writeup

Weixi Chen

2021/9/30

Introduction

In this project, we still focus on the World Series. And we are going to use simulation and analytic methods to compare the probability of winning the World Series with and without home field advantage. The World Series is a first-to-4-wins match-up between the champions of the American and National Leagues of Major League Baseball.

Suppose that the Braves and the Yankees are teams competing in the World Series, and the home field advantage is the edge which a team may have when playing a game at its home stadium. For example, it is the edge the Braves may have over the Yankees when the head-to-head match-up is in Atlanta. It is the advantage the Yankees may have when the head-to-head match-up is in New York.

The table below has the two possible schedules for each game of the series. (NYC = New York City, ATL = Atlanta):

Overall advantage	Game 1	Game 2	Game 3	Game 4	Game 5	Game 6	Game 7
Braves	ATL	ATL	NYC	NYC	NYC	ATL	ATL
Yankees	NYC	NYC	ATL	ATL	ATL	NYC	NYC

Let P_B be the probability that the Braves wins a single head-to-head match-up with the Yankees, under the assumption that home field advantage doesn't exist. Let P_B^H denotes the probability that the Braves wins a single head-to-head match-up with the Yankees as the home team (H for home). Let P_B^A denotes the probability that the Braves wins a single head-to-head match-up with the away team (A for away).

Game location	No advantage	Advantage
ATL	P_B	$P_B^H = P_B * 1.1$
NYC	P_B	$P_B^A = 1 - (1 - P_B) * 1.1$

Question1

Compute analytically the probability that the Braves win the world series when the sequence of game locations is {NYC, NYC, ATL, ATL, ATL, NYC, NYC}. (The code below computes the probability for the alternative sequence of game locations. Note: The code uses data.table syntax, which may be new to you. This is intentional, as a gentle way to introduce data.table.) Calculate the probability with and without home field advantage when $P_{B}=$0.55. What is the difference in probabilities?

With home field advantage

```
require(dplyr)
```

```
##      dplyr

##
##    'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
require(data.table)
```

```
##      data.table

##
##    'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
# Get all possible outcomes
apo <- fread('all-possible-world-series-outcomes.csv')

# Home field indicator
hfi <- c(0, 0, 1, 1, 1, 0, 0) #{NYC, NYC, ATL, ATL, ATL, NYC, NYC}

# P_B
pb <- 0.55
advantage_multiplier <- 1.1 # Set = 1 for no advantage
pbh <- 0.55*advantage_multiplier
pba <- 1 - (1 - 0.55)*advantage_multiplier

# Calculate the probability of each possible outcome
apo[, p := NA_real_] # Initialize new column in apo to store prob
for(i in 1:nrow(apo)){
  prob_game <- rep(1, 7)
  for(j in 1:7){
    p_win <- ifelse(hfi[j], pbh, pba)
    prob_game[j] <- case_when(
        apo[i, j, with=FALSE] == "W" ~ p_win
      , apo[i, j, with=FALSE] == "L" ~ 1 - p_win
      , TRUE ~ 1
    )
  }
  apo[i, p := prod(prob_game)] # Data.table syntax
}

# Check if sum(p) == 1
apo[, sum(p)] # This is data.table notation
```

```
## [1] 1
```

```
# Probability of overall World Series outcomes
apo[, sum(p), overall_outcome]
```

```
##    overall_outcome       V1
## 1:               W 0.604221
## 2:               L 0.395779
```

The probability of Braves win the World Series with home field advantage is 0.604221.

Without home field advantage

```
pnbinom(3, 4, 0.55)
```

```
## [1] 0.6082878
```

The probability of Braves win the World Series without home field advantage is 0.6082878.

```
0.6082878-0.604221
```

```
## [1] 0.0040668
```

The probabilities in these two conditions are different, and the probability of Braves win the World Series without home field advantage is 0.0040668 larger than the probability of Braves win the World Series with home field advantage.

Question2

Calculate the same probabilities as the previous question by simulation.

With home field advantage

```r
set.seed(1)
# Home field indicator
hfi <- c(0, 0, 1, 1, 1, 0, 0) #{NYC, NYC, ATL, ATL, ATL, NYC, NYC}

# P_B
pb <- 0.55
advantage_multiplier <- 1.1 # Set = 1 for no advantage
pbh <- 0.55*advantage_multiplier
pba <- 1 - (1 - 0.55)*advantage_multiplier

# Use for loop to simulate the sorld series with home field advantage.
win = 0
for (i in 1:1e6) {
  win_num = 0
  result = 0
  for (j in 1:7) {
    result = rbinom(1, 1, ifelse(hfi[j], pbh, pba))
    win_num = win_num + result
    if(win_num == 4){
      break
    }
  }
  win = win + ifelse(win_num == 4, 1, 0)
}

win/1e6
```

```
## [1] 0.604548
```

The simulated probability of Braves win the World Series with home field advantage is 0.604548.

Without home field advantage

```r
set.seed(2)
# Home field indicator
hfi <- c(0, 0, 1, 1, 1, 0, 0) #{NYC, NYC, ATL, ATL, ATL, NYC, NYC}

# P_B
pb <- 0.55

# Use for loop to simulate the sorld series with home field advantage.
win = 0
for (i in 1:1e7) {
  win_num = 0
  result = 0
  for (j in 1:7) {
    result = rbinom(1, 1, pb)
    win_num = win_num + result
    if(win_num == 4){
      break
    }
  }
  win = win + ifelse(win_num == 4, 1, 0)
}

win/1e7
```

```
## [1] 0.6081044
```

The simulated probability of Braves win the World Series without home field advantage is 0.6081044.

Question3

What is the absolute and relative error for your simulation in the previous question?

Absolute error

Absolute error is equal to $|\hat{P} - P|$

```
# With home field advantage
abs(0.604548-0.604221)
```

```
## [1] 0.000327
```

```
# Without home field advantage
abs(0.6081044-0.6082878)
```

```
## [1] 0.0001834
```

The absolute error is 0.000327, 0.0001834, respectively.

Relative error

Relative error is equal to $|\hat{P} - P|$/P

```
# With home field advantage
abs(0.604548-0.604221)/0.604221
```

```
## [1] 0.0005411927
```

```
# Without home field advantage
abs(0.6081044-0.6082878)/0.6082878
```

```
## [1] 0.000301502
```

The relative error is 0.0005411927, 0.000301502, respectively.

Question4

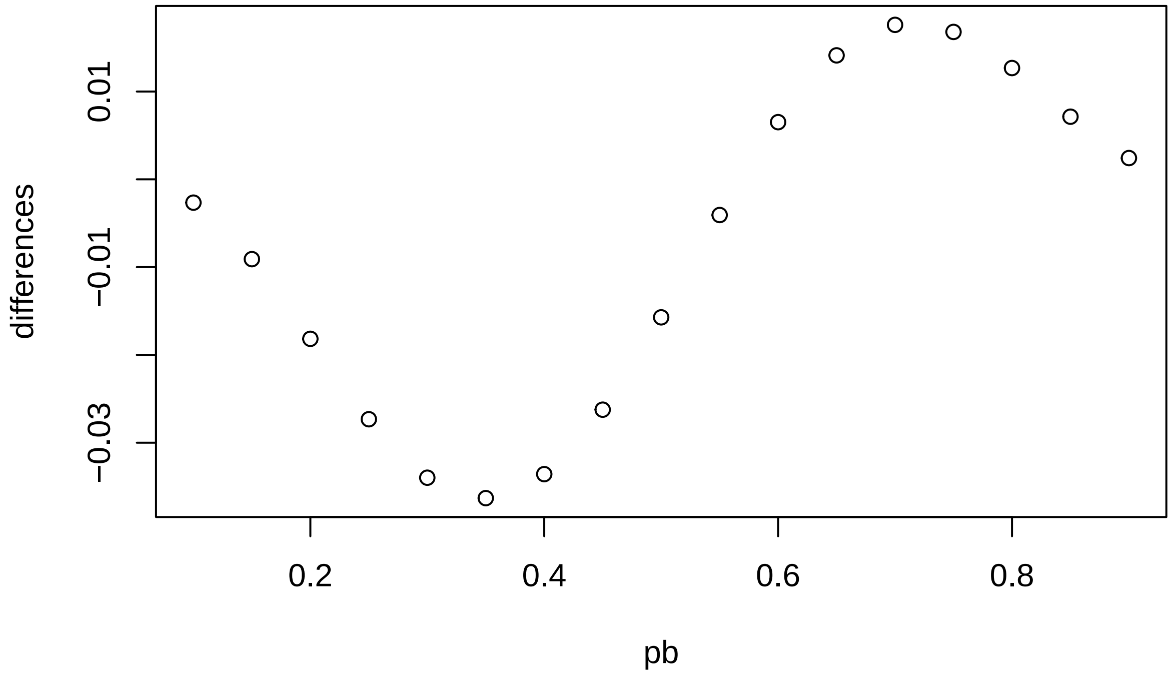
Does the difference in probabilities (with vs without home field advantage) depend on P_B? (Generate a plot to answer this question.)

```r
# Home field indicator
hfi <- c(0, 0, 1, 1, 1, 0, 0) #{NYC, NYC, ATL, ATL, ATL, NYC, NYC}

# P_B
pb <- seq(0.1, 0.9, 0.05) # To make sure the values of pbh and pba are between 0 and 1
advantage_multiplier <- 1.1 # Set = 1 for no advantage
differences <- c()

# Calculate the probability of each possible outcome and each probability
for (h in pb) {
   pbh <- h*advantage_multiplier
   pba <- 1 - (1 - h)*advantage_multiplier
   apo[, p := NA_real_] # Initialize new column in apo to store prob
   for(i in 1:nrow(apo)){
        prob_game <- rep(1, 7)
        for(j in 1:7){
           p_win <- ifelse(hfi[j], pbh, pba)
           prob_game[j] <- case_when(
              apo[i, j, with=FALSE] == "W" ~ p_win
            , apo[i, j, with=FALSE] == "L" ~ 1 - p_win
            , TRUE ~ 1
           )
         }
        apo[i, p := prod(prob_game)] # Data.table syntax
      }
    p_with_win = apo[, sum(p), overall_outcome][1,2]
    p_without_win = pnbinom(3, 4, h)
    differences <- c(differences, p_with_win - p_without_win)
}

plot(x = pb, y = differences)
```



The difference between probabilities of the Braves win World Series with and without home field advantage depends on P_B. The differences goes larger as P_B being larger, and then the difference goes smaller at around $P_B = 0.35$, then the difference goes larger again at around $P_B = 0.55$, in the last, the difference goes smaller again at until the end.

Question5

Does the difference in probabilities (with vs without home field advantage) depend on the advantage factor? (The advantage factor in P_B^H and P_B^A is the 1.1 multiplier that results in a 10% increase for the home team. Generate a plot to answer this question.)

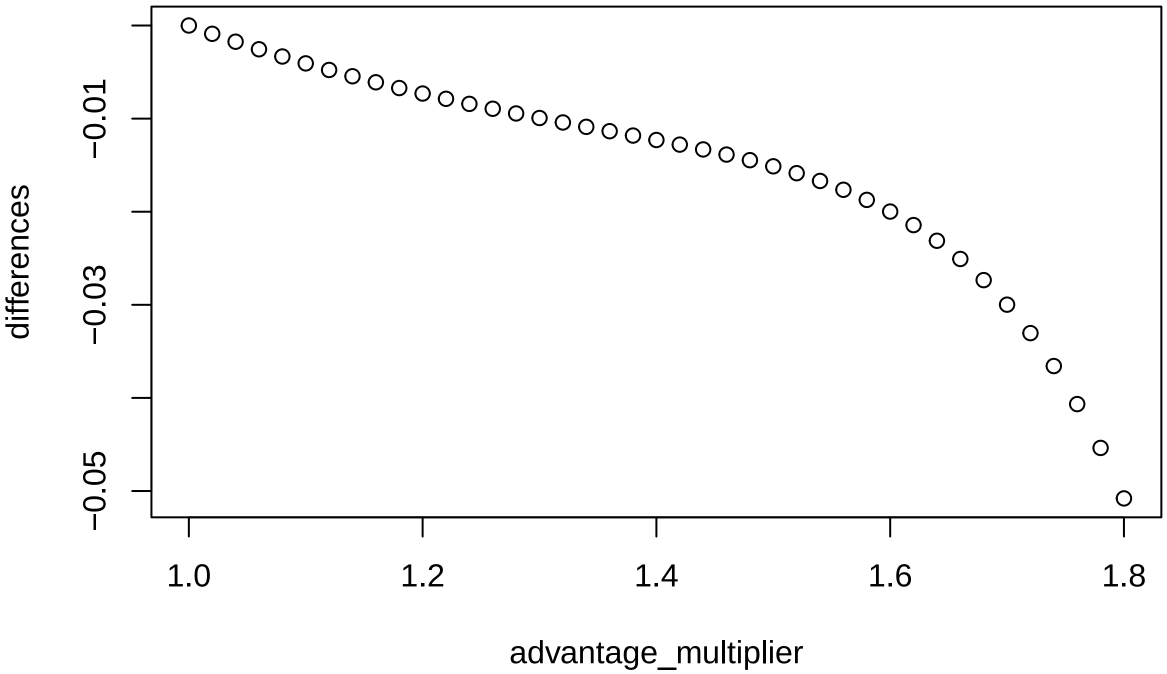
We will fix P_B at 0.55 and explore the relationship between difference and the advantage factor.

```
# Home field indicator
hfi <- c(0, 0, 1, 1, 1, 0, 0) #{NYC, NYC, ATL, ATL, ATL, NYC, NYC}

# P_B
pb <- 0.55
advantage_multiplier <- seq(1, 1.8, 0.02)
# Set = 1 for no advantage, and set this range to make sure that the values of pbh and pba are between
differences <- c()

# Calculate the probability of each possible outcome and each advantage factor
for (h in advantage_multiplier) {
   pbh <- 0.55*h
   pba <- 1 - (1 - 0.55)*h
   apo[, p := NA_real_] # Initialize new column in apo to store prob
   for(i in 1:nrow(apo)){
        prob_game <- rep(1, 7)
        for(j in 1:7){
           p_win <- ifelse(hfi[j], pbh, pba)
           prob_game[j] <- case_when(
              apo[i, j, with=FALSE] == "W" ~ p_win
            , apo[i, j, with=FALSE] == "L" ~ 1 - p_win
            , TRUE ~ 1
           )
         }
        apo[i, p := prod(prob_game)] # Data.table syntax
      }
    p_with_win = apo[, sum(p), overall_outcome][1,2]
    p_without_win = pnbinom(3, 4, 0.55)
    differences <- c(differences, p_with_win - p_without_win)
}

plot(x = advantage_multiplier, y = differences)
```



The difference between probabilities of the Braves win World Series with and without home field advantage depends on the advantage factor. The differences goes larger as the advantage factor being larger.