# DS-5620 HW4

Jingyu Ruan

## Contents

## Q1

```
library(data.table)
library(dplyr)
library(ggplot2)

apo <- fread("C:/Users/Ruan Jingyu/OneDrive - PennO365/Desktop/Academic Courses/DS 5620-01 Probability S
setnames(apo, old = names(apo), new = sub("^\\s+|\\s+$", "", names(apo)))  # trim names

# clamp probabilities to [0,1] to be safe under multipliers.
clamp01 <- function(x) pmax(0, pmin(1, x))

# build a per-game Braves win-prob vector from a schedule, base P_B, and advantage multiplier m.
per_game_probs <- function(schedule, P_B, m){
  P_home <- clamp01(P_B * m)
  P_away <- clamp01(1 - (1 - P_B) * m)
  # map locations to Braves home (ATL) or away (NYC).
  if(!all(schedule %in% c("ATL","NYC"))) stop("I expect schedule values only in {'ATL','NYC'}.")
  ifelse(schedule == "ATL", P_home, P_away)
}

# compute the analytic probability that the Braves win the series by enumerating apo.
# assume apo has game outcomes as "W"/"L" per column.
series_prob_analytic <- function(schedule, P_B, m, apo_df = apo){
  p_win_vec <- per_game_probs(schedule, P_B, m)
  # ensure the first 7 columns correspond to games 1..7.
  game_cols <- names(apo_df)[1:7]
  # compute the path probability row by row.
  probs <- apply(apo_df[, ..game_cols], 1, function(row_outcome){
```

```r
    per_game <- ifelse(row_outcome == "W", p_win_vec, 1 - p_win_vec)
    prod(per_game)
  })
  # detect Braves series win by counting W >= 4.
  braves_win <- rowSums(apo_df[, ..game_cols] == "W") >= 4
  sum(probs[braves_win])
}

# simulate one series given schedule, P_B, and m.
sim_one_series <- function(schedule, P_B, m){
  p_win_vec <- per_game_probs(schedule, P_B, m)
  b <- 0; y <- 0
  for(g in seq_along(p_win_vec)){
    if(runif(1) < p_win_vec[g]) b <- b + 1 else y <- y + 1
    if(b == 4) return(TRUE)
    if(y == 4) return(FALSE)
  }
  b > y
}

# simulate many series to estimate the championship probability.
series_prob_sim <- function(schedule, P_B, m, n_sims = 200000L, seed = 42L){
  set.seed(seed)
  mean(replicate(n_sims, sim_one_series(schedule, P_B, m)))
}

schedule_yankees_adv <- c("NYC","NYC","ATL","ATL","ATL","NYC","NYC")

P_B <- 0.55
m_with <- 1.10
m_none <- 1.00

p_with  <- series_prob_analytic(schedule_yankees_adv, P_B, m_with)
p_none  <- series_prob_analytic(schedule_yankees_adv, P_B, m_none)
delta   <- p_with - p_none

tibble(
  schedule = paste(schedule_yankees_adv, collapse = "-"),
  P_B = P_B,
  m = c(m_with, m_none, NA_real_),
  case = c("with_adv", "no_adv", "difference"),
  prob = c(p_with, p_none, delta)
)
```

```
## # A tibble: 3 x 5
##   schedule                      P_B     m case          prob
##   <chr>                       <dbl> <dbl> <chr>        <dbl>
## 1 NYC-NYC-ATL-ATL-ATL-NYC-NYC  0.55   1.1 with_adv   0.293
## 2 NYC-NYC-ATL-ATL-ATL-NYC-NYC  0.55   1   no_adv     0.292
## 3 NYC-NYC-ATL-ATL-ATL-NYC-NYC  0.55  NA   difference 0.00125
```

Home field makes almost no difference. Advantage adds only 0.1%.

## Q2

```r
p_with_sim <- series_prob_sim(schedule_yankees_adv, P_B, m_with, n_sims = 200000, seed = 123)
p_none_sim <- series_prob_sim(schedule_yankees_adv, P_B, m_none, n_sims = 200000, seed = 123)
delta_sim  <- p_with_sim - p_none_sim

tibble(
  schedule = paste(schedule_yankees_adv, collapse = "-"),
  P_B = P_B,
  m = c(m_with, m_none, NA_real_),
  case = c("with_adv_sim", "no_adv_sim", "difference_sim"),
  prob = c(p_with_sim, p_none_sim, delta_sim)
)
```

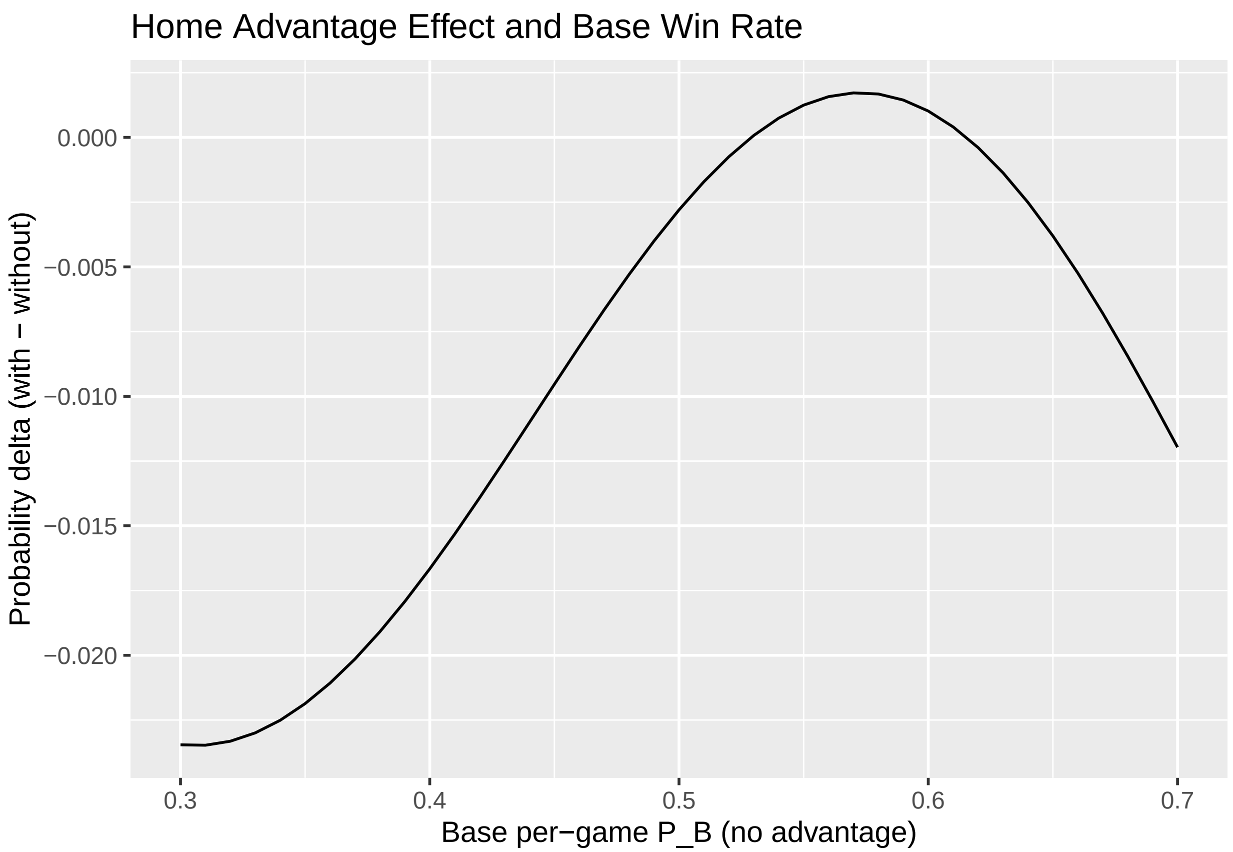
```
## # A tibble: 3 x 5
##   schedule                        P_B     m case               prob
##   <chr>                         <dbl> <dbl> <chr>             <dbl>
## 1 NYC-NYC-ATL-ATL-ATL-NYC-NYC  0.55   1.1 with_adv_sim    0.605
## 2 NYC-NYC-ATL-ATL-ATL-NYC-NYC  0.55   1   no_adv_sim      0.609
## 3 NYC-NYC-ATL-ATL-ATL-NYC-NYC  0.55  NA   difference_sim -0.00377
```

Simulation shows almost no difference, too.

## Q3

```r
grid_pb <- seq(0.30, 0.70, by = 0.01)
df_pb <- tibble(
  P_B = grid_pb,
  prob_with = sapply(grid_pb, function(p) series_prob_analytic(schedule_yankees_adv, p, 1.10)),
  prob_none = sapply(grid_pb, function(p) series_prob_analytic(schedule_yankees_adv, p, 1.00))
) |>
  mutate(delta = prob_with - prob_none)

ggplot(df_pb, aes(P_B, delta)) +
  geom_line() +
  labs(
    title = "Home Advantage Effect and Base Win Rate",
    x = "Base per-game P_B (no advantage)",
    y = "Probability delta (with - without)"
  )
```



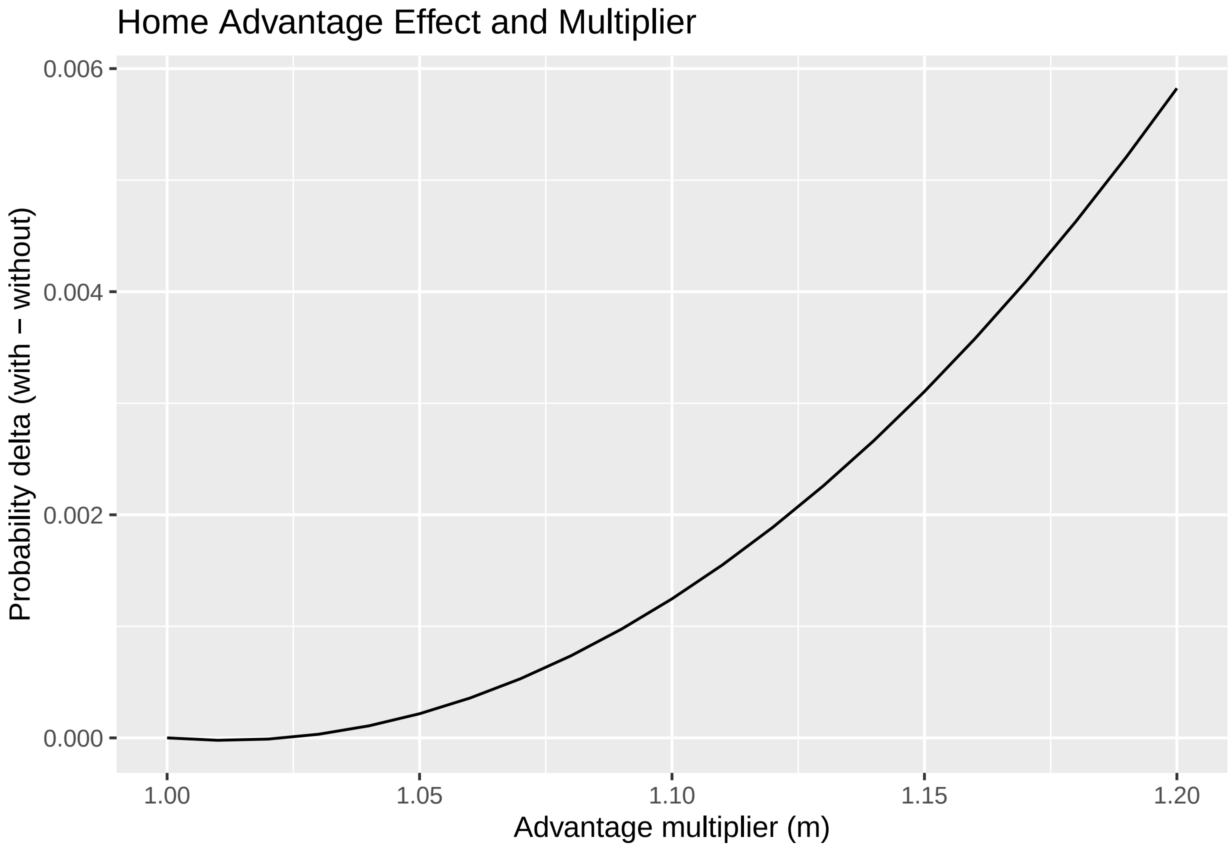
Effect depends on base strength. At most Braves lose about 2%. Advantage does not always help.

## Q4

```r
grid_m <- seq(1.00, 1.20, by = 0.01)

df_m <- tibble(
  m = grid_m,
  prob_with = sapply(grid_m, function(mm) series_prob_analytic(schedule_yankees_adv, 0.55, mm)),
  prob_none = series_prob_analytic(schedule_yankees_adv, 0.55, 1.00)
) |>
  mutate(delta = prob_with - prob_none)

ggplot(df_m, aes(m, delta)) +
  geom_line() +
  labs(
    title = "Home Advantage Effect and Multiplier",
    x = "Advantage multiplier (m)",
    y = "Probability delta (with - without)"
  )
```



Stronger advantage multiplier gives larger effect. Gain grows but still small relatively.