rebounding) + 4.27(def\_freethrows) The four factors cover 94.1% of variation in the games won. Using a Hypothesis test for multiple linear regression with 0.05 significance level, a null hypothesis of all the beta coefficients = 0 and alternative of at least one of the beta coefficients != 0. For this model we get that the p-value of 1.211e-12 < 0.05 so we reject the null and conclude that there is statistical evidence of correlation between wins and at least one of the variables. For hypothesis tests for individual predictors in multiple linear regression, the null hypothesis is that beta = 0 and alternative is beta != 0. Using a 0.05 significance level for the 8 predictors defensive free throws has a p-value > 0.05 and so there is no statistically significant evidence of correlation between defensive free throws and wins. The other seven predictors all have p-value < 0.05 so there is statistically significant evidence of correlation between each of those predictors and wins.

## Conclusion

The overall model seems strong which is backed up by previous literature, but there are so differences between this model and previous ones(Mandić et al. [2019](#ref-mandic); WINSTON [2009](#ref-winston)). For every 1% increase in effective field goal percentage teams will win an average 4.27 more games and for every 1% increase in opponent’s effective field goal percentage teams will lose an average of 3.33 more games. A team’s own effective field goal percentage has a greater impact on wins than their opponents. Shooting well is more important for teams, then defending well, but they are both still important. The effect of opponents shooting is similar to a previous model but, team shooting has a larger effect than the previous model. (WINSTON [2009](#ref-winston)) This makes sense because the literature has shown that shooting has been trending upward (Mandić et al. [2019](#ref-mandic)) and having a greater impact on the dynamics of the game. (García et al. [2013](#ref-garcia)) For every additional turnover per 100 possessions teams will lose an average 4.84 more games and for every additional turnover caused per 100 possessions teams will win an average of 2.16 more games. The effects of turnovers committed and turnovers forced are different from the previous model by themselves, but their average is similar to the previous model. (WINSTON [2009](#ref-winston)) The total effect of turnovers is similar, but taking care of the ball and not turning it ball over is a greater impact on wins than forcing turnovers.  
For every additional offensive rebound per 100 missed shots teams will win an average of 1.41 more games and for every additional defensive rebound per 100 missed shots teams will win an average if 1.11 more games. The impact of rebounding is similar to the previous model. (WINSTON [2009](#ref-winston)) Offensive rebounding seems to have a slightly higher impact.  
For every 1 free throw made per 100 field goal attempts teams will win an average of .87 more games and for every 1 opponent free throw made per 100 field goal attempts teams will win an average of .04 more games. The opponent free throw makes having a positive effect on wins doesn’t make sense, but the effect is small and the hypothesis test shows no correlation between defensive free throws and wins. Compared to the previous model free throws are not correlated with wins, and there estimated effect is similar (WINSTON [2009](#ref-winston)). Limitations from this model is that it doesn’t complete the normality assumption, so while the inference is similar to previous literature, it the data isn’t valid for inference. Future studies could look into possible data transformations for the data to try to fit the normality assumption. This model also wasn’t able to compare the impact of each of the factors individually, so there could be more work done on which of the factors have the greatest impacts on winning.

## Bibliography

García, Javier, Sergio J. Ibáñez, Raúl Martinez De Santos, Nuno Leite, and Jaime Sampaio. 2013. “Identifying Basketball Performance Indicators in Regular Season and Playoff Games.” *Journal of Human Kinetics* 36 (1): 161–68. <https://doi.org/10.2478/hukin-2013-0016>.

Mandić, Radivoj, Saša Jakovljević, Frane Erčulj, and Erik Štrumbelj. 2019. “Trends in NBA and Euroleague Basketball: Analysis and Comparison of Statistical Data from 2000 to 2017.” *PLOS ONE* 14 (10): e0223524. <https://doi.org/10.1371/journal.pone.0223524>.

Oliver, Dean. 2004. *Basketball on Paper : Rules and Tools for Performance Analysis*. Vol. 1st ed. Washington, D.C.: Potomac Books. <http://search.ebscohost.com/login.aspx?direct=true&db=e000xna&AN=388578&site=ehost-live>.

WINSTON, WAYNE L. 2009. “BASKETBALL STATISTICS 101: The Four-Factor Model.” In *Mathletics*, 187–94. How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football. Princeton University Press. <https://www.jstor.org/stable/j.ctt7sj9q.33>.