Basketball efficiency metrics

Michael Lopez, Skidmore College

Overview

In this lab, we'll gain experience implementing traditional and more novel approaches to measuring shooter accuracy. We'll also implement newer approach for looking at the repeatability of a statistic. First, our preamble to get shot level data from the 2014-2015 NBA season.

```
library(RCurl)
library(tidyverse)
url <- getURL("https://raw.githubusercontent.com/JunWorks/NBAstat/master/shot.csv")
nba_shot <- read.csv(text = url)
nba_shot <- na.omit(nba_shot)%>%
  filter(SHOT_DIST>=21 |PTS_TYPE==2)
```

Let's get player specific rate statistics. These include field goal percentage (`FGP`) and effective field goal percentage (`eFGP`). (Note: I don't have free throw info, so we can't look at true shooting percentage).

Additionally, `n.shots` stores the number of shots for each player. We'll filter so that we are only looking at players with at least 100 shots.

```
player_shoot <- nba_shot %>%
  group_by(playerName) %>%
  summarize(n.shots = length(FGM),
            FGP = sum(FGM)/(n.shots),
            eFGP = sum(PTS)/(2*n.shots))
```

Let's look at some of our top shooters and our worst shooters, as judged by `eFGP`.

```
player_shoot %>%
  arrange(eFGP) %>%
  head()

player_shoot %>%
  arrange(-eFGP) %>%
  head()
```

Before we go much further, it's pretty clear there's an issue with players who took a small number of shots. We can also visualize this using a scatter plot.

```
ggplot(data = player_shoot, aes(n.shots, eFGP)) +
  geom_point()
```

1. Why does the scatter plot fan in? Is that what we expected?

Let's restrict our sample to only shooters with at least 500 shots.

```
player_shoot <- player_shoot %>%
  filter(n.shots >= 500)

ggplot(data = player_shoot, aes(n.shots, FGP)) +
  geom_point()

ggplot(data = player_shoot, aes(n.shots, eFGP)) +
  geom_point()
```

2. Does there appear to be a link between FGP and number of shots or eFGP and number of shots?

Next, we look at the link between eFGP and FGP.

```
ggplot(data = player_shoot, aes(FGP, eFGP)) +
  geom_point()
```

3. Why is there a line in the scatter plot above at the line y = x? What players are reflected in those points?

Repeatability of shooting percentage metrics.

With baseball and football metrics, we frequently looked at how well a certain statistic correlated with itself (or another statistic) in the future. This is often useful for decision makers in sports who are looking to predict a player's future given his or her past.

In the NBA data set, however, we *don't* have multiple years of data. Instead, we'll split our larger data frame of shots into two smaller data frames, each with half of the shots, and compare the associations between shooting metrics between the two smaller subsets. In essense, this looks at the repeatability of a metric within the same time frame, in place of looking forward.

First, we split the larger data set into two smaller ones, using a random number generator and the filter command. In this example, nba_shot1 stores the first subset, while nba_shot2 stores the second.

```
set.seed(1)
nba_shot$test <- rnorm(nrow(nba_shot))>0
nba_shot1 <- filter(nba_shot, test)
nba_shot2 <- filter(nba_shot, !test)
nrow(nba_shot1)
nrow(nba_shot2)
```

It's fine if there are a few more shots in one of the data sets – that's just due to chance.

Next, we repeat the same procedure as above within each the two smaller data frames.

```
player_shoot1 <- nba_shot1 %>%
  group_by(playerName) %>%
  summarize(n.shots1 = length(FGM),
            FGP1 = sum(FGM)/(n.shots1),
            eFGP1 = sum(PTS)/(2*n.shots1))

player_shoot2 <- nba_shot2 %>%
  group_by(playerName) %>%
  summarize(n.shots2 = length(FGM),
            FGP2 = sum(FGM)/(n.shots2),
            eFGP2 = sum(PTS)/(2*n.shots2))

head(player_shoot1)
head(player_shoot2)
```

4. How many of A.J. Price's shots ended up in the first subset of shots? And how many in the second?

Finally, we join the two player-level shot subsets.

```
player_shoot_all <- inner_join(player_shoot1, player_shoot2)
player_shoot_all <- filter(player_shoot_all, n.shots1 >=100, n.shots2 >=100)
```

5. Identify the correlations between (FGP1 and FGP2) and (eFGP1 and eFGP2). Which is more repeatable: a player's shooting percentage or his effective field goal percentage?

6. Provide a logical explanation for why eFGP may be less repeatable than field goal percentage.
7. Why might effective field goal percentage still be a preferred metric, even if it is less repeatable?

Pretty graphs

```
fit.1 <- glm(SHOT_RESULT == "made" ~ SHOT_DIST + TOUCH_TIME +
             DRIBBLES + SHOT_CLOCK + CLOSE_DEF_DIST,
              data = nba_shot, family = "binomial")

nba_shot <- nba_shot %>%
  mutate(predicted.probs = fitted(fit.1),
  expected.pts = predicted.probs * PTS_TYPE,
         epa = PTS - expected.pts)

shot_group <- nba_shot %>%
  group_by(playerName) %>%
  summarise(total.epa = sum(epa)) %>%
  arrange(total.epa)

all_shooters <- inner_join(shot_group, player_shoot)

ggplot(all_shooters, aes(FGP, total.epa, label = playerName)) +
  geom_text() + scale_y_continuous("Expected Points Added") +
  scale_x_continuous("eFGP") +
  ggtitle("Expected points added ~ FGP, 2014-15 season") +
  theme_bw()

ggplot(all_shooters, aes(eFGP, total.epa, label = playerName)) +
  geom_text() + scale_y_continuous("Expected Points Added") +
  scale_x_continuous("eFGP") +
  ggtitle("Expected points added ~ effective FGP, 2014-15 season") +
  theme_bw()
```

8. Is the association between expected points added (`EPA`) stronger when compared to field goal percentage or effective field goal percentage. Why does this follow our intuition?