Analyzing Consistency of Three-Point Defense by NBA Teams

Aristaeus Chang, Malcolm Ehlers, Tyshanti Montgomery, Malek Shafei

2024-04-29

Introduction

Since the early 2010's, the significance of three-point shooting in the NBA has surged dramatically. One of the main reasons for this popularity was that researchers found that the expected value of three-pointers to be higher than two-point shots. Three-pointers became an efficient shot, which led to teams prioritizing it as a core element of offensive strategy. This evolution has compelled NBA teams to prioritize their strategies for defending against three-point shots. This is necessary, as taking away a team's three-point shot in the modern NBA disrupts the offensive flow for a team. Consequently, this prompted several questions that we wanted to figure out. Firstly, how can we accurately track three-point defense? Is it primarily a matter of poor shooting by opponents or effective defense by teams? We wanted to find out if being good at defending three-pointers was something that stayed the same for teams over time.

To answer these questions, we used ridge regression models to analyze the defensive coefficients of each NBA team. By comparing three-point defense between the first and second halves of the season, we found that the correlation coefficient between the models trained on the first and second halves of the season was 0.080. This indicates that there is a weak positive linear relationship between the models, and suggests that there is little consistency in the three-point defensive performance of NBA teams over the course of the season. The weak correlation between the models suggests that the factors influencing three-point defense do not remain stable throughout the season.

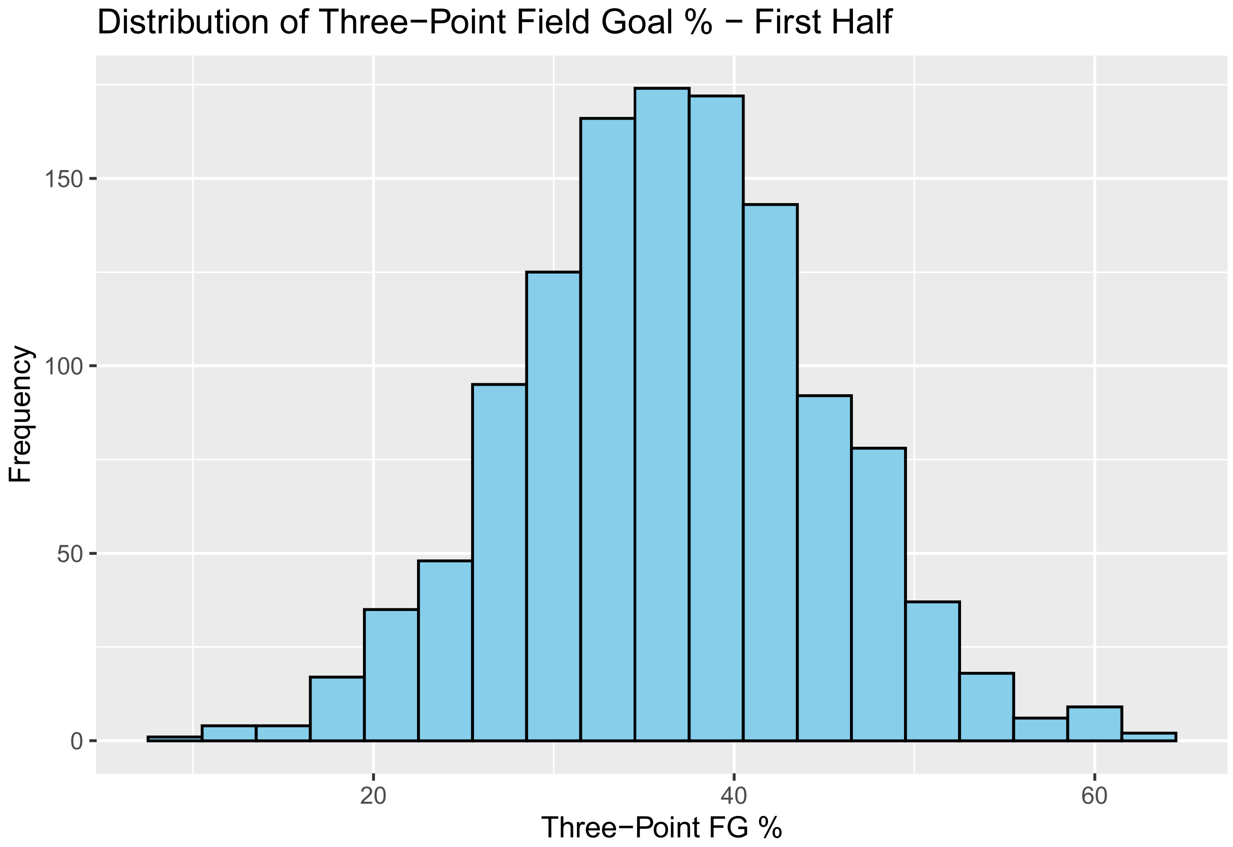
Data

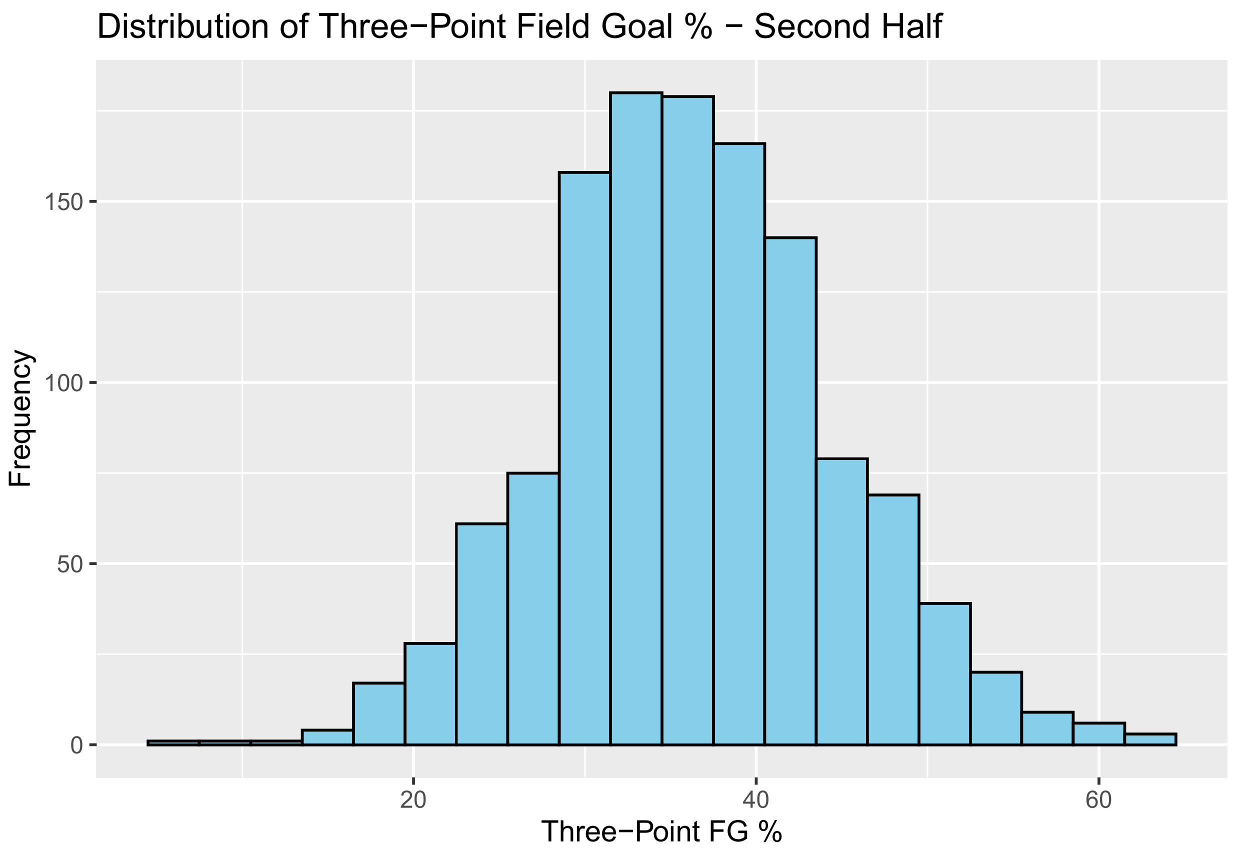
```
## 1.08 sec elapsed
```

To begin our analysis, we load NBA team box score data from the 2023-24 season using the load_nba_team_box() function from the hoopR package. For every game, there were two rows, each with basic data like points, field goal attempts and makes on the home and away team. For the purpose of our analysis, we selected only the team's name, the opponent's team name, and the team's three-point field goal percentage which we reframe to use as the opponent's three-point field goal defense. We also excluded the All-Star game as well as playoff games to focus solely on regular season games for this current NBA season. We proceeded to create a design matrix for the model, and encoded the defensive team as -1, the offensive team as 1, and all other teams as 0 for every game in our filtered dataset.

Next, we split the seasons into first and second halves.

We visualize the distribution of three-point field goal percentage for both the first and second halves of the season using histograms. This helps us understand the variability and skewness of three-point shooting performance.


Distribution of Three−Point Field Goal % − First Half
Frequency
150
100
50
0
20
40
60
Three−Point FG %



We observe the distribution of all three-point percentages by a team in a game across the first and second half of the season. Both histograms reveal a unimodal beta distribution centered around 0.35, with the majority of the data falling between 0.2 and 0.5. There appears to be a slight positive skew in the distribution.

Three-point percentage is bounded between 0 and 1, so we apply a log odds transformation to the three-point field goal percentage to normalize the data and improve model performance.

For each half of the season, we create a design matrix with 60 new columns representing each team's defense and their offense. For each row, if a team's offense is represented, their offense column is assigned a value of "1". If a team's defense is represented, their defense column is assigned a value of "-1". All other team offense and defense columns are assigned a value of "0". In our design matrix, we also include the log odds of the offensive team's three-point field goal percentage.

Methods

The next step in our analysis involves building statistical models to predict three-point shooting percentages.

For our model, we split the 2023-2024 season into halves and plan to train the same model on each half of the season. Our response variable is the log odds transformed percentage of three-point shots made relative to total shots attempted. Our primary metric of interest is the correlation of parameters between the two halves of the season which allows us to assess the consistency of the factors influencing three-point defense over time.

We will employ ridge regression as our primary modeling technique to analyze the consistency of NBA teams in defending three-point shots. Ridge regression addresses multicollinearity, which is prevalent in our data due to repeated match ups between teams. This is appropriate for our model because it allows us to examine the relationship between various factors influencing three-point defense while managing the potential correlation among these factors. The regularization provided by ridge regression helps prevent overfitting, ensuring that our model generalizes well to new data and is useful when dealing with multicollinearity.

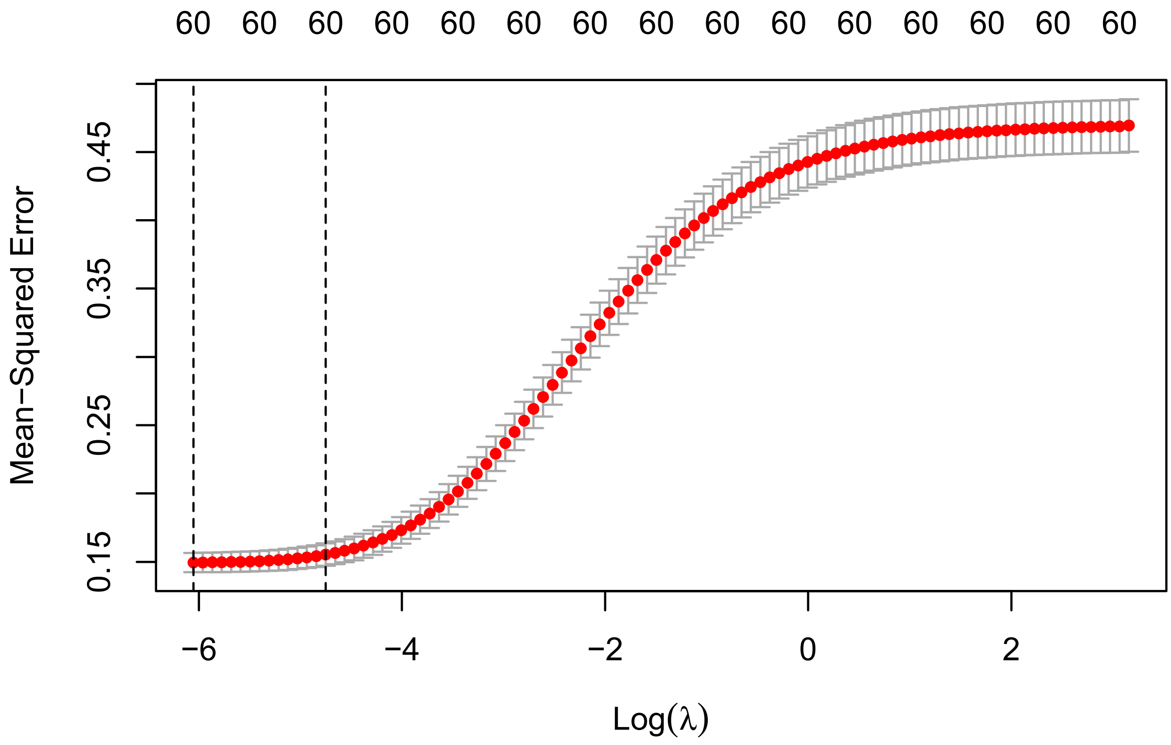
The assumptions of ridge regression include linearity, constant variance, and independence. Linearity assumes that the relationship between the predictors and the response variable can be adequately captured by a linear model. Constant variance implies that the variance of the errors is consistent across all levels of the predictors. Independence assumes that the observations are independent of each other. While the independence assumption seems false due to repeated match ups, ridge regression will address this by shrinking parameter estimates towards zero, in effort to reduce the impact of multicollinearity.

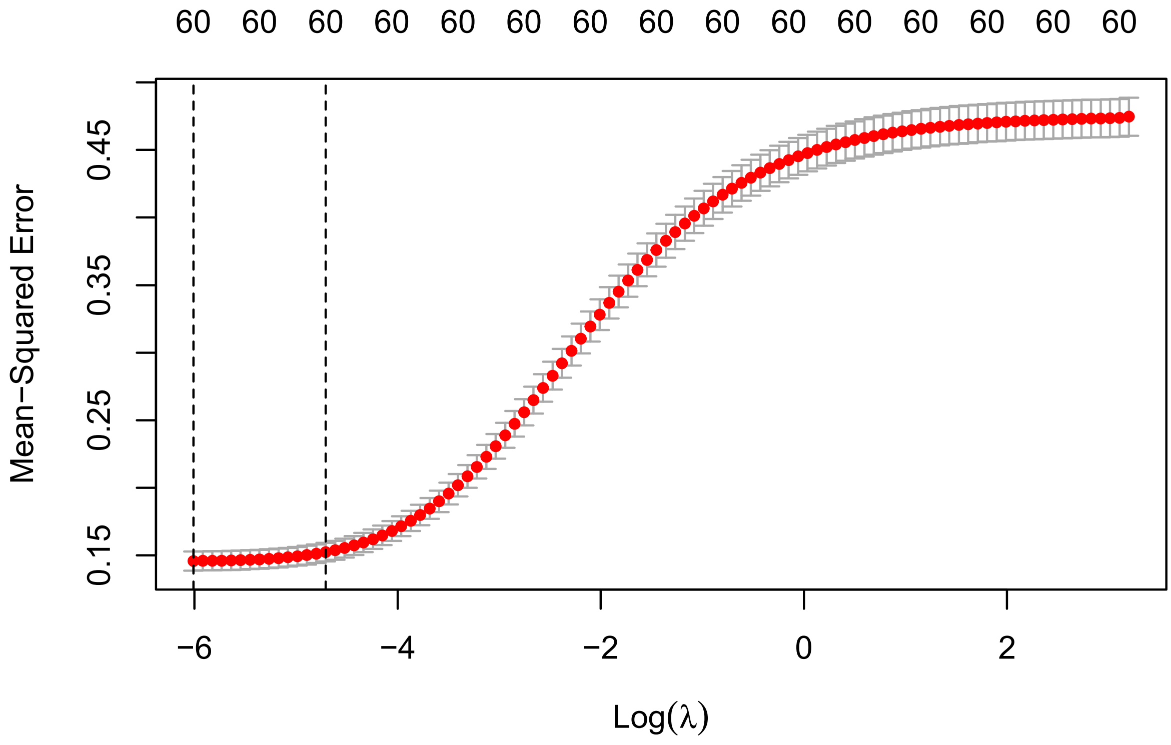
We utilize the glmnet package in R to conduct our ridge regression analysis with cross-validation on each half of the season. These models use whether each team is present on offense and defense as the predictor variables, and log odds three-point percentage as the response variable.

```
## Loaded glmnet 4.1-8
```

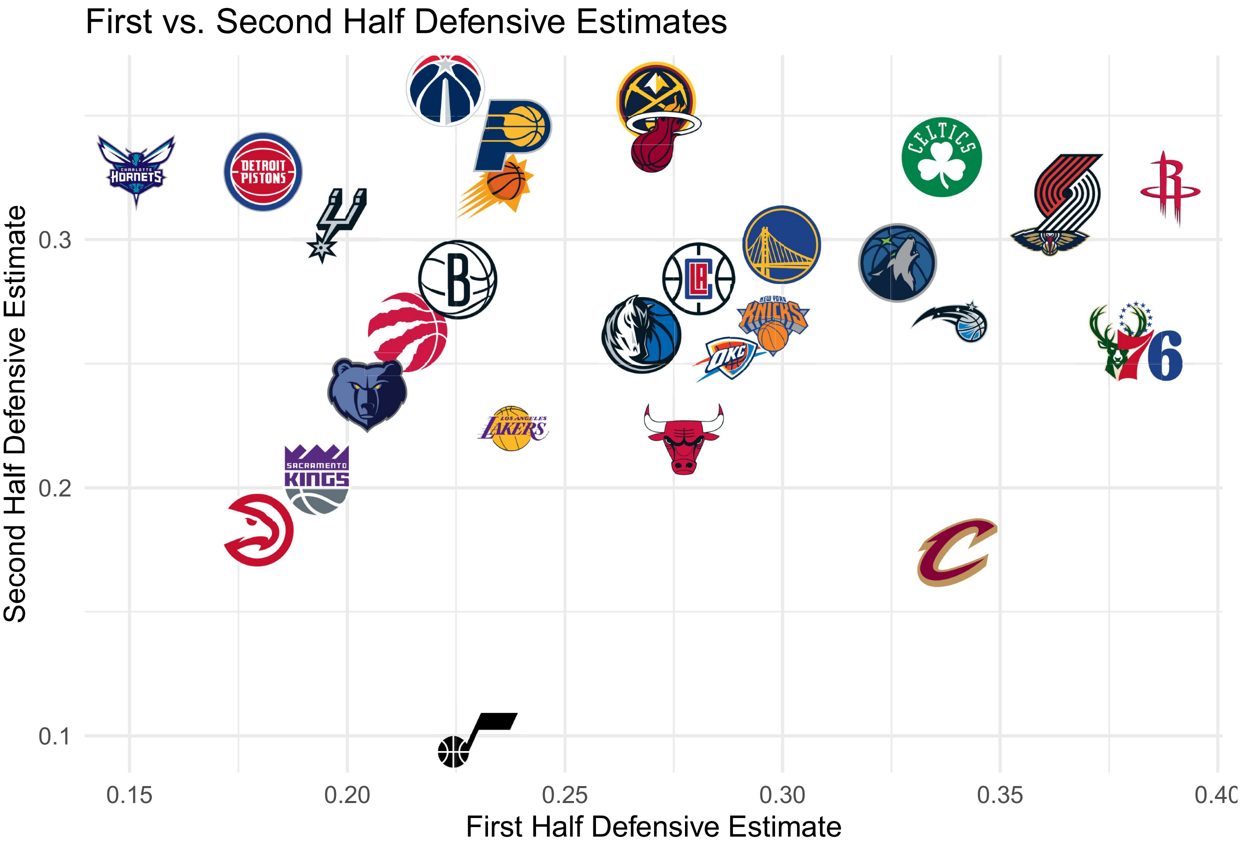
To quantify uncertainty in our parameter estimates and model predictions, we will employ bootstrap resampling. Bootstrapping involves repeatedly sampling from the observed data with replacement to create multiple bootstrap samples, from which we estimate the sampling distribution of a statistic of interest. We perform bootstrap resampling on our data to estimate confidence intervals for regression coefficients, providing a range of values for the true coefficients with a specified level of confidence. These intervals help us understand the uncertainty surrounding our estimates and interpret the model's coefficients.

Results

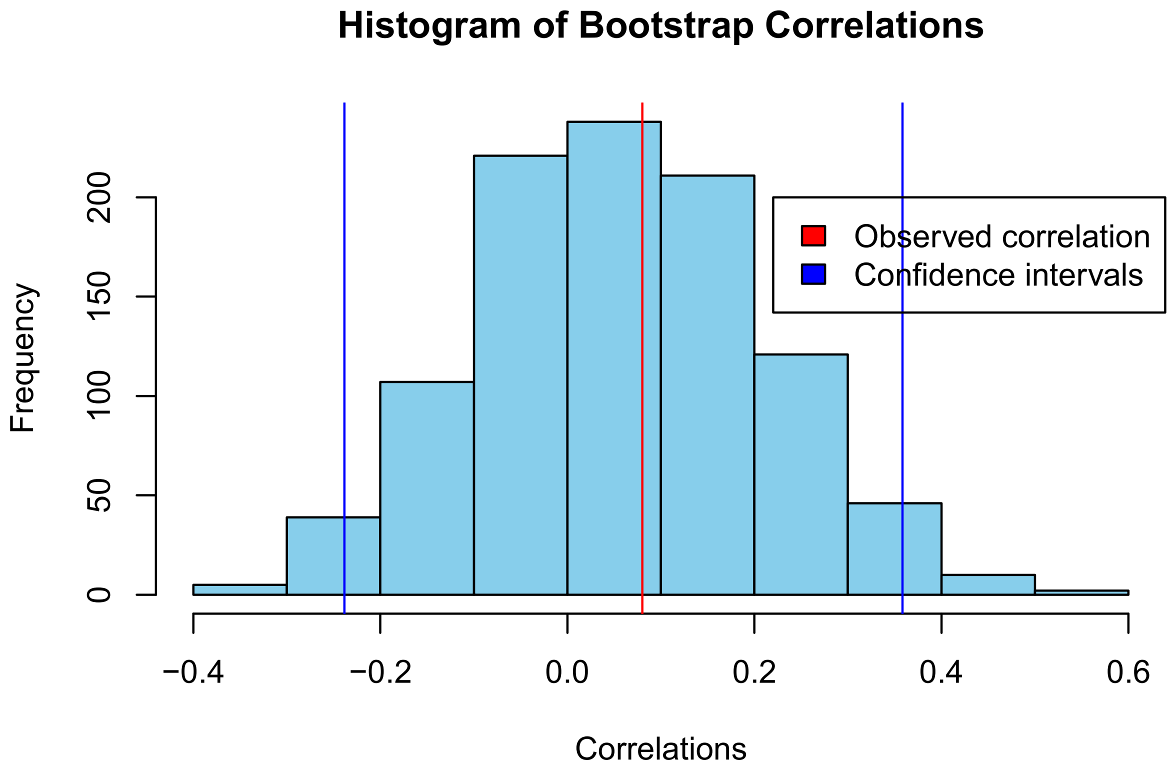

60 60 60 60 60 60 60 60 60 60 60 60 60 60 60
Mean−Squared Error
0.15
0.25
0.35
0.45
−6
−4
−2
0
2
Log(λ)


60 60 60 60 60 60 60 60 60 60 60 60 60 60 60
Mean−Squared Error
0.15
0.25
0.35
0.45
−6
−4
−2
0
2
Log(λ)

We see that the optimal lambda for both models is around -6. We use the coefficients calculated from training the model on the optimal lambdas.



We see no visible correlation between the NBA teams' first and second halves, with a calculated correlation coefficient of 0.080. This indicates that three-point defense is not a consistent metric; a team's three-point defense in the first half of the season tells us very little about their three-point defense in the second half of the season.



Our bootstrapping produces a 95% confidence interval of -0.238 to 0.358. This indicates that our observed correlation is within a reasonable range, and that the true correlation between three-point defense in the first and second halves of the season is likely indeed close to 0, meaning they're uncorrelated.

Discussion

Conclusions, Accuracy, and Limitations

The primary objective of our study was to analyze the consistency of NBA teams in defending three-point shots across the first and second halves of the 2023-24 season. Using a ridge regression model, we observed a very low correlation (0.080) in three-point defensive performance between the first and second halves of the season. This suggests that a team's ability to defend against three-point shots is not consistent throughout the season, indicating that teams do not necessarily maintain similar effectiveness in defending three-pointers throughout the season.

Our analysis highlights that three-point defense may not be a persistent quality within teams across the season. Teams that performed well in defending three-point shots in the first half do not necessarily continue doing so in the second half. This finding challenges the assumption that three-point defense is a stable trait and suggests that other factors, possibly including changes in team dynamics, injuries, or adjustments by opponents, may significantly influence defensive effectiveness.

The claims regarding the consistency of three-point defense are not supported by the low correlation observed. Ridge regression, chosen for its ability to avoid multicollinearity and overfitting, and the use of cross-validation in the regression analysis, were expected to enhance the reliability of the model. However, the low correlation indicates that the model's predictions do not demonstrate consistency in three-point defensive performance across different parts of the season.

Moreover, the bootstrap analysis, intended to support our conclusions by providing a distribution of correlation coefficients through resampling, actually underscores the variability of the defense metric across different samples rather than its stability. This finding suggests that our initial expectations for defensive consistency were not met.

The claims regarding the consistency of three-point defense are well-supported by the analytical methods used. Ridge regression was chosen for its ability to avoid multicollinearity and overfitting. The use of cross-validation in the regression analysis enhances the reliability of the model, ensuring that the model's predictions are not fitted to a specific subset of data but hold generalizability across different sets of game data.

The analysis presented in this report has several limitations, including the use of data from only one NBA season, which may not fully capture long-term trends in three-point defense. Variability in team strategies and rosters due to trades, injuries, and coaching changes could significantly influence defensive effectiveness, but these factors were not accounted for. The ridge regression model, while robust against multicollinearity, assumes linearity and constant variance, which may not accurately reflect the dynamic nature of the data. Furthermore, the study does not consider other potentially influential factors such as home-court advantage or opponent strength, and relies solely on the three-point percentage allowed to measure defensive effectiveness, ignoring the quality of shots or defensive schemes used. Moreover, it is possible that a team concedes a lot of three-pointers but at a low accuracy, which would make their three-point defense look better than reality.

Future Work:

While this analysis provided insights into the variability of three-point defense from a team perspective, we believe that the logical next step would be to apply a similar process to player-level data. This would include metrics such as three-pointers contested and the percentage made when defended. Unfortunately, we could not access such data for this project. Analyzing player-level data would allow us to determine if three-point defending is more a factor of team strategy/quality or individual player skill. Furthermore, we could identify which players excel or struggle in three-point defense, which would have broad implications for basketball at all levels, from the NBA to high school.

From a recruitment standpoint, just as teams prioritize acquiring excellent three-point shooters, it could be argued that signing adept three-point defenders is equally important. Moreover, a better evaluation of three-point defending could provide players, agents, executives, and the media with a more comprehensive understanding of player contributions and qualities, potentially impacting awards, contracts, and discussions regarding defensive ability and impact. Lastly, it would be interesting to explore the effects of adding a skilled three-point defender to a team that struggles with three-point defense and vice versa. Such an analysis could add significant nuance to team selection and squad building across basketball leagues, perhaps highlighting that the impact a player has in defending three-pointers could be limited by the quality of the team they join.

Code Appendix

```
# To load the data:
tictoc::tic()
progressr::with_progress({
  nba_team_box <- hoopR::load_nba_team_box(2024:hoopR::most_recent_nba_season())
})
tictoc::toc()

library(lme4)

library(tidyverse)
```

```
# Getting rid of the All-Stars team
not_nba <- which(nba_team_box$team_name=="All-Stars")
nba_team_box <- nba_team_box[-not_nba,]

# Filter to exclude playoff games
nba_team_box <- nba_team_box %>%
  filter(game_date <= as.Date("2024-04-14"))


# Split the season into first and second halves
first_half <- subset(nba_team_box, game_date < median(nba_team_box$game_date))
second_half <- subset(nba_team_box, game_date >= median(nba_team_box$game_date))


# Histogram of three-point field goal percentage for the first half of the season
ggplot(first_half, aes(x = three_point_field_goal_pct)) +
  geom_histogram(binwidth = 3, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Three-Point Field Goal % - First Half",
       x = "Three-Point FG %", y = "Frequency")

# Histogram of three-point field goal percentage for the second half of the season
ggplot(second_half, aes(x = three_point_field_goal_pct)) +
  geom_histogram(binwidth = 3, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Three-Point Field Goal % - Second Half",
       x = "Three-Point FG %", y = "Frequency")

# Log odds transformation
first_half$log_odds <- log(first_half$three_point_field_goal_pct/
                             (100-first_half$three_point_field_goal_pct))

second_half$log_odds <- log(second_half$three_point_field_goal_pct/
                             (100-second_half$three_point_field_goal_pct))


# Extract log odds from the halved datasets
first_log_odds <- first_half$log_odds
second_log_odds <- second_half$log_odds

# Create data frames for the design matrices with all zeros initially
first_design_def <- data.frame(matrix(0,nrow=nrow(first_half),ncol=30))
first_design_off <- data.frame(matrix(0,nrow=nrow(first_half),ncol=30))
second_design_def <- data.frame(matrix(0,nrow=nrow(second_half),ncol=30))
second_design_off <- data.frame(matrix(0,nrow=nrow(second_half),ncol=30))

# Assign appropriate values to the design matrices based on team names and opponent names
team_names <- unique(nba_team_box$team_name)

colnames(first_design_def) <- paste0("def_",team_names)
colnames(first_design_off) <- paste0("off_",team_names)
colnames(second_design_def) <- paste0("def_",team_names)
colnames(second_design_off) <- paste0("off_",team_names)

for (i in 1:length(team_names)) {
  team <- team_names[i]
  def_name <- paste0("def_",team)
  off_name <- paste0("off_",team)
  first_design_def[,def_name] <- ifelse(first_half$opponent_team_name==team,-1,0)
  first_design_off[,off_name] <- ifelse(first_half$team_name==team,1,0)
  second_design_def[,def_name] <- ifelse(second_half$opponent_team_name==team,-1,0)
  second_design_off[,off_name] <- ifelse(second_half$team_name==team,1,0)
}

# Combine the design matrices with the halved datasets
first_design <- cbind(first_log_odds,first_design_def,first_design_off)
second_design <- cbind(second_log_odds,second_design_def,second_design_off)


# Train the ridge regression model
library(glmnet)
first_teams <- as.matrix(first_design[-1])
second_teams <- as.matrix(second_design[-1])
first_ridge_cv <- cv.glmnet(first_teams,first_design$first_log_odds,alpha=0,intercept=FALSE,
                            standardize=FALSE)
second_ridge_cv <- cv.glmnet(second_teams,second_design$second_log_odds,alpha=0,intercept=FALSE,
                             standardize=FALSE)


# Plot lambdas and MSEs
plot(first_ridge_cv)
plot(second_ridge_cv)


# Plot defensive coefficients between halves of season
library(broom)
tidy_first_ridge_coef <- tidy(first_ridge_cv$glmnet.fit)
tidy_second_ridge_coef <- tidy(second_ridge_cv$glmnet.fit)
filt_first_ridge_coef <- tidy_first_ridge_coef[which(tidy_first_ridge_coef$lambda==
                                                       first_ridge_cv$lambda.min),]
filt_second_ridge_coef <- tidy_second_ridge_coef[which(tidy_second_ridge_coef$lambda==
                                                         second_ridge_cv$lambda.min),]

nba_team_box <- nba_team_box[, c("team_name", "team_logo")]
nba_team_box <- unique(nba_team_box)

filt_first_ridge_coef <- filt_first_ridge_coef %>%
  mutate(team_name = substring(term,5))

filt_second_ridge_coef <- filt_second_ridge_coef %>%
  mutate(team_name = substring(term,5))

filt_first_ridge_coef <- filt_first_ridge_coef %>%
  left_join(nba_team_box, by = "team_name")

filt_second_ridge_coef <- filt_second_ridge_coef %>%
  left_join(nba_team_box, by = "team_name")

data_for_plot <- data.frame(
  Estimate1 = filt_first_ridge_coef$estimate[1:30],
  Estimate2 = filt_second_ridge_coef$estimate[1:30],
  Logo = filt_second_ridge_coef$team_logo
)
```

```
library(ggimage)

ggplot(data_for_plot, aes(x = Estimate1, y = Estimate2)) +
  geom_image(aes(image = Logo), size = 0.12) +
  labs(x = "First Half Defensive Estimate",
       y = "Second Half Defensive Estimate",
       title = "First vs. Second Half Defensive Estimates") +
  theme_minimal()

# Calculate correlation between defensive coefficients between halves of season
real_cor <- cor(filt_first_ridge_coef$estimate[1:30],filt_second_ridge_coef$estimate[1:30])

# Bootstrapping the data and rerunning the model
first_boot_defs <- matrix(nrow=1000,ncol=30)
second_boot_defs <- matrix(nrow=1000,ncol=30)
boot_cors <- rep(0,1000)
first_boot <- first_design
second_boot <- second_design
for (ii in 1:1000) {
  set.seed(ii)
  first_indices <- sample(1:1226,1226,replace=TRUE)
  second_indices <- sample(1:1236,1236,replace=TRUE)
  for (jj in 1:1226) {
    first_boot[jj,] <- first_design[first_indices[jj],]
  }
  for (jj in 1:1236) {
    second_boot[jj,] <- second_design[second_indices[jj],]
  }
  first_boot_teams <- as.matrix(first_boot[-1])
  second_boot_teams <- as.matrix(second_boot[-1])
  first_ridge_boot <- cv.glmnet(first_boot_teams,first_boot$first_log_odds,alpha=0,
                                intercept=FALSE,standardize=FALSE)
  second_ridge_boot <- cv.glmnet(second_boot_teams,second_boot$second_log_odds,alpha=0,
                                 intercept=FALSE,standardize=FALSE)
  tidy_first_ridge_boot <- tidy(first_ridge_boot$glmnet.fit)
  tidy_second_ridge_boot <- tidy(second_ridge_boot$glmnet.fit)
  filt_first_ridge_boot <- tidy_first_ridge_boot[which(tidy_first_ridge_boot$lambda==
                                                         first_ridge_boot$lambda.min),]
  filt_second_ridge_boot <- tidy_second_ridge_boot[which(tidy_second_ridge_boot$lambda==
                                                           second_ridge_boot$lambda.min),]
  boot_cors[ii] <- cor(filt_first_ridge_boot$estimate[1:30],
                       filt_second_ridge_boot$estimate[1:30])
  first_boot_defs[ii,] <- filt_first_ridge_boot$estimate[1:30]
  second_boot_defs[ii,] <- filt_second_ridge_boot$estimate[1:30]
}
for (ii in 1:1000) {
  boot_cors[ii] <- cor(first_boot_defs[ii,],second_boot_defs[ii,])
}

# Getting confidence intervals
bottom_ci <- as.numeric(quantile(boot_cors,0.025))
top_ci <- as.numeric(quantile(boot_cors,0.975))
```

```
# Plotting the bootstrapped correlation distribution
hist(boot_cors,xlab="Correlations",main="Histogram of Bootstrap Correlations",col="skyblue")
abline(v=real_cor,col="red")
abline(v=bottom_ci,col="blue")
abline(v=top_ci,col="blue")
legend(0.22,200,legend=c("Observed correlation","Confidence intervals"),fill=c("red","blue"))
```