Harper vs Trout

Dr. Ekstrom

Purpose: In this activity, we will compare the offensive performance of Bryce Harper with Mike Trout.

Sports Analytics: Introduce some ideas of examining offensive perfomance in baseball.

Data Science: Finding and downloading data from an appropriate website.

R: Introduction to *R* - including loading data, data frames and variables, and manipulating variables.

Finding, Downloading, and Uploading Data

There are many sites that have baseball data. Here we will use Baseball Reference. Go to the site. Type *Bryce Harper* in the search box, and press the search button.

Once you get to the Bryce Harper page, scroll down to the **Standard Batting** section. Go to **Share & more** - a menu should pop up and select **Get table as CSV (for Excel)** from that menu. The table turns into *csv* which you can copy. I copy the table, paste it into a blank TextEdit file (I have a Mac), and then save it as a *csv* file. In this case, I named the file *BryceHarper.csv*. Be sure you put this new file in a place you can find it.

To load the data, we use the *read.csv* command.

```
# Reads the CSV file and stores it in harper
harper=read.csv("BryceHarper.csv")
```

The command above will load the *BryceHarper.csv* file and store it in `harper` - which becomes a dataframe. A dataframe in *R* is similar to an array, but it is far more versatile.

Note the windows in the right margin. In the top right window, select **Environment** tab, under **Data** you should see `harper` - click on this to see the dataframe `harper`.

Manipulating the Data

Each column in the `harper` is easily accessible. For example, to access the `AB` (at-bats) column

```
harper$AB
```

```
##  [1]  533  424  352  521  506  420  550  573 3879  580   NA 3306  573
```

Note that Harper's career total in `AB` is the 9th entry (in the table, the row is labeled 8 yrs). To access this particular entry:

```
harper$AB[9]
```

```
## [1] 3879
```

How many extra base hits - doubles (`X2B`), triples (`X2B`), and homeruns (`HR`) - has Harper hit in his career?

```
# we create a new variable xbh - which is the sum of the variables for doubles,
# triples and homeruns
xbh=harper$X2B+harper$X3B+harper$HR
# the career total for xbh
xbh[9]
```

```
## [1] 457
```

So Harper has hit 457 extra base hits in his career (as of December 31, 2019).

How many singles has Harper hit in his career?

```
harperX1B=harper$H[9]-xbh[9]
harperX1B
```

```
## [1] 614
```

Harper has hit 614 singles as of May 31, 2019.

If we want to see (or access) all of Harper's career totals, we can use

```
harper[9,]
```

```
##    Year   Age    Tm    Lg    G   PA   AB   R    H X2B X3B  HR RBI SB CS
## 9 8 Yrs 8 Yrs 8 Yrs 8 Yrs 1084 4639 3879 708 1071 219  19 219 635 90 34
##    BB   SO    BA   OBP   SLG   OPS OPS.   TB GDP HBP SH SF IBB Pos Awards
## 9 684 1012 0.276 0.385 0.512 0.897  137 1985  76  29  9 38  81
```

A Bit of Analysis

What kind of numbers would Harper have if he took every at-bat for his team during the season? We will assume that Harper's team will get exactly 9 innings worth of at-bats every game.

How many outs does a team get in a season? There are 27 outs in a game (again, assusming the team gets exactly 9 innings worth of at-bats every game), and 162 games in a season, so there are 4374 total outs in a season.

How many outs has Harper made thus far in his career? We subtract the number of hits (`H`) from his total at-bats (`AB`), and add sacrifice bunts (`SH`), sacrifice flies (`SF`), times caught stealing (`CS`), and number of double plays he grounded into (`GDP`).

```
# Create a variable to hold the number of outs Harper has made thus far
harperouts=harper$AB[9]-harper$H[9]+harper$SH[9]+harper$SF[9]+harper$CS[9]+
  harper$GDP[9]
harperouts
```

```
## [1] 2965
```

We divide the number of outs Harper has made thus far by the number of outs a team gets in a season:

```
# Create a variable to hold the season factor for Harper
harperfactor=harperouts/4374
harperfactor
```

```
## [1] 0.6778692
```

So Harper has made 0.6778692 seasons worth of outs. So if Harper took every at-bat for his team, how many homeruns would Harper hit in a season?

```
harper$HR[9]/harperfactor
```

```
## [1] 323.0712
```

How many homeruns would Harper hit per game?

```
harper$HR[9]/(162*harperfactor)
```

```
## [1] 1.994266
```

The next question is how all these hits turn into runs. There are many ways to do this - one way is to create a model using linear weighting (something we will consider later in the semester). One such model is below:

$$\begin{pmatrix}\text{Runs}\\ \text{per}\\ \text{Game}\end{pmatrix} = -3.07429 + 0.55615 * \begin{pmatrix}\text{Singles}\\ \text{per}\\ \text{Game}\end{pmatrix} + 0.77286 * \begin{pmatrix}\text{Doubles}\\ \text{per}\\ \text{Game}\end{pmatrix}$$
$$+1.20914 * \begin{pmatrix}\text{Triples}\\ \text{per}\\ \text{Game}\end{pmatrix} + 1.49853 * \begin{pmatrix}\text{Homeruns}\\ \text{per}\\ \text{Game}\end{pmatrix} + 0.33919 * \begin{pmatrix}\text{Walks}\\ \text{per}\\ \text{Game}\end{pmatrix}$$
$$+0.13084 * \begin{pmatrix}\text{Stolen Bases}\\ \text{per}\\ \text{Game}\end{pmatrix} + 0.67992 * \begin{pmatrix}\text{Caught Stealing}\\ \text{per}\\ \text{Game}\end{pmatrix}$$

```
Harperruns=-3.07429+0.55615*(harperX1B/(162*harperfactor))+
  0.77286*(harper$X2B[9]/(162*harperfactor))+
  1.20914*(harper$X3B[9]/(162*harperfactor))+
  1.49853*(harper$HR[9]/(162*harperfactor))+
  0.33919*(harper$BB[9]/(162*harperfactor))+
  0.13084*(harper$SB[9]/(162*harperfactor))+
  0.67992*(harper$CS[9]/(162*harperfactor))
Harperruns
```

```
## [1] 7.204678
```

```
# To get the total number of runs in a season
162*Harperruns
```

```
## [1] 1167.158
```

How would these runs tranlate to wins? Again, there are several ways to do this. The Pythagorean Theorem of Baseball (see Pythagorean Theorem of Baseball) says

$$\text{Winning Percentage} = \frac{(\text{Runs Scored})^{1.81}}{(\text{Runs Scored})^{1.81} + (\text{Runs Allowed})^{1.81}}$$

In 2018, teams allowed 4.45 runs per game (see Major League Baseball Batting Year-by-Year Averages) or a total of 720.9 runs. Using the Pythagorean Theorem for baseball,

```
HarperWPercent=(1169.615)^(1.81)/((1169.615)^(1.81)+(720.9)^(1.81))
HarperWPercent
```

```
## [1] 0.7059744
```

we would expect Harper's team to win around 70.6% of their games. This equates to

```
162*HarperWPercent
```

```
## [1] 114.3679
```

around 114 wins.

Class Work

1. Go through a similar analysis for Mike Trout.
   - Find and download career data for Mike Trout.
   - Upload data into *R*.
   - Use the linear weighting and Pythagorean Theorem for baseball above to answer the following questions.
     - Determine how many runs a team would score if Mike Trout took every at-bat (again, assume the team gets exactly 27 outs each game). This is the total number of runs for the season, not the runs per game.
     - Determine how many games a team would win if Mike Trout took every at-bat (again, assume the team allows 720.9 runs).
2. Find the data for the 2019 Major League Baseball season. Use the Pythagorean Theorem for baseball to predict the number of wins for each MLB team. Compare your numbers with the actual number of wins for each MLB team. Who overperformed the most? Who underperformed the most?

For additional help, check out:

- Video: Read a CSV into R
- Video: Linear Weights on R
- Video: R Notebooks Introduction Part 1: Background
- Video: R Notebooks Introduction Part 2: Creating and Using R Notebooks

Further Reading:

- Article: A New Formula to Predict a Team's Winning Percentage