Final Report

Tim Vigers

04 December 2019

# Introduction

I’ve been watching basketball since 2011 and following the National Basketball Association (NBA) semi-obsessively since 2014. In that time the game has changed enormously with increased focus on the three point shot league-wide, a trend that has been well documented by analyst/journalists like Kirk Goldsberry and is part of a larger movement towards increased efficiency (i.e. taking good shots and avoiding bad ones) [1]. Based purely on the “eye test,” it seems that teams across the league are passing the ball more than ever in order to create better, more open shots for their most efficient offensive players.

So, this project was split into two distinct questions of interest: 1) Has total passing rate increased over time? and 2) Do teams that pass more tend to win more (i.e. does passing actually help)?

# The Data

Total number of passes made per season were manually downloaded from https://stats.nba.com/teams/passing/ and concatenated into a “long” dataset. These data were relatively well-organized to begin with and required minimal cleaning, but unfortunately only go back as far as 2013.

Traditional statistics, such as points, rebounds, etc. going back to the beginning of the NBA and the American Basketball Associatoin (ABA) were downloaded using an HTML scraping tool developed for this project (see Appendix for code). These data were also relatively clean, but teams that moved or changed names were assigned a unique three letter code corresponding to their current location (e.g. observations from the New Orleans Jazz were given the code “UTA” in order to group them with the rest of the Jazz data after their move to Salt Lake City). Also, seasons were designated using the numeric year of the first game of the season, (e.g. 2018 for the 2018-2019 season) in order to treat time as a continuous variable. There were no missing or excluded observations in these data, and counting statistics such as points, turnovers, etc. were converted to per-game measures in order to account for shortened seasons in 1998 and 2011. For these analyses I considered only data from after the ABA and NBA merger in 1976.

# Passing

## Mixed Model Selection

Prior to modeling the number of passes over time, I created a spaghetti plot with a line for each team (see Figure A1). There did not appear to be much of an overall trend. The total number of passes in a season appears to follow a normal distribution (Figure A2), so this outcome was modeled using a simple linear mixed model.

In order to test for a fixed effect of season on total number of passes made, I compared four linear mixed models. In the following models i indexes team, j indexes season, and x represents the season variable.

### Model 1: Random Intercept Only

### Model 2: Random Intercept and AR(1) Structure for Repeated Measures

### Models 3 & 4: Random Slope for Season

The last two models are the same as models 1 and 2, but with the addition of a random slope, so the random effects are:

with

The model with random intercept and random slope did not converge without the AR(1) structure for repeated measures, and the model with random intercept and AR(1) structure was the best by the Akaike information criterion (AIC) (Table A1).

Using loess smoothing to plot total number of passes suggested a potential cubic trend in the data (Figure A3). So once the final model was selected, I also tested the polynomial effects of season up to a quadratic term:

## Piecewise Model

In addition to a linear mixed model, I also tried a linear spline model with a knot at 2015, including random intercept and AR(1) structure for repeated measures:

The year 2015 was chosen based on the estimated break point according to Muggeo’s method [2] and implemented using his R package “segmented” [3] (see Appendix for code).

## Results

### Table 1: The Effect of Time on Total Passes Made

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Value | Std.Error | DF | t-value | p-value |
| (Intercept) | 24349.989 | 235.835 | 146 | 103.250 | <1e-04 |
| Season | -40.362 | 2004.510 | 146 | -0.020 | 0.984 |
| Season^2 | -1941.549 | 1404.499 | 146 | -1.382 | 0.169 |
| Season^3 | 360.020 | 1088.741 | 146 | 0.331 | 0.741 |
| Season^4 | -465.829 | 925.730 | 146 | -0.503 | 0.616 |

According to the linear mixed model, passing has not changed significantly since 2013.

### Table 2: Change in Total Passes Made After the 2015 Season

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Value | Std.Error | DF | t-value | p-value |
| (Intercept) | 164836.019 | 213588.290 | 148 | 0.772 | 0.441 |
| Season | -69.854 | 106.029 | 148 | -0.659 | 0.511 |
| Change in Slope | 0.181 | 0.154 | 148 | 1.178 | 0.241 |

Passing appears to increase slightly after 2015 according to the linear spline model, but the change in slope is not statistically significant (p = 0.24).

# Assists

## Model Selection

Winning percentage appears to be reasonably normally distributed (Figure A4), so I used normal theory linear mixed models to determine whether increasing the percentage of baskets assisted results in more wins. Model selection for this question followed a similar process to the passing question. I compared models with random intercept for team to models with random intercept for team and random slope, both with and without an AR(1) structure for repeated measures. However, in these models the outcome was regular season win percentage and the fixed effects were percetange of baskets assisted (“AST%”); average team age (“Age”); average team height (“Ht.”); average team weight (“Wt.”); team field goal percentage (“FG%”); and steals (“SPG”), blocks (“BPG”), points (“PPG”), and turnovers (“TPG”) per game. I chose to use percentage of baskets assisted because it should be a reasonable indicator of how much a team passes compared to others (a team with lower AST% is relying more on isolation plays).

Once again, the model with random intercept for team and AR(1) structure for repeated measures was the best by AIC (Table A2). However, during model selection I realized that there was a significant positive association between AST% and winning percentage, but that this effect goes away when adjusting for field goal percentage (Table A3). So, I conducted a mediation analysis (see Appendix for code, Figure A5 for mediation diagram) to try and determine whether field goal percentage mediates the effect of assists on winning [4]. The “mediation” package in R requires models without the AR(1) structure for repeated measures, so the mediation analysis was conducted using only a random intercept for team.

## Results

### Table 3: The Effect of Assists per Game on Winning Percentage

Fixed Effects

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Value | Std.Error | DF | t-value | p-value |
| (Intercept) | -277.331 | 44.178 | 1110 | -6.278 | <1e-04 |
| AST% | 0.372 | 0.085 | 1110 | 4.397 | <1e-04 |
| Age | 3.878 | 0.222 | 1110 | 17.471 | <1e-04 |
| Ht. | 0.508 | 0.566 | 1110 | 0.898 | 0.369 |
| Wt. | 0.338 | 0.062 | 1110 | 5.440 | <1e-04 |
| SPG | 2.349 | 0.345 | 1110 | 6.804 | <1e-04 |
| BPG | 3.766 | 0.350 | 1110 | 10.763 | <1e-04 |
| TPG | -2.280 | 0.216 | 1110 | -10.568 | <1e-04 |
| PPG | 0.847 | 0.052 | 1110 | 16.396 | <1e-04 |

Without adjusting for FG%, increasing AST% by 10 points can lead to a statistically significant (p = <1e-04) increase in winning of 3.72 percentage points on the season (or about 3.1 games). After adjustment for FG%, this effect is no longer significant (Table A3), although AST% still has a positive effect estimate.

### Mediation Analysis

There is a significant (p < 0.0001) mediation effect of FG% on the relationship between AST% and winning percentage. Field goal percentage accounts for approximately 67.5% of the association, meaning that on average a 10 point increase in AST% directly results in closer to a 1.217 point increase in winning percentage (about 1 game) on the season.

# Discussion and Limitations

The passing analysis appears to indicate that the overall passing rate has not increased significantly since 2013. However, it’s possible that the trend is more recent and there simply isn’t enough data to parse it out yet. Also, it’s possible that a subset of the elite teams have increased their passing rate, but this is being averaged out by teams without as many good passers. It would be interesting to look into passing trends among specific teams, rather than at a league-wide level.

As one might expect, the percentage of baskets assisted has a significant positive association with winning percentage. On average a 10 point increase in AST% can lead to an additional 1 - 3 games won per season. This may not sound like a great return on investment but for some teams, particularly in the competitive Western conference, this could be the difference between making the playoffs or not. Playoff games are an enormous source of revenue, with some teams earning as much as $2 million per game [5]. So those 1-3 extra wins could be worth millions of dollars, although increasing AST% by 10 points may be no easy task (particularly for teams that already have a high AST%).

It’s also remarkable that FG% mediates approximately two thirds of the postive effect of AST%. I repeated the mediation analysis with both FG% and AST% as the mediator, and the results suggest that FG% is mediating the positive effect of AST% and not the other way around. This is interesting because it implies in a general way that passing leads to better shooting, which in turn leads to increased wins. This makes some intuitive sense, because in general the “eye test” suggests that assisted baskets tend to come from more open shots than baskets scored in isolation plays. On the other hand, because the effect of AST% is no longer statistically significant after adjusting for FG%, one could argue that this is a case of complete mediation. This would imply that good shooting far outweighs passing in importance. Determining directionality in mediation analysis is difficult, particularly in the basic Baron and Kenny [6] framework that I used, so this relationship requires additional study. There are likely better, more modern methods but they were beyond the scope of this work.

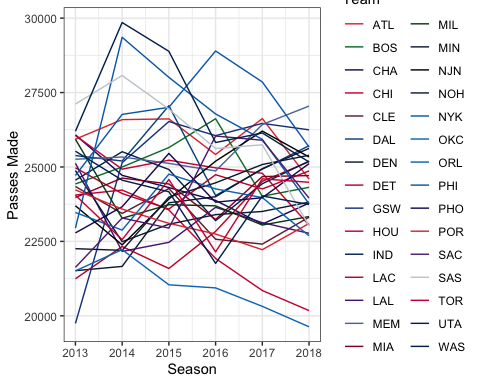
Finally, there are some general limitations to these analyses that would are important to address in future work. First, NBA teams do not play a balanced schedule, which means that playing bad teams more frequently should increase win percentage. It would have been helpful to adjust for strength of schedule, but these data were unfortunately not available to me. Perhaps most importantly though, it’s highly likely that looking at trends at a league-wide level results in vital information being averaged out. For example, average team height was not associated with increased wins (when we know that height is important in basketball), but this is likely because there isn’t much variation when comparing average team height. To really look at the effect of height it would be better to compare, for example, the offensive efficiency of specific tall lineups to a specific small ball lineups. To really determine the effects of passing, we need much more detailed information than is currently available.

# References

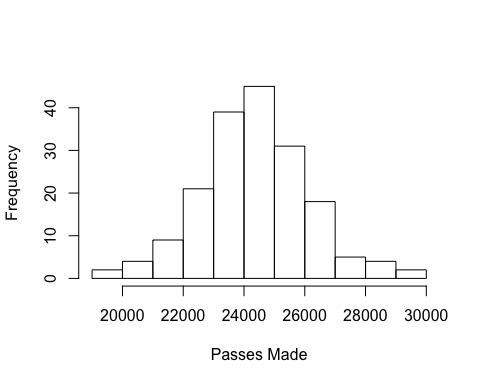
1. Goldsberry, Kirk. SPRAWLBALL: a Visual Tour of the New Era of the NBA. Mariner Books, 2020.
2. Muggeo, V.M.R. (2003) Estimating regression models with unknown break-points. Statistics in Medicine 22, 3055–3071.
3. Vito M. R. Muggeo (2008). segmented: an R Package to Fit Regression Models with Broken-Line Relationships. R News, 8/1, 20-25. URL <https://cran.r-project.org/doc/Rnews/>.
4. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R package for causal mediation analysis. UCLA Stat Stat Assoc. August 2014. <https://dspace.mit.edu/handle/1721.1/91154>. Accessed December 2, 2019.
5. Dignam, Devin, et al. “The Best NBA Markets of 2012-13.” The Wages of Wins Journal, <http://wagesofwins.com/2013/03/28/the-best-nba-markets-of-2012-13/>.
6. Baron, R.M. and D.A. Kenny, The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of personality and social psychology, 1986. 51(6): p. 1173.

# Appendix

## Figure A1: Total Passes by Season



## Figure A2: Distribution of Total Passes

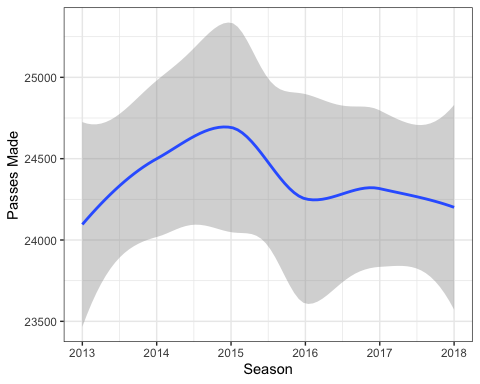


### Table A1: AIC of Passes Made Models

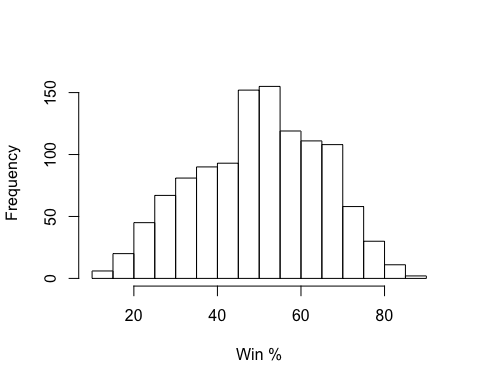
All models fit using ML estimation.

|  |  |  |
| --- | --- | --- |
|  | df | AIC |
| RI Only | 4 | 3159.781 |
| RI and AR(1) | 5 | 3120.225 |
| RI, RS, and AR(1) | 7 | 3124.225 |

## Figure A3: Overall Trend in Total Passes by Season, Loess Smoothing



## Figure A4: Distribution of Win Percentage



## Table A2: AIC of Win Percentage Models

All models fit using ML estimation.

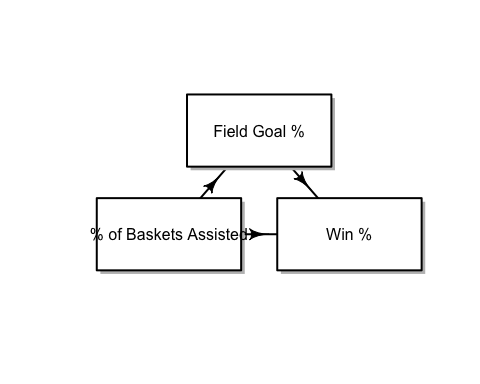
|  |  |  |
| --- | --- | --- |
|  | df | AIC |
| RI Only | 12 | 8495.325 |
| RI and RS | 14 | 8499.289 |
| RI and AR(1) | 13 | 8264.485 |
| RI, RS, and AR(1) | 15 | 8268.486 |

## Table A3: Effect of Assists on Win Percentage, Adjusted for FG%

Fixed Effects

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Value | Std.Error | DF | t-value | p-value |
| (Intercept) | -288.599 | 40.557 | 1109 | -7.116 | <1e-04 |
| AST% | 0.122 | 0.079 | 1109 | 1.530 | 0.126 |
| Age | 3.203 | 0.209 | 1109 | 15.331 | <1e-04 |
| Ht. | -0.195 | 0.521 | 1109 | -0.373 | 0.709 |
| Wt. | 0.436 | 0.058 | 1109 | 7.588 | <1e-04 |
| SPG | 2.937 | 0.319 | 1109 | 9.192 | <1e-04 |
| BPG | 3.447 | 0.322 | 1109 | 10.707 | <1e-04 |
| TPG | -2.983 | 0.204 | 1109 | -14.635 | <1e-04 |
| PPG | 0.099 | 0.070 | 1109 | 1.427 | 0.154 |
| FG% | 351.702 | 24.059 | 1109 | 14.618 | <1e-04 |

## Figure A5: Mediation Diagram



## Code

### HTML Scraping Tool

library(rvest)  
library(tidyverse)  
teams <- c("ATL","BOS","NJN","CHA","CHI","CLE","DAL","DEN","DET","GSW","HOU",  
 "IND","LAC","LAL","MEM","MIA","MIL","MIN","NOH","NYK","OKC","ORL",  
 "PHI","PHO","POR","SAC","SAS","TOR","UTA","WAS")  
# Scrape each team page  
all\_seasons <- data.frame()  
for (team in teams) {  
 url <- paste0("https://www.basketball-reference.com/teams/",team,  
 "/stats\_basic\_totals.html")  
 table <- url %>%  
 read\_html() %>%  
 html\_nodes("table") %>%  
 html\_table()  
 df <- as.data.frame(table[[1]])   
 df <- df[colnames(df) != ""] %>%  
 filter(Season != "Season",Season != "2019-20")  
 df[df == ""] <- NA  
 df <- as.data.frame(lapply(df, as.character))  
 colnames(df) <- c("Season","Lg","Tm","W","L","Finish","Age","Ht.","Wt.",  
 "G","MP","FG","FGA","FG%","3P","3PA",  
 "3P%","2P","2PA","2P%","FT","FTA","FT%","ORB","DRB","TRB",  
 "AST","STL","BLK","TOV","PF","PTS")  
 df$Team <- team  
 all\_seasons <- rbind.data.frame(all\_seasons,df)  
}

### Break Point

linmod <- lm(Passes.Made ~ Season,data = passing)  
segmented(linmod)

## Call: segmented.lm(obj = linmod)  
##   
## Meaningful coefficients of the linear terms:  
## (Intercept) Season U1.Season   
## -547034.4 283.7 -421.6   
##   
## Estimated Break-Point(s):  
## psi1.Season   
## 2015

### Mediation

# Mediation with FGP as mediator  
mod.y <- lmer(w\_perc ~ AST\_perc + Age + Ht. + Wt. + STL\_game + BLK\_game + TOV\_game +   
 PTS\_game + FGP + (1|Team),data = post\_merger)  
mod.m <- lmer(FGP ~ AST\_perc + Age + Ht. + Wt. + STL\_game + BLK\_game + TOV\_game +   
 PTS\_game + (1|Team),data = post\_merger)  
med\_fgp <- mediate(mod.m,mod.y,treat = "AST\_perc",mediator = "FGP")  
# Mediation with AST as mediator  
mod.m <- lmer(AST\_perc ~ FGP + Age + Ht. + Wt. + STL\_game + BLK\_game + TOV\_game +   
 PTS\_game + (1|Team),data = post\_merger)  
med\_ast <- mediate(mod.m,mod.y,treat = "FGP",mediator = "AST\_perc")  
# Mediation summary  
summary(med\_fgp)

##   
## Causal Mediation Analysis   
##   
## Quasi-Bayesian Confidence Intervals  
##   
## Mediator Groups: Team   
##   
## Outcome Groups: Team   
##   
## Output Based on Overall Averages Across Groups   
##   
## Estimate 95% CI Lower 95% CI Upper p-value   
## ACME 0.2514 0.1809 0.33 <2e-16 \*\*\*  
## ADE 0.1219 -0.0305 0.28 0.13   
## Total Effect 0.3733 0.2106 0.54 <2e-16 \*\*\*  
## Prop. Mediated 0.6747 0.4515 1.12 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Sample Size Used: 1148   
##   
##   
## Simulations: 1000

summary(med\_ast)

##   
## Causal Mediation Analysis   
##   
## Quasi-Bayesian Confidence Intervals  
##   
## Mediator Groups: Team   
##   
## Outcome Groups: Team   
##   
## Output Based on Overall Averages Across Groups   
##   
## Estimate 95% CI Lower 95% CI Upper p-value   
## ACME 7.94980 -2.15036 18.91 0.13   
## ADE 352.03485 305.92526 401.18 <2e-16 \*\*\*  
## Total Effect 359.98466 314.24555 404.83 <2e-16 \*\*\*  
## Prop. Mediated 0.02125 -0.00603 0.05 0.13   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Sample Size Used: 1148   
##   
##   
## Simulations: 1000