Lab 2

Due on 02/17/23 at 11:59 pm

```
library(ggplot2)
library(dplyr)
library(Lahman)
library(tidyverse)
library(retrosheet)
```

Question 1

- Construct a data frame which includes the following variables from the `Teams` data frame in the `Lahman` package: `yearID`, `teamID`, `AB`, `SO`, `H`, `HR`, `R`, `RA`, `W`, and `L`. Only keep seasons dating back to 1990, and remove the 1994, 1995, and 2020 seasons.

```
newTeams <- Teams %>%
  select(yearID, teamID, AB, SO, H, HR, R, RA, W, L) %>%
  filter(yearID >= "1990") %>%
  filter(yearID != "1994") %>%
  filter(yearID != "1995") %>%
  filter(yearID != "2020")

#852

#question 1b
bwar_bat = readr::read_csv("https://www.baseball-reference.com/data/war_daily_bat.txt", na = "NULL")
```

```
## Rows: 119945 Columns: 49
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): name_common, player_ID, team_ID, lg_ID, pitcher
## dbl (44): age, mlb_ID, year_ID, stint_ID, PA, G, Inn, runs_bat, runs_br, run...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
bwar_pit = readr::read_csv("https://www.baseball-reference.com/data/war_daily_pitch.txt", na = "NULL")
```

```
## Rows: 53884 Columns: 43
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): name_common, player_ID, team_ID, lg_ID
## dbl (39): age, mlb_ID, year_ID, stint_ID, G, GS, IPouts, IPouts_start, IPout...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
bwar_bat <- bwar_bat %>%
  filter(year_ID >= "1990") %>%
  filter(year_ID != "1994") %>%
  filter(year_ID != "1995") %>%
  filter(year_ID != "2020") %>%
  filter(year_ID != "2022")

bwar_pit <- bwar_pit %>%
  filter(year_ID >= "1990") %>%
  filter(year_ID != "1994") %>%
  filter(year_ID != "1995") %>%
  filter(year_ID != "2020") %>%
  filter(year_ID != "2022")

WARdef_pull <- bwar_bat %>%
  select(name_common, team_ID, year_ID, WAR_def)
#40291

#need to groupby(team_ID, year_ID)
BRruns_pull <- bwar_bat %>%
  select(name_common, team_ID, year_ID, runs_br)
#40291
bullpen_war_pull <- bwar_pit %>%
  select(team_ID, year_ID, IPouts,
         IPouts_relief, WAR) %>%
  filter((IPouts_relief/ IPouts) >= 0.75)
#12298

teamdWAR <- WARdef_pull %>%
  group_by(year_ID, team_ID) %>%
  na.omit(WAR_def) %>%
  summarise(dWAR = sum(WAR_def))
```

```
## `summarise()` has grouped output by 'year_ID'. You can override using the
## `.groups` argument.
```

```r
teamBRruns <- BRruns_pull %>%
  group_by(year_ID, team_ID)  %>%
  summarise(BRruns = sum(runs_br))
```

```
## `summarise()` has grouped output by 'year_ID'. You can override using the
## `.groups` argument.
```

```r
teamPenWAR <- bullpen_war_pull %>%
  group_by(year_ID, team_ID) %>%
  summarise(penWAR = sum(WAR))
```

```
## `summarise()` has grouped output by 'year_ID'. You can override using the
## `.groups` argument.
```

```r
teamPenWAR$penWAR <- round(teamPenWAR$penWAR ,digits = 2)
teamBRruns$BRruns <- round(teamBRruns$BRruns ,digits = 2)
teamdWAR$dWAR <- round(teamdWAR$dWAR ,digits = 2)
#need to figure out all the rounding stuff

#question 1c

newTeams$teamID <- str_replace(newTeams$teamID, "CHA", "CHW")
newTeams[newTeams == "CHN"] <- "CHC"
newTeams[newTeams == "KCA"] <- "KCR"
newTeams[newTeams == "LAN"] <- "LAD"
newTeams[newTeams == "ML4"] <- "MIL"
newTeams[newTeams == "NYA"] <- "NYY"
newTeams[newTeams == "NYN"] <- "NYM"
newTeams[newTeams == "SDN"] <- "SDP"
newTeams[newTeams == "SFN"] <- "SFG"
newTeams[newTeams == "TBA"] <- "TBR"
newTeams[newTeams == "WAS"] <- "WSN"
newTeams[newTeams == "FLO"] <- "FLA"
newTeams$teamID <- str_replace(newTeams$teamID, "SLA", "STL")

teams <- newTeams %>%
  group_by(yearID, teamID)

newTeams <- cbind(teams, teamdWAR$dWAR, teamBRruns$BRruns, teamPenWAR$penWAR)
```

```
## New names:
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
```

```r
colnames(newTeams)[colnames(newTeams) == "...11"] ="dWAR"
colnames(newTeams)[colnames(newTeams) == "...12"] ="BRruns"
colnames(newTeams)[colnames(newTeams) == "...13"] ="penWAR"


#question 1d
newTeams <- newTeams %>%
  mutate(RD = R - RA)

#Compute and add winning percentage \texttt{Wpct} to your data frame. Use an equation in your notes and

#question 1e

q1e <- newTeams %>%
  mutate(Wpct = W / (W + L))

dat_aug <- newTeams %>%
  mutate(logWratio = log(W / L),
         logRratio = log(R / RA))

pyFit <- lm(logWratio ~ 0 + logRratio, data = dat_aug)
pyFit
```

```
##
## Call:
## lm(formula = logWratio ~ 0 + logRratio, data = dat_aug)
##
## Coefficients:
## logRratio
##     1.858
```

```
#Display the rows of this data frame corresponding to the 2014-2015 Royals seasons.

royals <- dat_aug %>%
  filter(yearID == "2014" | yearID == "2015") %>%
  filter(teamID == "KCR")
```

Question 2 In this problem we will perform analyses that investigate strengths and peculiarities of the 2014-2015 Royals. Do the following:

- Fit and analyze a regression model of `residuals_pytk` on `penWAR`. Determine how many wins one would expect the Royals to obtain above their Pythagorean expectations on the basis of their bullpen.

```
dat_aug <- dat_aug %>%
  mutate(Wpct = W / (W + L)) %>%
  mutate(Wpct_pyt = R^2 / (R^2 + RA^2)) %>%
  mutate(residuals_pytk = Wpct - Wpct_pyt)

m3 <- lm(penWAR ~ 0 + Wpct_pyt, data = dat_aug)
m3
```

```
##
## Call:
## lm(formula = penWAR ~ 0 + Wpct_pyt, data = dat_aug)
##
## Coefficients:
## Wpct_pyt
##    8.008
```

```
0.02991 *162
```

```
## [1] 4.84542
```

```
#The Royals' bullpen WAR outpaced their Pythagorean wins by 4.85 wins, on average.
```

Question 3 Do the following:

- Select a period of your choice (at least 20 years) and fit the Pythagorean formula model (after finding the optimal exponent) to the run-differential, win-loss data.

```
q3a <- dat_aug %>%
  filter (yearID >= "2000") %>%
  mutate(logWLratio = log(W/L),
         logRDratio = log(R/RA))

fitted <- lm(logWLratio ~ 0 + logRDratio, data = q3a)
fitted
```

```
##
## Call:
## lm(formula = logWLratio ~ 0 + logRDratio, data = q3a)
##
## Coefficients:
## logRDratio
##      1.845
```

```
q3a <- q3a %>%
  mutate(W_pyt = (W ^ 1.845 / (W ^ 1.845 + L ^ 1.845)) *162 )  %>%
  mutate(RD_pyt = (R ^ 1.845 / (R ^ 1.845 + RA ^ 1.845)) *162 ) %>%
  mutate(RD_resid = RD - RD_pyt) %>%
  mutate(W_resid = W - W_pyt)
```

- On the basis of your fit in the previous part and the list of managers obtained from Retrosheet, compile a top 10 list of managers who most overperformed their Pythagorean winning percentage and a top 10 list of managers who most underperformed their Pythagorean winning percentage.

```
#underperforming managers
underperform <- q3a[order(q3a$W_resid), ]
head(underperform, 10)
```

```
## # A tibble: 10 x 25
## # Groups:   yearID, teamID [10]
##    yearID teamID    AB    SO     H    HR     R    RA     W     L  dWAR BRruns
##     <int> <chr>  <int> <int> <int> <int> <int> <int> <int> <int> <dbl>  <dbl>
##  1   2001 SEA     5680   989  1637   169   927   627   116    46 10.2   18.1
##  2   2018 BOS     5623  1253  1509   208   876   647   108    54 -0.03   4.93
##  3   2019 HOU     5613  1166  1538   288   920   640   107    55  9.5   -5.35
##  4   2021 SFG     5462  1461  1360   241   804   594   107    55 -1.39   1.57
##  5   2019 LAD     5493  1356  1414   279   886   613   106    56  3.29   5.45
##  6   2021 LAD     5445  1408  1330   237   830   561   106    56 -1.05  -2.92
##  7   2004 SLN     5555  1085  1544   214   855   659   105    57  2.41   4.05
##  8   2002 NYY     5601  1171  1540   223   897   697   103    58 -2.68   2.08
##  9   2016 CHC     5503  1339  1409   199   808   556   103    58 -0.39 -12.2
## 10   2002 ATL     5495  1028  1428   164   708   565   101    59  7.84  -5.3
## # ... with 13 more variables: penWAR <dbl>, RD <int>, logWratio <dbl>,
## #   logRratio <dbl>, Wpct <dbl>, Wpct_pyt <dbl>, residuals_pytk <dbl>,
## #   logWLratio <dbl>, logRDratio <dbl>, W_pyt <dbl>, RD_pyt <dbl>,
## #   RD_resid <dbl>, W_resid <dbl>
```

Managers: 2001 Seattle: pinielo01 & mclarjo99 2018 Boston: coraal01 2019 Houston: hinchaj01 2021 San Francisco: kaplega01 2019 Los Angeles Dodgers: roberda07 2021 Los Angeles Dodgers: roberda07 2004 St. Louis: larusto01 2002 New York Yankees: torrejo01 2016 Chicago Cubs: maddojo99 2002 Atlanta: coxbo01

```
#overperforming managers

overperform <- q3a[order(-q3a$W_resid), ]
head(overperform, 10)
```

```
## # A tibble: 10 x 25
```

```
## # Groups:   yearID, teamID [10]
##    yearID teamID    AB    SO     H    HR     R    RA     W     L  dWAR BRruns
##     <int> <chr>  <int> <int> <int> <int> <int> <int> <int> <int> <dbl>  <dbl>
##  1   2003 DET     5466  1099  1312   153   591   928    43   119 -2.84  -7.35
##  2   2018 BAL     5507  1412  1317   188   622   892    47   115 -2.82   2.24
##  3   2019 DET     5549  1595  1333   149   582   915    47   114 -9.44   1.58
##  4   2004 ARI     5544  1022  1401   135   615   899    51   111 -0.85  -1.86
##  5   2013 HOU     5457  1535  1307   148   610   848    51   111 -3.57  -7.63
##  6   2021 ARI     5489  1465  1297   144   679   893    52   110 -5.4    3.44
##  7   2021 BAL     5420  1454  1296   195   659   956    52   110 -3.05  -3.2
##  8   2019 BAL     5596  1435  1379   213   729   981    54   108 -2.08   1.51
##  9   2012 HOU     5407  1365  1276   146   583   794    55   107 -4.93  -4.73
## 10   2002 MIL     5415  1125  1369   139   627   821    56   106 -1.82 -18.9
## # ... with 13 more variables: penWAR <dbl>, RD <int>, logWratio <dbl>,
## #   logRratio <dbl>, Wpct <dbl>, Wpct_pyt <dbl>, residuals_pytk <dbl>,
## #   logWLratio <dbl>, logRDratio <dbl>, W_pyt <dbl>, RD_pyt <dbl>,
## #   RD_resid <dbl>, W_resid <dbl>
```

Managers: 2003 Detroit: trammal01 2018 Baltimore: showabu99 2019 Detroit: gardero01 2004 Arizona: brenlbo01 & pedrial01 2013 Houston: portebo03 2021 Arizona: lovulto01 2021 Baltimore: hydebr99 2019 Baltimore: hydebr99 2012 Houston: millsbr01 & defrato99 2002 Milwaukee: lopesda01 & roystje01 **Question 4** The first question on page 21 in Section 1.4.3 of Analyzing Baseball Data with R.

- During the McGwire/Sosa home run race, which player was more successful at hitting homers with men on base?

Mark McGwire hit 37 home runs in 313 plate appearances with runners on base, while Sammy Sosa hit 29 in 367. Once walks (both intentional and unintentional) and hit by pitches are removed, the number of opportunities become 223 for McGwire and 317 for Sosa.

```
#fields <- Batting %>%
#fields <- read.csv("fields.csv")

#I am very stressed and confused
#I think we'll end up needing year-by-year data, similar to the dataset we worked on in class
```