Analyzing Trends in the NCAA 3 Point Shooting from 2003 to 2019

Seth Billiau, Jess Eng, Anna Li

December 12, 2019

Github Repo: `https://github.com/sethbilliau/stat139finalproject`

1 Introduction and Motivation

In the twenty-first century, the three-point shot rules the basketball court. Famed basketball executive and former Naismith Basketball Hall of Fame player Jerry West once said, "The 3-point line has changed the game so much."[1] And recently, reporters have even dubbed two-time MVP and NBA Champion Stephen Curry the "greatest shooter of all time" – largely for his accuracy and his sharp shooting at the three-point range.[2]

Not only do three-pointers rule the NBA, but they're also salient shots in unforgettable college basketball games. From last minute three-pointers that win crucial games (see Michigan's insane buzzer-beater win in 2018 NCAA tournament) to college players consistently showcasing remarkable three point shots, coaches and players devise game winning strategies around the three-point ring. Furthermore, professional basketball teams have started shaping the future of their team around top drafts and key three-point shooters from the NCAA; recently in the 2018 NBA Draft, star three-point shooter Trae Young from the University of Oklahoma went to the Atlanta Hawks at pick 5 where he's already been an outstanding contributor and shooter, averaging 28.7 points on 8.7 three point attempts per game so far this year.[3] All in all, three-point shooting exemplifies one of the most difficult and exciting aspects of the modern basketball game.

The introduction of the three-pointer has also had an impact on team strategy in the college game. Since its debut in 1987, the popularity of the shot continued to grow until it comprised over one fourth of the total number of shots made in the 2008-2009 NCAA Men's Basketball season. At this point, however, the NCAA changed the three-point line just before the 2008-2009 season by pushing it back from 19 feet and 9 inches to 20 feet and

1. "The 3-point line has changed the game so much...." - Jerry West quotes from BrainyQuote.com
2. Fromal, "Are Klay Thompson and Stephen Curry Really the Greatest Shooting Backcourt Ever?"
3. Young was actually drafted by the Dallas Mavericks, but was immediately dealt to the Hawks on Draft Day. "Trae Young Stats," ESPN.com.

9 inches. The number of three-point attempts dropped from 35.2 percent of all field goal attempts in 2008 to 34.4 percent in 2009 – the lowest percentage since 2000. Since then, however, three-point shots as a percentage of all field goal attempts have begun to increase once again making up over 37 percent of overall field goals in the 2018-19 season. In fact, three-pointers have rebounded so well that the NCAA Playing Rules Oversight Panel in June 2019 decided to move the Division I three-point line again to 22 feet, $1\frac{3}{4}$ inches (the same distance used in international basketball), adding a further degree of difficulty to the shot.

This paper, drawing upon data from 2003 to the present, analyzes how three-point line rule changes and other exogenous changes in the sport have been associated with changes in Division I NCAA teams three-point shooting tendencies. Specifically, we hope to answer the following questions: Is the 2008-2009 NCAA rule change associated with a statistically significant drop in three-point shots in the following years? Furthermore, if the rule change is associated with a significant decrease in three-point shots, when did the frequency of three-point shots begin to increase again? And what other changes in the sport may have led to this increase? To answer these questions, we will now turn to the data.

2 Data and Methods

Data Collection

We looked into various college basketball data sets and found `https://www.sports-reference.com/` to be the most comprehensive source of statistics that also required minimal data cleaning. From this site, we collected basic and advanced statistics for all NCAA Division I Schools from the 2003-2004 season to the 2018-2019 season.[4]

After collecting all the data, we took a few measures to clean it. First, we merged the data across all the years into one data frame. Then, we removed certain predictors that we knew wouldn't be helpful in looking at three-point data.[5] We also removed predictors that were highly colinear. Among these were the wins and losses broken down by whether the game was in conference, at home, or away. These variables were highly correlated with the total number of games won, which we kept in our data.

Next, from this full data set, we created three new files. Some schools didn't have data for all the years between 2003-2019, so we labeled that data as "incomplete", whereas we have a "complete" data set for schools that aren't lacking data for any of the years. Finally, the last last file is a subset of the "complete" data set, only including schools that made ever made it to the NCAA tournament during the time period. We did this for two main reasons: this gives us a less complex data set that might result in models that have a higher likelihood of converging, and it also gives us a subset of stronger teams that the public may

4. We started at 2003 to include an initial time period before the rule change. We realized later that it would have been better to start later on in order to have three equal six year blocks instead of one four year block with two six year blocks.

5. Removed predictors: Pace, ORtg, STL., ORB., MP, Rk, and X (added by R), full codebook in appendix.

have a stronger interest in.

Another factor that we took into consideration was the increase in the total number of games per year over time. To account for this, we divided certain predictors by the number of games, leaving us with a proportion rather than the original statistic. We also created our own time-independent predictor: an indicator for whether or not a team had one coach for all of the years between 2003-2019. We hypothesized that teams that did not experience a coaching change may perform better than their counterparts, as they may have a had more cohesive team due to more stable leadership. In addition, we created a new variable called time which is the original year subtracted by 2003, our earliest year. Times 0 and 15 correspond to the years 2003-2004 and 2018-2019. We did this because by making time start at 0, the intercepts of all our models are easily interpretable in reference to the 2003-2004 season. In addition, if we were to include polynomial terms, they would be less correlated for the time variable than for the year variable.

Methods

We originally wanted to find a best model that would predict the results of NCAA Division 1 teams for this upcoming season after the recent shift in the three-point shift. Using winning percentage (W-L%), we trained various models on data from the 2007-2008 season, validated on the 2008-2009 season (after the first three-point line shift), and planned on testing the models on the 2017-2019 data set to predict the W-L% of the 2019-2020 season. The idea was to look at whether there were certain predictors in the season leading up to the initial shift in the three-point line that affected team results after the shift, which would ultimately help us in determining how teams would perform in the upcoming 2019-2020 season. In this original analysis, we found that among a linear regression model, backward sequential variable selection model, forward sequential variable selection regression, ridge model, lasso model, decision tree model, and a pruned regression tree model, our backward sequential model performed the best on the validation data, with the lasso model coming in second. If we had continued the analysis, we would have tested the models upon the completion of the 2019-2020 season to find the best performing model.

However, upon further deliberation, we found that rather than focusing on prediction, it would be more informative to shift more towards interpretation and inference. Some key things we wanted to discover were characteristics that might make a team more or less affected by the shift in the three-point line and also which teams are most and less affected by the change. Therefore, instead of using W-L%, we changed our response variable to 3PAr, the percentage of field goal attempts from the three-point range. This allows us to look closer at how results may have been affected by strategy shifts that resulted from the shift in the three-point line and potential other outside factors.

EDA

As mentioned, earlier, We removed predictors from our original full data set that were highly colinear. Figure 1 shows the correlation matrix visualization of the variables related to wins and losses. As we can see from the dask red and purple, these variables are highly correlated with one another. Therefore, we ended up just keeping the total number of games won to reduce the colinearity.

In our EDA process, we found correlations between time and the different predictors from our data set. We noticed a strange result since the number of games was highly correlated with the time. Once we plotted the mean number of games played by teams over time in the NCAA, we noticed that number of games played each season in the NCAA steadily increased. Upon discovering this trend, we realized some of our original data needed to be modified. For instance, showing a team's number of threes made per season would naturally be higher in a season with more games since there are more attempts to make threes. We ended up fixing this by dividing all the game dependent predictors by the number of games.

Before we decided what models to explore, we explored how the (altered) variables 3P made, 3P attempted, and 3PAr (3-Point Attempt Rate: Percentage of field goal Attempts from 3-Point Range) varied over time, deciding on 3PAr as the response variable for our project.[6] We noticed that in our initial trend plot of the mean 3PAr for schools in a given year, the 3PAr at first increased, decreased, remained steady, and then increased in the last few years. Given this non-linear pattern, we knew that just fitting one simple linear regression might

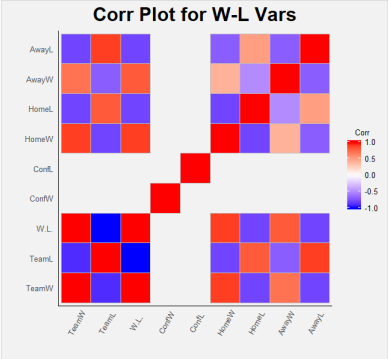

Figure 1: Correlation plot between predictors of wins and losses

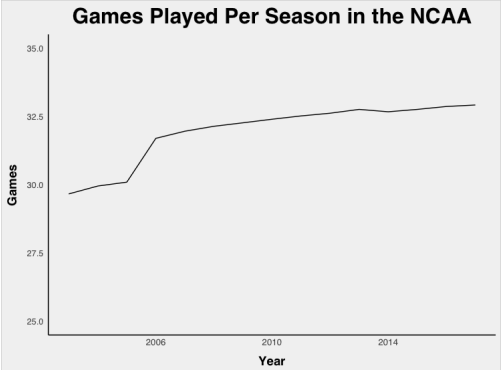

Figure 2: Games Played Continued to rise every Year

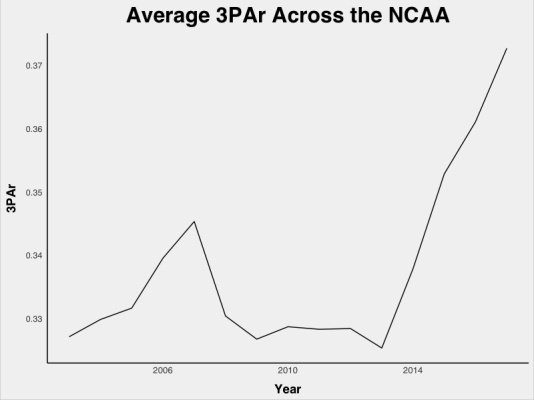

Figure 3: Field goals from 3-Point vs. Time

6. EDA for 3P and 3PA in appendix

not be the best model. Instead, that led us to explore other modeling decisions (highlighted in the section below) including mixed models and piecewise linear models (slopes that change for different chunks of time).

3 Models and Results

Building our model

We briefly considered a pooled OLS model relating time to 3PAr and an unpooled OLS model relating time to 3PAr with an indicator for School to provide the intercept. However, these OLS models involved several hundred different estimates for coefficients, becoming unwieldy very quickly. Additionally, these OLS models did not allow for variation among School-level coefficients across groups, fixing the time coefficient to be constant.[7]

Abandoning OLS models, we turned to hierarchical mixed-effects models. Given the longitudinal nature of our data with season-level statistics being collected for a single school in many seasons over time, we decided to pursue several different mixed-effects models clustered on the school variable. By clustering on schools, these model allowed us to specifically isolate the effect of time at the individual level and vary all coefficients accordingly if necessary (with random intercepts and/or slopes).

In the realm of mixed-effects models, we began by considering both a random intercepts model and a random slopes and intercepts model. After fitting both models and performing an ANOVA test, it became clear that the random slopes and intercepts model performed significantly better. This was to be expected as it is only natural for different teams with different systems and coaches to change the frequency with which they take three-point shots at different rates over time.

All of these models were fit as if there should be some constant linear rate of change in three-point shots over time. However, we have already established in the introduction of our paper that we do not expect this to be the case. Rather, we expect exogenous shocks (like rule changes or other cultural changes) to lead to a drastic shift in this rate of change. To model these exogenous shocks, we turned to piece-wise/segmented regression.

We chose to use the "segmented" library. Unfortunately, that meant that we had to use our simple pooled OLS model (Pooled 3PAr vs. time) given the constraints of this library to find breakpoints. However, we knew that if found significant breakpoints in our OLS model, we could also attempt to segment our random slopes and intercepts model using those same breakpoints. Then, we could conduct contrast t-tests to see if the breakpoints found with the OLS model were also significant for our mixed model.

With this strategy in mind, we searched for breakpoints using the segmented command which requires initial, a priori guesses for breakpoints in the time variable to make a determination. Using our EDA graphs, we suggested and successfully discovered two breakpoints: one between the 2006-2007 season and the 2007-2008 season (times 3 and 4) since the

7. Plots in Appendix.

rule change was announced during this point in time and one around the 2012-2013 and 2013-2014 seasons. Using the Davies Test, we also successfully verified that these breakpoints were statistically significant. This means that at both of these inflection points, the difference between the slopes of the two lines was significantly different from zero.

Given these breakpoints for our OLS model, we included a categorical indicator called "era" for whether or not a measurement was in one of the three periods of time in our OLS segmented model: 2003-2006, 2007-2012, and 2013-2018. We included "era" as a fixed, time-constant predictor called "era" in our random slopes and random intercepts model. This assumes that the effect of being in a particular era is uniform across teams and also allows us to perform inferences directly on our beta coefficients.[8] We then conducted contrast t-tests, determining that the slopes of all three fixed effects lines were significantly different than each other. All slope coefficients were also statistically different from zero, meaning that at no point was 3PAr simply remaining constant over time. The values of these slopes were 0.00386, -0.00255 and 0.0118 for each era respectively. These slopes are easily interpretable: for example, the last slope means that the average team between 2012 and 2018 experienced a 1.118% increase in the proportion of shots from 3 point range relative to the total number of shots per year. To make such a strategic change, a coach could instruct their players to take around 20 more shots from three point range in a season, holding the number of two-point shots constant.

This piece-wise, random slopes and intercepts model yielded both the lowest AIC of all models tested.[9] We also ran an ANOVA F-test which showed that the segmented, random slopes and intercepts model was a significant improvement on our unsegmented random slopes and intercepts model. Our final model is plotted with our team data in Figure 4.

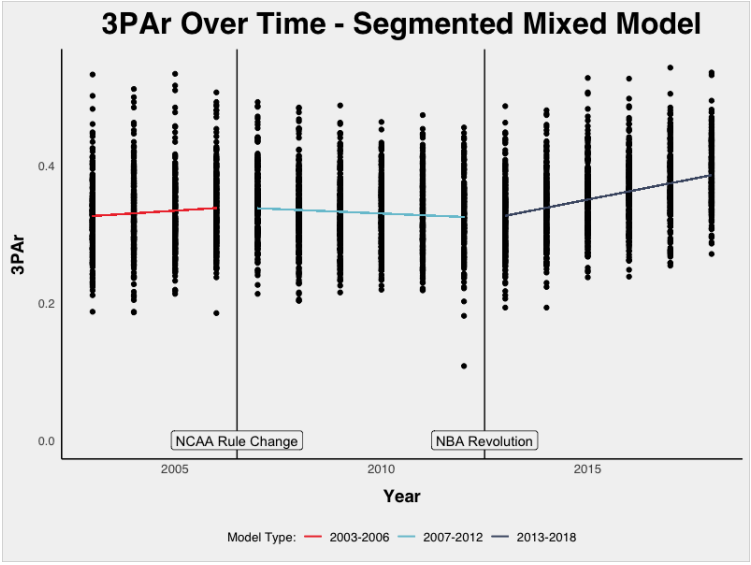

Figure 4: Final Piece-wise, random slopes and intercepts model

8. Since we only cared about inference with our betas, we did not include era as a random effect.

9. Table 1 in appendix.

Explaining the Breakpoints

We hypothesize that both of these statistically significant breakpoints in our model were caused by exogenous shocks in the world of basketball. We believe the existence of the first breakpoint can be attributed to the announcement of the 2008-2009 rule change. Logically, moving the line should effect how many three point shots are being taken and how the three is being used as a strategic shot.

The second breakpoint between the 2012-2013 and 2013-2014 seasons is harder to explain. We hypothesize that its existence could be due to what we will call the NBA three point revolution a cultural revolution in pro-basketball that seems to be mirrored in college, perhaps with some sort of lagging effect. In the 2012-2013 season, the duo of Stephen Curry and Klay Thompson lit up the NBA with a record 483 between the two of them. Curry also set another NBA record with 270 individual threes, leading the Golden State Warriors to the 6 seed in the Western Conference.[10]

The emergence of the Splash Brothers was indicative of a larger cultural change in the NBA, and perhaps the world of basketball as a whole, with players beginning to shoot more threes at an accelerated pace. From 2002 to 2011, the number of three point attempts per game in the NBA increased by just 4 attempts per game. Then from 2011 to 2015, the number of three point attempts increased by 5 attempts per game.[11] Though our group has not taken a deep dive into NBA statistics, we theorize that this change in the styles of NBA teams could be association with a lead-lag effect in college threes, especially given that the NBA a change appears to begin a year before the observed change in college teams.

To be clear, our group has not done any rigorous analysis of NBA statistics to prove this lead-lag theory, nor can we claim to the prove the causality of either these claims. We offer these explanations as plausible theories to explain the empirical patterns observed in our dataset and reflected in our final model.

Additional Analysis and Exploration: Analyzing Teams

Using this model, we also attempted to answer some of the following questions: What characteristics made teams more prone or less prone to the rule change at the 3-point line? In order to do so, we attempted a few different methods. We wanted to find out whether game wins or the number of 3pt shots affected if the team was more or least prone to the change. In the first method, we took beta coefficients from the overall mixed model with the random slopes and random intercepts with school clusters. We looked at two groups: 50 schools with the highest absolute value for slopes and 50 schools with a slope value closest to zero for the response variable 3PAr over time. Then through a formal hypothesis test t-test, we compared whether the average number of wins and number of three-pointers made over time for that specific cluster of 50 schools was different than the average number of wins and three-pointers for the other group. The null hypothesis was that the two means

10. Buckley, "Visualizing Splash Brothers' Climb Up NBA's All-Time 3-Point List."

11. "NBA League Averages - Per 100 Possessions." Basketball-Reference.com.

(average number of wins/three-pointers made) would be the same between the groups and the alternative hypothesis is that they would be different. Ultimately, we found a p value of 0.8064 and 0.3619 for difference in number of wins and 3 pointers made overtime respectively. Therefore, we failed to reject the null hypothesis and concluded that the differences between the two averages for these groups are not statistically significant.

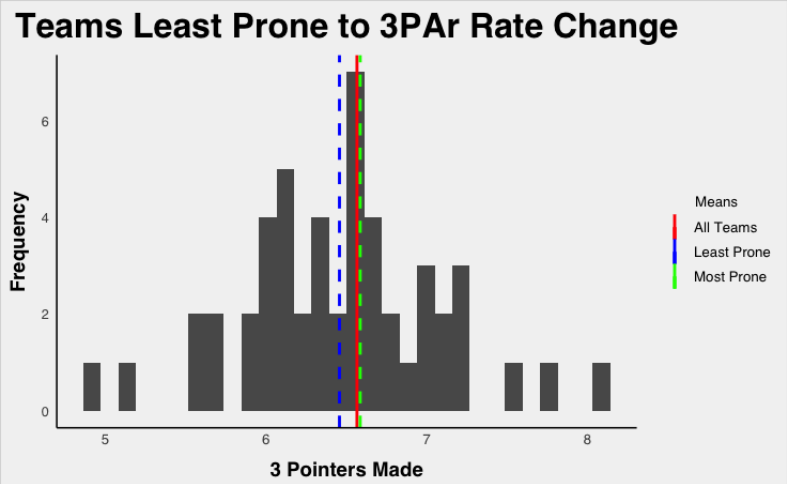

Figure 5: Histogram comparing mean 3 Pointers Made Per Game for All Teams, and Teams Least Prone and Most Prone Rule Change

In the second method, we considered the segmented regression model with breakpoints. We wanted to determine whether the teams most affected in each of the three eras separated by the breakpoints had slopes that differed from the mean number of wins and number of three pointers made for teams least affected by the rule change. As we tested all eras through a t-test, we obtained p values 0.8171 and 0.05021 for difference in number of wins and 3 pointers made overtime respectively. Therefore, since we set our alpha to be $\alpha = 0.05$ we failed to reject the null hypothesis that the averages for the teams most prone and least prone to the changes differed significantly.

Another approach that we took was to include an indicator of whether there was a coaching change during the 2003-2018 time period. We thought this could add more information to our models, as teams that have more steady leadership may perform better. However, we were unable to get the models including this predictor to converge, and thus gained no additional information.

We also wanted to look at similar models, but for individual teams. To do this, we created linear models only on subsetted data from the University of Virginia and Villanova.[12] We decided on these two teams because they are the past two NCAA Tournament champions, but we could could use the same type of model to compare any teams of our choosing. We see this as a further extension to our project, proving that it is possible to use this analysis

12. We fitted a random slopes and intercepts model here as well, but with only two teams the linear model actually performed better in terms of AIC.

to highlight how teams may have differed in strategies across these different eras.

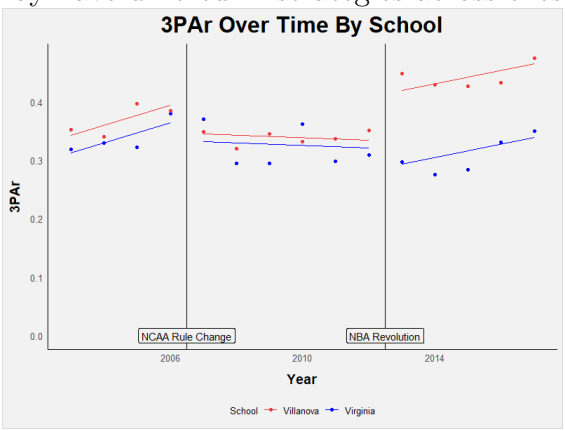

Figure 6: 3PAr Over Time By School

4 Conclusion and Discussions

Some main takeaways from our analysis and possible extensions to our project are as follows:

1. The rate at which the number of attempted three point shots relative to the total number of three point shots changed over time exhibited three distinct non-zero trends between 2003-2006, 2007-2012, and 2012-2018. While initially increasing over time from 2003-2006, 3PAr decreased between 2007-2012. In 2013, 3PAr began to increase again at an accelerated rate - a rate even larger than the annual rate of growth between 2003-2006.

2. We hypothesize that the first breakpoint and subsequent decrease in 3PAr over time beginning in 2007 is due in part to the announcement of the 2007-2008 NCAA rule change. We also hypothesize that the second breakpoint in 2013 could be due to a lead-lag relationship between the NBA basketball and NCAA basketball. We are more confident in our prior claim than our latter claim, though further research could be done to flesh out both of these hypothesis. It would be interesting to do more research into other exogenous factors that may have led to the breakpoints. We would be especially interested in analyzing any possible lead-lag relationship between the NBA and NCAA more in depth in a future project.

3. We attempted to use our model to analyze teams that were more or less similar to the average trend for a typical team in each era. While we were able to identify which teams were extreme, we were unable to pinpoint any commonalities between such teams. However, we have demonstrated how a future extension to our project could build off of the work we have done to compare groups of teams like in the example of Villanova and Virginia.

4. If future studies want to determine how trends in three-point averages vary over time and become affected by different predictors, there are various ways to build upon our study. Instead of clustering on school as we did in this project, another method could consider conference clusters to see if there are any statistically significant results. One reason we did not attempt to cluster on conferences is that many conferences are transient – they often add or subtract teams such as the PAC 12 with the additions of Colorado and Utah in 2011. However, there are certainly ways to use conference clustering to draw conclusions from these data sets.

A First Appendix: Codebook

The following variables are in the our "complete_data_clean.csv", "incomplete_data_raw.csv", and "tourney_data_clean.csv" files. Each row in the dataframe contains season-level statistics for a specific NCAA Division I team:

School: NCAA Division I School for a given row

year: The year that in which the season for a given row began (the 2003-04 season corresponds to year 2003)

G Games played by the team in a given year

TeamW Proportion of games won by the team in a given year

TeamL Proportion of games lost by the team in a given year

SRS Simple Rating System - A rating that takes into account average point differential and strength of schedule. The rating is denominated in points above/below average, where zero is average. Non-Division I games are excluded from the ratings.

SOS Strength of Schedule - A rating of strength of schedule. The rating is denominated in points above/below average, where zero is average. Non-Division I games are excluded from the ratings.

Tm. Points scored by the team Per Game in a given season

Opp. Total points conceded by the team Per Game in a given season or total points scored by opponents of the team in a given season

FTr. Free Throw Attempt Rate - Number of Free Throw Attempts Per Field Goal Attempt

X3PAr 3-Point Attempt Rate: Percentage of field goal Attempts from 3-Point Range

TS. True Shooting Percentage - A measure of shooting efficiency that takes into account 2-point field goals, 3-point field goals, and free throws.

TRB. Total Rebound Percentage - An estimate of the percentage of available rebounds a player grabbed while he was on the floor.

AST. Assist Percentage - An estimate of the percentage of teammate field goals a player assisted while he was on the floor.

BLK. Block Percentage - An estimate of the percentage of opponent two-point field goal attempts blocked by the player while he was on the floor.

eFG.	Effective Field Goal Percentage - this statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal.
TOV.	Turnover Percentage - an estimate of turnovers per 100 plays.
FT.FGA	Free Throws Per Field Goal Attempt
FG	Field Goals Made Per Game in a given year
FGA	Field Goals Attempted Per Game in a given year
FG.	Field Goals Percentage in a given year
X3P	3-Point Field Goals Made Per Game in a given year
X3PA	3-Point Field Goals Attempted Per Game in a given year
X3P.	3-Point Field Goals Percentage in a given year
FT	Free Throws Made Per Game in a given year
FTA	Free Throws Attempted Per Game in a given year
FT.	Free Throw Percentage in a given year
ORB	Offensive Rebounds Per Game in a given year
TRB	Total Rebounds Per Game in a given year
AST	Assists Per Game in a given year
STL	Steals Per Game in a given year
BLK	Blocks Per Game in a given year
TOV	Turnovers Per Game in a given year
PF	Personal Fouls Per Game in a given year

B Second Appendix: Additional EDA

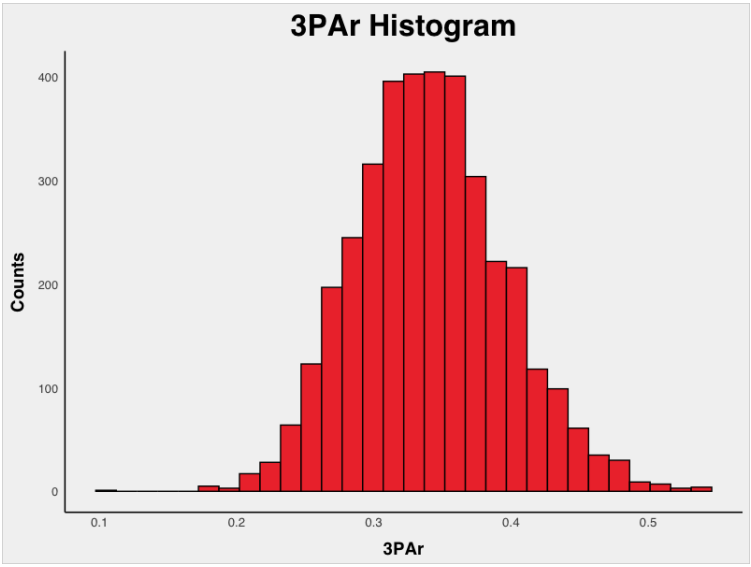

Figure 7: Histogram of response variable Normality Check 1

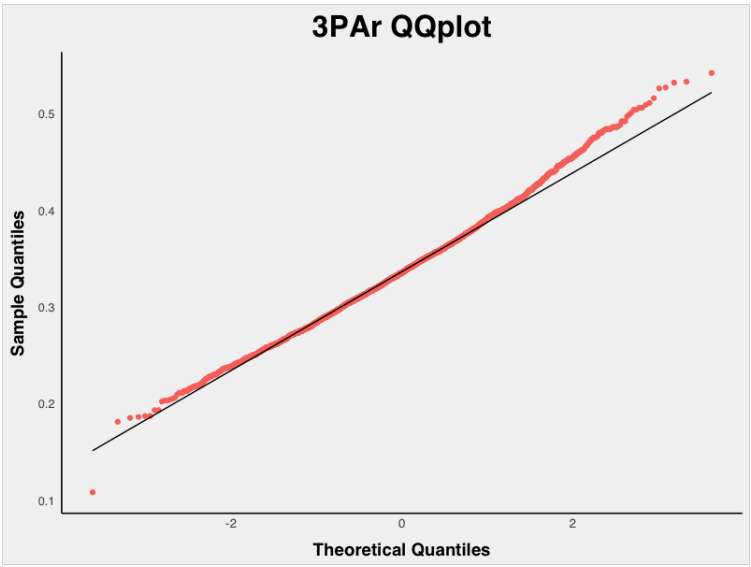

Figure 8: QQplot of response variable Normality Check 1

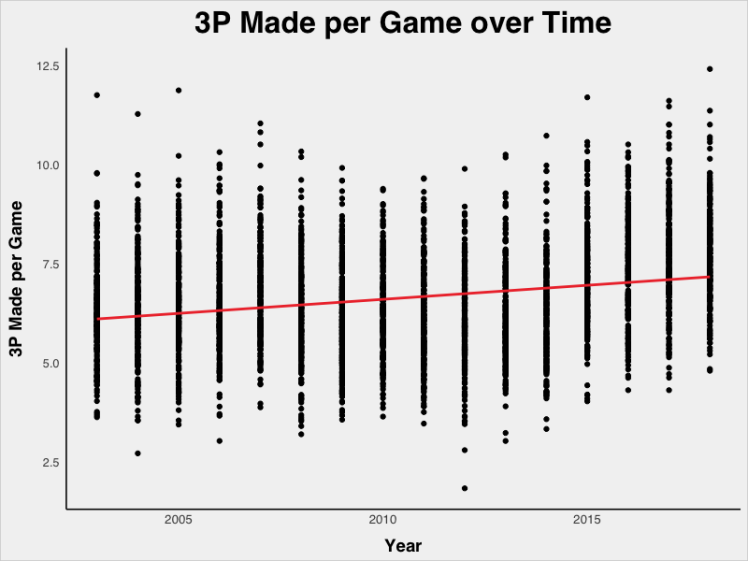

Figure 9: 3P Made EDA

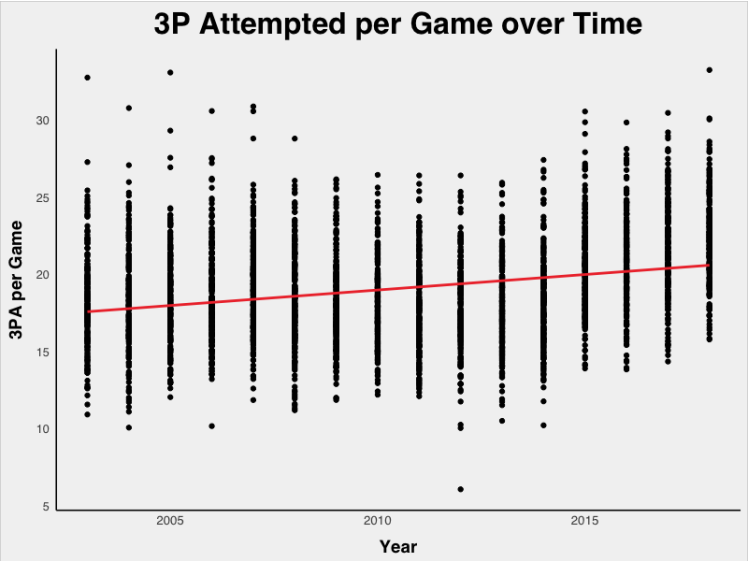

Figure 10: 3P Attempted EDA

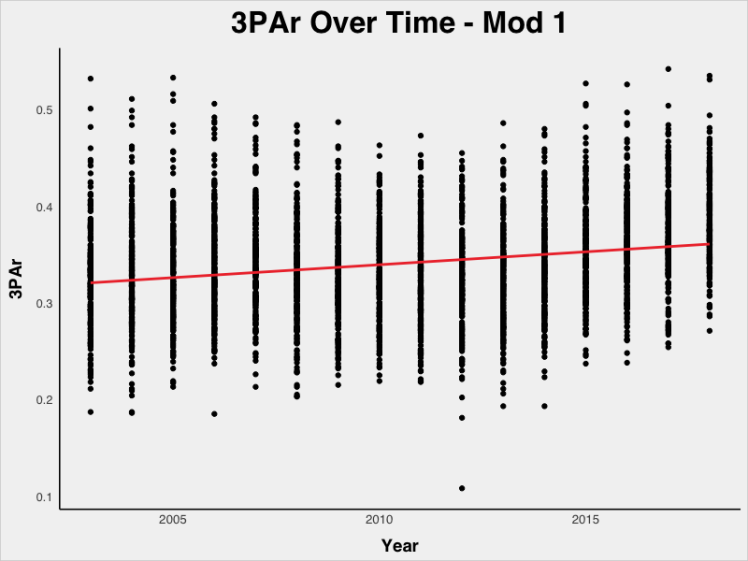

Figure 11: Baseline Pooled Linear Model

C Third Appendix: Additional Tables

Table 1: AIC for each Model

Model Name	AIC
Simple OLS unsegmented	-11243.57
Simple Random Intercepts	-11979.96
Simple Random Slopes and Intercepts	-12163.78
Segmented OLS	-11444.87
Segmented Random Slopes and Intercepts	-12450.06

Table 2: ANOVA Tests

Model 1	Model 2	AIC Change (mod2 - mod1)	BIC Change (mod2 - mod1)	p-value
Simple Random Intercepts	Simple Random Slopes and Intercepts	-183	-171	2.2e-16
Simple Random Slopes and Intercepts	Segmented Random Slopes and Intercepts	-330	-305	2.2e-16

D Fourth Appendix: Explaining our R files

We have submitted a number of .R files with our .pdf. Here is a guide to navigating them.

styleguide.R: Styling for our ggplot graphics

packages.R: Loading all packages for our project

cleaner.R: Code to clean our raw .csv files.

eda.R: Code with all of our visualizations. We also made this in rmd format and included a knitted .pdf as an additional appendix.

models.R: Fitting and evaluating our models. We also made this in rmd format and included a knitted .pdf as an additional appendix.

project.R: All other project-related exploration. We also made this in rmd format and included a knitted .pdf as an additional appendix.