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. Data 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 plotg 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 trend:

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

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

Piecewise Model

In addition to a linear mixed model, I also tried a piecewise regression

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

Appendix

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)
}
```

Figure A1: Total Passes by Season

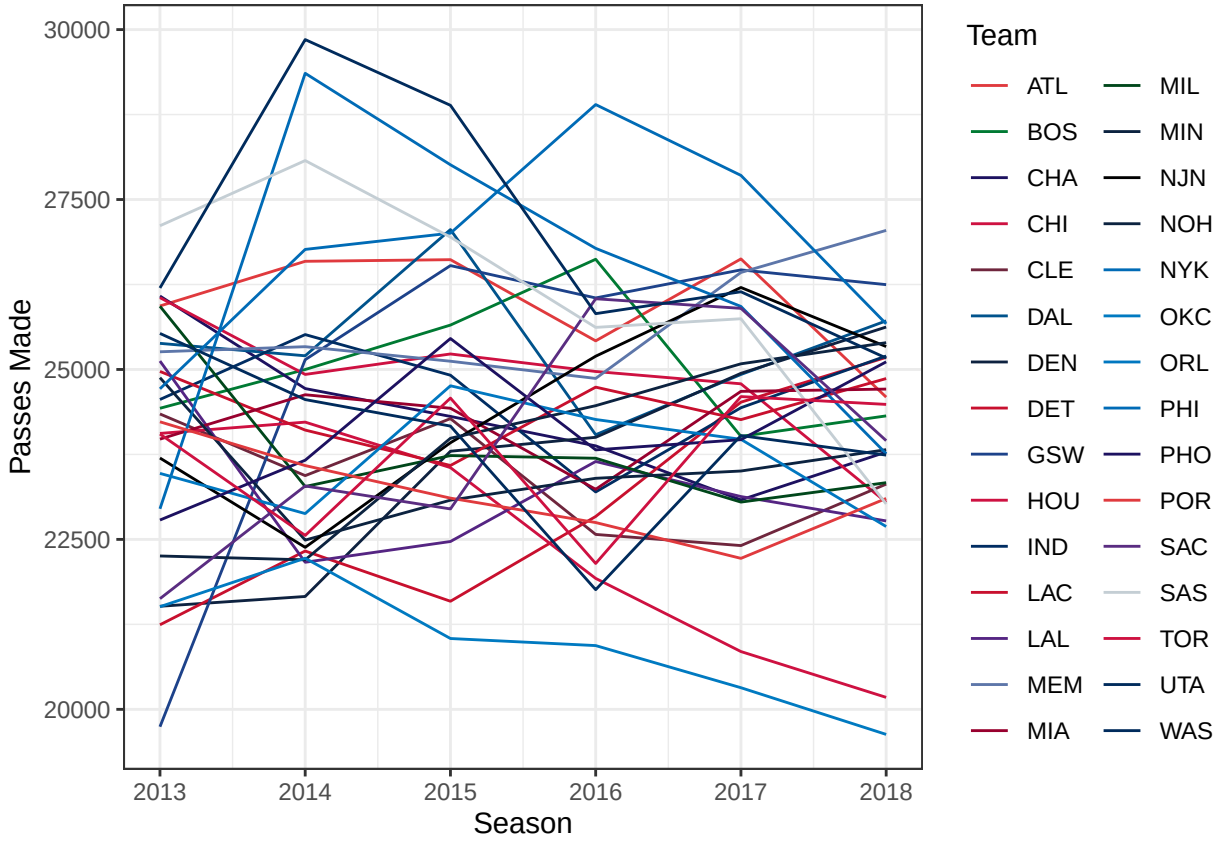

Figure A2: Distribution of Total Passes

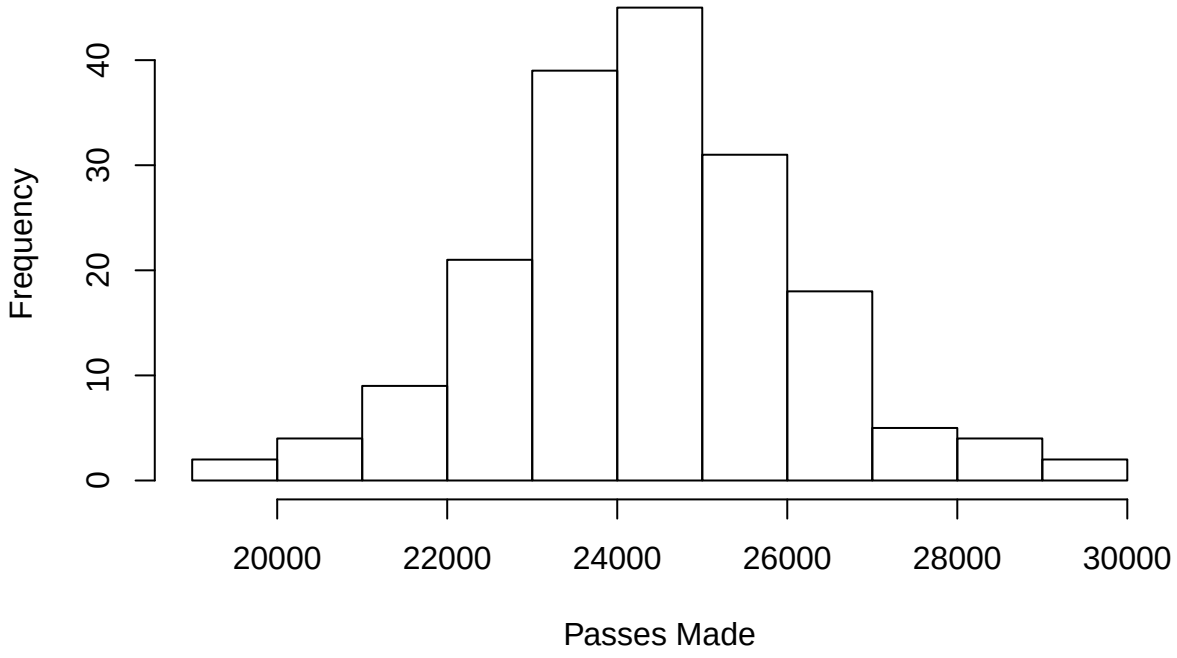

Table A1: AIC of Passes Made Models

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