Lab 7: The James-Stein estimator

Michael Lopez, Skidmore College

Important note 1

First, we'll open RStudio by going to http://r.skidmore.edu/.

Important note 2

Open a new R Markdown file (File / New File / R Markdown...). You can create a basic name – Lab0, for example – and that'll set you up with a new file ready to go.

Overview

Today's focus is on the implementation of the James-Stein estimator to estimate shooting percentages in hockey.

Recall:

Stein's Paradox: Circumstances in which there are estimators better than the arithmetic average - `better` defined by accuracy - `better` estimators use combination of individual ones

However, note that this process is applicable to variables across sports.

1. Identify metrics in two other sports that are similar to shooting percentage in hockey – e.g., ones where implementation of the J-S estimator would be appropriate.

Summary of shooting percentages

First, we load in the data, which contains information on player-level hockey metrics since the 2012-2013 season. Our focus today is on defensemen, and identifying better representations of defenseman shooting percentages.

```
library(RCurl)
library(tidyverse)
url <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/NHL.csv")
nhl.data <- read_csv(url)
nhl.data <- nhl.data %>% filter(TOI > 500)
nhl.data <- na.omit(nhl.data)
nhl.data$ShP <- nhl.data$Goals/nhl.data$Shots

first.season <- nhl.data %>% filter(Season==20122013)
first.defenders <- first.season %>%
  group_by(Name) %>%
  filter(Shots > 20, Position == "D") %>%
  select(Name, Position, Goals, Shots, ShP)
dim(first.defenders)
```

2. Use a histogram to plot the shooting percentages of the defensemen. Describe the center, shape, and spread of this distribution. How does these numbers compare to the forwards we looked at last class?

Calculating the James-Stein estimator

3. Each of the below calculates one element of the J-S estimator. What is being calculated in this step?

```
p.bar <- mean(first.players$ShP)
p.bar
p.hat <- first.players$ShP
p.hat
```

4. Each of the below calculates one element of the J-S estimator. What is being calculated in each step?

```
N <- first.players$Shots
N
sigma.sq <- sd(p.hat)^2
sigma.sq
```

5. The following code calculates c, which is often referred to as a shrinkage or skill factor. The first value of c is 0.11. Interpret this number.

```
c <- (N/0.25)/(N/0.25 + 1/sigma.sq)
c
```

6. What values of c are typical? How does the c for defensemen compare to the value we found for forwards?

7. The following code implements the J-S estimator, as well as the MLE estimator (`ShP.MLE`) and the constant c for each player. How does each players c compare to their number of shots?

```
first.players$c <- c
first.players$ShP.MLE <- first.players$ShP
first.players$ShP.JS <- p.bar + c*(p.hat - p.bar)

set.seed(0)
sample.players <- first.players %>% ungroup() %>% sample_n(14)
sample.players
```

Estimation accuracy

A final step is to identify how our estimators have done at estimating a players' shooting percentage. To do that, we'll link to each players shooting percentage over the remainder of their career.

```
all.players <- nhl.data %>%
  group_by(Name) %>%
  filter(Name %in% first.players$Name, Season >= 20122013) %>%
  summarise(ShP.Career = sum(Goals)/sum(Shots), n.shots = sum(Shots))
sample.players1 <- inner_join(sample.players, all.players) %>%
  select(Name, ShP, ShP.MLE, ShP.JS, ShP.Career)
sample.players1
```

8. Find a player where the J-S estimator was more accurate as far as predicting career shooting rate. Find a player where the MLE was more accurate in predicting future shooting rate.

9. The following code estimates the MAE for each estimator. Which estimator was more accurate? By how much? Interpret each of the MAE's below.

```
sample.players1 %>%
  ungroup() %>%
  mutate(abs.error.mle = abs(ShP.MLE - ShP.Career),
```

```
        abs.error.js = abs(ShP.JS - ShP.Career)) %>%
  summarise(mae.mle = mean(abs.error.mle),
            mae.js = mean(abs.error.js))
```

Finally, we visualize the predictions.

```
sample.players1 %>%
  gather(shot.type, shot.rate, ShP.MLE:ShP.Career) %>%
  ggplot(aes(x = rev(shot.type), y = shot.rate, group = Name, colour = Name)) +
  geom_line() +
  geom_point() +
  scale_colour_discrete(guide = FALSE) +
  geom_text(data = sample.players1, aes(x = 0.2, y  = jitter(ShP.MLE), label = Name), hjust = 0) +
  xlab("Estimator") + ylab("Shooting pct")
```

10. Explain where predictions hit and missed using the chart. Don't be afraid to print out the list of players in `sample.players1` to better investigate.