# Final Report

*Tim Vigers*

*03 December 2019*

## Introduction

Basketball has come a long way since James Naismith

## The Data

Team passing data 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 going back to the beginning of the NBA and ABA were downloaded using an HTML scraping tool developed for this project (see Appendix). These data were also relatively clean, but teams which 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 Utah Jazz data). 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 of passes over time 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 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 and j indexes season.

#### Model 1: Random Intercept Only

$$Y_{ij} = \beta_0 + \beta_1 x_j + b_{0i} + \epsilon_{ij}$$

$$b_{0i} \sim N(0, \sigma^2_{Team}) \text{ and } \epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$$

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

$$Y_{ij} = \beta_0 + \beta_1 x_j + b_{0i} + \epsilon_{ij}$$

$$b_{0i} \sim N(0, \sigma^2_{Team}) \text{ and } \epsilon_{ij} \sim N(0, R_i)$$

$$R_i = \sigma_\epsilon^2 \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 & \dots \\ \phi & 1 & \phi & \phi^2 & \\ \phi^2 & \phi & 1 & \phi & \\ \phi^3 & \phi^2 & \phi & 1 & \\ \vdots & & & & \ddots \end{bmatrix}$$

### 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 for season, so the random effects were

$$b_{0i} + b_{1j}x_j$$

with

$$b_{0i} \sim N(0, \sigma^2_{Team}) \text{ and } b_{1j} \sim N(0, \sigma^2_{Season})$$

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 made suggested a potential cubic trend in the data. So once the final model was selected, I also tested the polynomial effects of season, up to a quadratic term:

$$Y_{ij} = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \beta_3 x_j^3 + \beta_4 x_j^4 + b_{0i} + \epsilon_{ij}$$

$$b_{0i} \sim N(0, \sigma^2_{Team}) \text{ and } \epsilon_{ij} \sim N(0, R_i)$$

## 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:

$$Y_{ij} = \beta_0 + \beta_1 x_j + \beta_2 max(x_j - 2015, 0) + b_{0i} + \epsilon_{ij}$$

$$b_{0i} \sim N(0, \sigma^2_{Team}) \text{ and } \epsilon_{ij} \sim N(0, R_i)$$

## Results

### Linear Mixed Model

| | 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.

**Spline Model**

| | 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, but the change in slope is not statistically significant (p = 0.24).

# Assists

## Model Selection

Winning percentage appears to be reasonably normally distributed (Figure A3), so I used normal theory linear mixed models to determine whether increasing assists 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 and random intercept for team and random slope for season, both with and without an AR(1) structure for repeated measures. However, in these models the outcome is regular season win percentage and the fixed effects are average team age ("Age"); average team height ("Ht."); average team weight ("Wt."); team field goal percentage ("FG%"); and assists ("APG"), steals ("SPG"), blocks ("BPG"), and turnovers ("TPG") per game. 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 assists per game and winning percentage, but that this effect goes away when adjusting for field goal percentage (Table A3). So, I conducted a mediation analysis to try and determine whether field goal percentage mediates the effect of assists on winning. 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: Without FG%

| | Value | Std.Error | DF | t-value | p-value |
|---|---:|---:|---:|---:|---|
| (Intercept) | -211.893 | 46.434 | 1111 | -4.563 | <1e-04 |
| APG | 1.640 | 0.176 | 1111 | 9.317 | <1e-04 |
| Age | 2.962 | 0.263 | 1111 | 11.267 | <1e-04 |
| Ht. | 1.061 | 0.604 | 1111 | 1.758 | 0.079 |
| Wt. | 0.279 | 0.076 | 1111 | 3.664 | <1e-04 |
| SPG | 2.594 | 0.370 | 1111 | 7.004 | <1e-04 |
| BPG | 3.327 | 0.392 | 1111 | 8.491 | <1e-04 |
| TPG | -2.325 | 0.248 | 1111 | -9.377 | <1e-04 |

Without adjusting for FG%, increasing assists by 5 per game can lead to a statistically significant (p = <1e-04) 8.2 point increase in winning percentage on the season (or about 6.7 games). After adjustment for FG%, this effect is no longer significant (Table A3).

# Appendix

## HTML Scraping Tool

```r
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)
}
```

**Figure A1: Total Passes by Season**

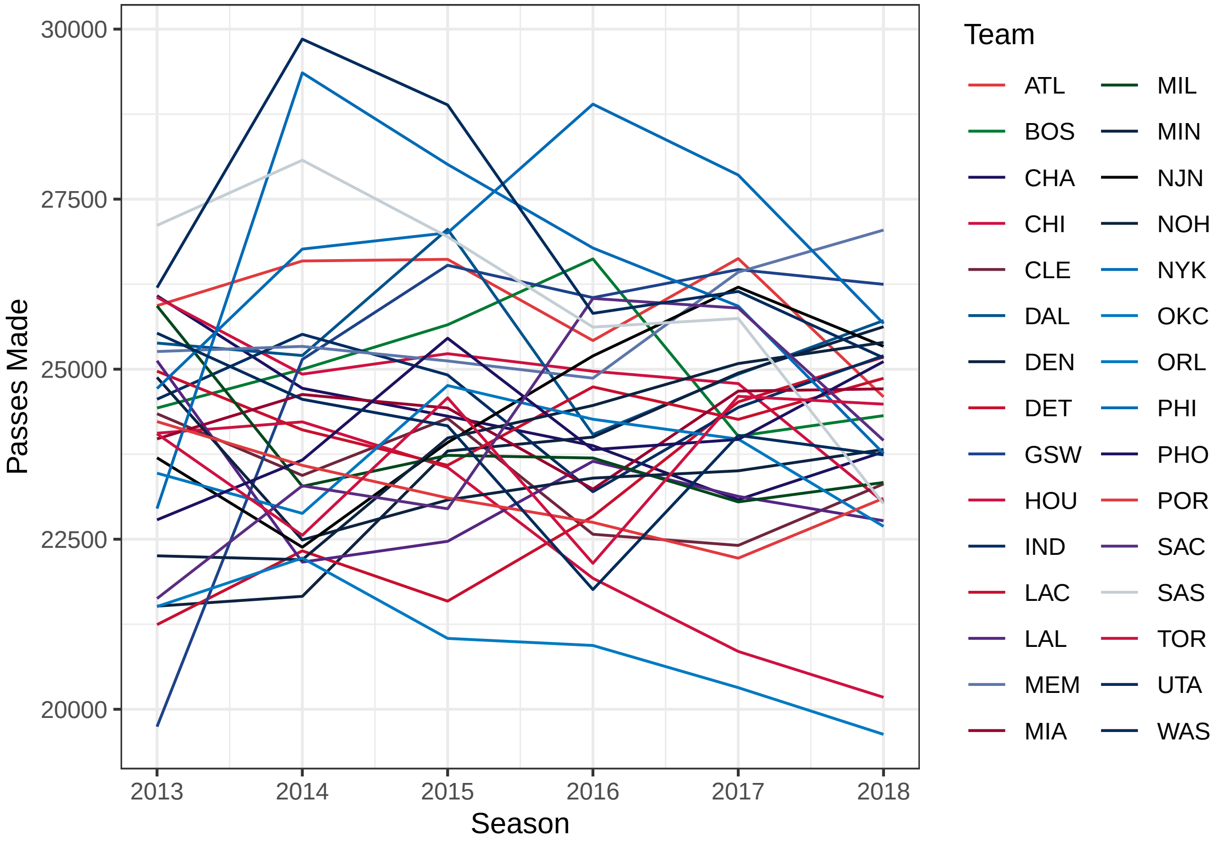
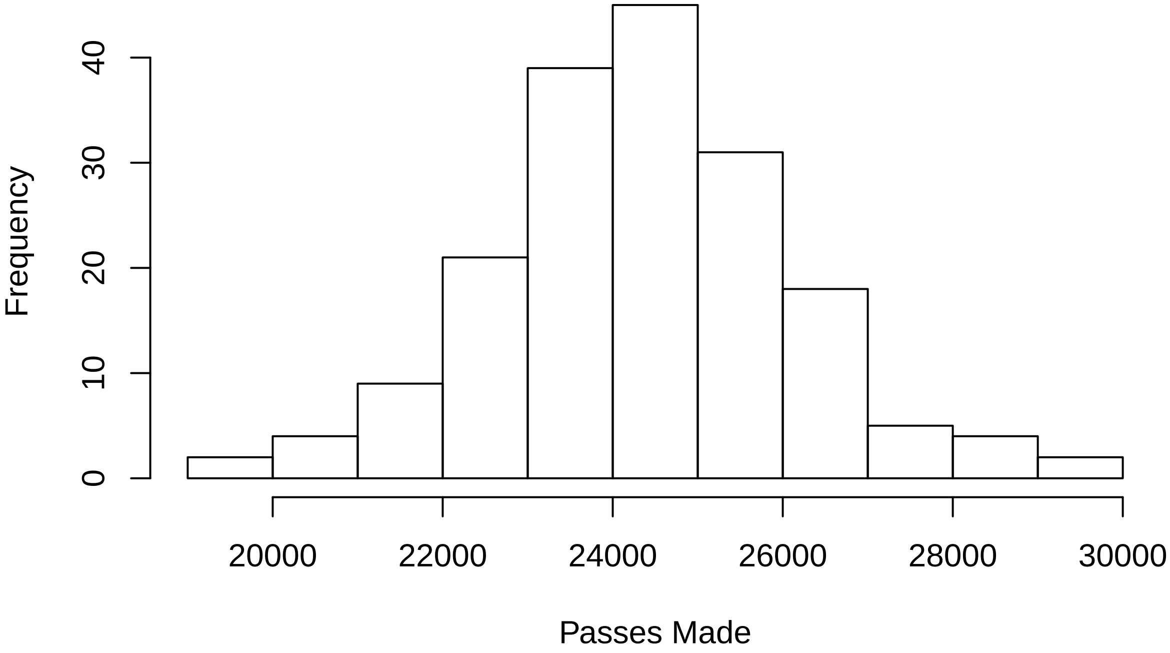


**Figure A2: Distribution of Total Passes**

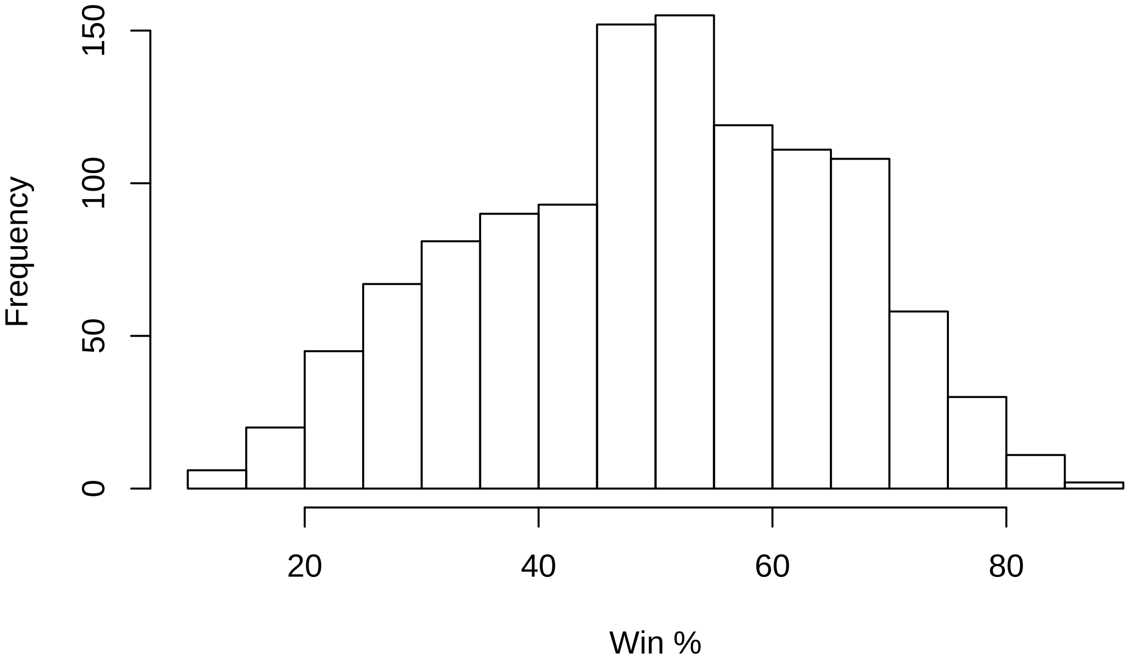


### Table A1: AIC of Passes Made Models

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: Distribution of Win Percentage



## Table A2: AIC of Win Percentage Models

| | df | AIC |
|---|---|---|
| RI Only | 4 | 9356.492 |
| RI and RS | 6 | 9360.492 |
| RI and AR(1) | 5 | 8873.876 |
| RI, RS, and AR(1) | 7 | 8877.876 |

## Table A3: Effect of Assists Adjusted for FG%

Table 6: Fixed Effects

| | Value | Std.Error | DF | t-value | p-value |
|---|---|---|---|---|---|
| (Intercept) | -288.491 | 41.698 | 1110 | -6.919 | <1e-04 |
| APG | -0.276 | 0.193 | 1110 | -1.429 | 0.153 |
| Age | 2.640 | 0.233 | 1110 | 11.329 | <1e-04 |
| Ht. | 0.298 | 0.541 | 1110 | 0.552 | 0.581 |
| Wt. | 0.335 | 0.067 | 1110 | 4.988 | <1e-04 |

|  | Value | Std.Error | DF | t-value | p-value |
|---|---|---|---|---|---|
| SPG | 3.075 | 0.332 | 1110 | 9.255 | <1e-04 |
| BPG | 3.132 | 0.349 | 1110 | 8.974 | <1e-04 |
| TPG | -2.863 | 0.222 | 1110 | -12.888 | <1e-04 |
| FG% | 395.964 | 23.331 | 1110 | 16.971 | <1e-04 |