Introduction, course setup, and data sets

Daniel J. Eck

Welcome

Welcome to STAT 430: Baseball Analytics!

Let’s have a fun semester exploring the ability of statistics to quantify winning games and evaluate players.

Background

This lecture is meant to supplement Chapter 1 in your textbook and to go over course resources, course logistics, and the course schedule.

GitHub

Course materials will be distributed from my GitHub organization which was built using CS's GitHub-repo-creator.

See setup.md for details.

Software

- The R Project for Statistical Computing: https://www.r-project.org/
- RStudio as an integrated development environment for R: https://www.rstudio.com/

Data sets used in class

- `Lahman` package:

```
install.packages("Lahman")
```

- Retrosheet data. Appendix A in your textbook provides an R script file for downloading and parsing all the game log files. A possible more convenient approach for obtaining retrosheets is included in this slide deck. Or you can occasionally use the `retrosheet` package for simple retrosheets:

```
install.packages("retrosheet")
```

- Statcast data obtained from the `baseballr` package:

```
install.packages("baseballr")
```

- baseball_R. Coding scripts and data sets that supplement your textbook.

- Data scraped from baseball reference:

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

- Era-adjusted data sets: Will be provided later.

More on Lahman

We will also use CSV files that are obtained online:

1. Access: https://www.seanlahman.com/baseball-archive/statistics/
2. Below the *Download Latest Version* section, there is a selection for dowloading the most recent version of the database. Click on the *2021 – comma-delimited version [Baseball Databank]* version
3. Save the file to a directory of your choice.
4. Extract the contents of the downloaded zip file

Lahman Data in R

The `Lahman` package contains several tables consisting of useful stat. We highlight a few tables below

```
library(Lahman)
data(Batting)
data(Pitching)
data(Fielding)
data(Teams)
```

Lahman Batting table

```
head(Batting)
```

```
##    playerID yearID stint teamID lgID  G  AB  R  H X2B X3B HR RBI SB CS BB SO
## 1 abercda01   1871     1    TRO   NA  1   4  0  0   0   0  0   0  0  0  0  0
## 2  addybo01   1871     1    RC1   NA 25 118 30 32   6   0  0  13  8  1  4  0
## 3 allisar01   1871     1    CL1   NA 29 137 28 40   4   5  0  19  3  1  2  5
## 4 allisdo01   1871     1    WS3   NA 27 133 28 44  10   2  2  27  1  1  0  2
## 5 ansonca01   1871     1    RC1   NA 25 120 29 39  11   3  0  16  6  2  2  1
## 6 armstbo01   1871     1    FW1   NA 12  49  9 11   2   1  0   5  0  1  0  1
##   IBB HBP SH SF GIDP
## 1  NA  NA NA NA    0
## 2  NA  NA NA NA    0
## 3  NA  NA NA NA    1
## 4  NA  NA NA NA    0
## 5  NA  NA NA NA    0
## 6  NA  NA NA NA    0
```

Lahman Pitching table

```r
head(Pitching)
```

```
##    playerID yearID stint teamID lgID  W  L  G GS CG SHO SV IPouts   H  ER HR BB
## 1 bechtge01   1871     1    PH1   NA  1  2  3  3  2   0  0     78  43  23  0 11
## 2 brainas01   1871     1    WS3   NA 12 15 30 30 30   0  0    792 361 132  4 37
## 3 fergubo01   1871     1    NY2   NA  0  0  1  0  0   0  0      3   8   3  0  0
## 4 fishech01   1871     1    RC1   NA  4 16 24 24 22   1  0    639 295 103  3 31
## 5 fleetfr01   1871     1    NY2   NA  0  1  1  1  1   0  0     27  20  10  0  3
## 6 flowedi01   1871     1    TRO   NA  0  0  1  0  0   0  0      3   1   0  0  0
##   SO BAOpp   ERA IBB WP HBP BK  BFP GF   R SH SF GIDP
## 1  1    NA  7.96  NA  7  NA  0  146  0  42 NA NA   NA
## 2 13    NA  4.50  NA  7  NA  0 1291  0 292 NA NA   NA
## 3  0    NA 27.00  NA  2  NA  0   14  0   9 NA NA   NA
## 4 15    NA  4.35  NA 20  NA  0 1080  1 257 NA NA   NA
## 5  0    NA 10.00  NA  0  NA  0   57  0  21 NA NA   NA
## 6  0    NA  0.00  NA  0  NA  0    3  1   0 NA NA   NA
```

Lahman Fielding table

```
head(Fielding)
```

```
##    playerID yearID stint teamID lgID POS  G GS InnOuts PO  A  E DP PB WP SB CS
## 1 abercda01   1871     1    TRO   NA  SS  1  1      24  1  3  2  0 NA NA NA NA
## 2  addybo01   1871     1    RC1   NA  2B 22 22     606 67 72 42  5 NA NA NA NA
## 3  addybo01   1871     1    RC1   NA  SS  3  3      96  8 14  7  0 NA NA NA NA
## 4 allisar01   1871     1    CL1   NA  2B  2  0      18  1  4  0  0 NA NA NA NA
## 5 allisar01   1871     1    CL1   NA  OF 29 29     729 51  3  7  1 NA NA NA NA
## 6 allisdo01   1871     1    WS3   NA   C 27 27     681 68 15 20  4 18 NA  0  0
##   ZR
## 1 NA
## 2 NA
## 3 NA
## 4 NA
## 5 NA
## 6 NA
```

Lahman Teams table

```
head(Teams, 3)
```

```
##   yearID lgID teamID franchID divID Rank  G Ghome  W  L DivWin WCWin LgWin
## 1   1871   NA    BS1      BNA  <NA>    3 31    NA 20 10   <NA>  <NA>     N
## 2   1871   NA    CH1      CNA  <NA>    2 28    NA 19  9   <NA>  <NA>     N
## 3   1871   NA    CL1      CFC  <NA>    8 29    NA 10 19   <NA>  <NA>     N
##   WSWin   R   AB   H X2B X3B HR BB SO SB CS HBP SF  RA  ER  ERA CG SHO SV
## 1  <NA> 401 1372 426  70  37  3 60 19 73 16  NA NA 303 109 3.55 22   1  3
## 2  <NA> 302 1196 323  52  21 10 60 22 69 21  NA NA 241  77 2.76 25   0  1
## 3  <NA> 249 1186 328  35  40  7 26 25 18  8  NA NA 341 116 4.11 23   0  0
##   IPouts  HA HRA BBA SOA   E DP    FP                    name
## 1    828 367   2  42  23 243 24 0.834    Boston Red Stockings
## 2    753 308   6  28  22 229 16 0.829 Chicago White Stockings
## 3    762 346  13  53  34 234 15 0.818  Cleveland Forest Citys
##                           park attendance BPF PPF teamIDBR teamIDlahman45
## 1          South End Grounds I         NA 103  98      BOS            BS1
## 2      Union Base-Ball Grounds         NA 104 102      CHI            CH1
## 3 National Association Grounds         NA  96 100      CLE            CL1
##   teamIDretro
## 1         BS1
## 2         CH1
## 3         CL1
```

Retrosheets

There is a lot of box score information contained in a retrosheet.

Basic retrosheets can be obtained from the `retrosheet` package (the following code chunk has `eval = FALSE` because the retrosheet will not fit on a single slide).

```
library(retrosheet)
getRetrosheet(type = "game", year = 2012)
```

More comprehensive retrosheets can be obtained from the `baseballr` package (the following code chunk has `eval = FALSE` because the retrosheet will take awhile to load and will be stored locally).

```
library(baseballr)
retrosheet_data(path_to_directory = "~/Desktop/baseball_course/retrosheet",
                years_to_acquire = 1998)
```

Obtaining retrosheets via `baseballr` requires some work outlined by Bill Petti here.

The steps in the hyperlink above require one to first download and install files from the Chadwick Bureau.

Follow the instructions in the INSTALL file in the downloaded Chadwick tarball (this course used `chadwick-0.9.5`).

Statcast

A Statcast scraper script is in the **stat430resources** repo.

This scraper requires the user to first load in `tidyverse` and `baseballr`.

We will use this data later in the course.