Assignment 4: Buster Posey

Jiang Li

9/17/2017

Contents

1. Introduction

You are a big baseball fan, and you enjoy looking at statistics of players and predicting which ones will do well. You have recently learned of a single metric, Weighted Runs Created, wRC+, that attempts to capture a player's total offensive value (how much they contribute to making runs). A complete explanation of wRC+ is beyond the scope of this class, but in summary, it combines every outcome (single, double, etc.), then adjusts the value to account for certain factors, such as the baseball parks where the player made the hits.

To learn more, go to the following sources:

http://www.fangraphs.com/library/offense/wrc/ (Links to an external site.)Links to an external site.

http://www.beyondtheboxscore.com/2014/5/26/5743956/sabermetrics-stats-offense-learn-sabermetrics (Links to an external site.)Links to an external site.

You are curious to see how standard baseball statistics, such as home runs and runs batted in, correlate to the more complex wRC+ score, so you gather some data. In this case, we study San Francisco Giants catcher Buster Posey. (For you baseball fans out there, I admit this is a dubious use of wRC+, but I still think it is an interesting statistical exercise)

See the associated dataset for the case, "DataScience_7_Case_Posey.xls". The screenshot below shows a portion of the data. It shows Buster Posey's batting performance from 2009 (the year he started with the Giants) to 2013.

2. Question 1

Using the data in the case, create a vector called "RBI" composed of the runs batted in by Buster Posey between 2009 and 2013 (i.e., 0, 67, 21, 103, 72). Find the mean, median, and range of the vector. Present the answers in an Adobe PDF or Microsoft Word document. Apply effective R coding practices, including comments embedded in the code. Include screenshots of your work in R.

```
## create RBI variable by manually entering the values
## In Question 2, we will show how to read it from csv or excel file
RBI = c(0,67,21,103,72)

## Calculate the mean
RBI.mean = mean(RBI)
cat("Mean is :",RBI.mean,"\n")
```

```
## Mean is : 52.6
```

```
## Calculate the median
RBI.median = median(RBI)
cat("Median is :",RBI.median,"\n")
```

```
## Median is : 67
```

```
## Calculate the range
RBI.range = range(RBI)
cat("Range is:",RBI.range,"\n")
```

```
## Range is: 0 103
```

3. Question 2

Read the entire dataset into R as a CSV file. Include the statement to read in the file, as well as a printout of the results to ensure the data was read in correctly. Present the answers in an Adobe PDF or Microsoft Word document. Apply effective R coding practices, including comments embedded in the code. Include screenshots of your work in R.

```
## Instead of converting the excel file into csv and later read it through read.csv
## We will use read_excel from the readxl package.
## Load the package (If not exists, install it)
if(!require("readxl")){
  install.packages("readxl")
}
```

```
## Loading required package: readxl
```

```
## Skip the first 22 rows int the original excel file
## If DataScience_7_Case_Posey.xls doesn't exist in your current directory, download it first
infile = "DataScience_7_Case_Posey.xls"
if(!file.exists(infile)){
  url.v = "http://www.stephansorger.com/content/DataScience_7_Case_Posey.xls"
  download.file(url = url.v,destfile = infile)
}
df = read_excel(path = infile,sheet = "Sheet1",skip = 22)
## print out the data
df
```

```
##   Year wRC+   G   AVG  AB  R   H 2B 3B HR RBI SB SO
## 1 2009  -51   7 0.118  17  1   2  0  0  0   0  0  4
## 2 2010  134 108 0.305 406 58 124 23  2 18  67  0 55
## 3 2011  116  45 0.284 162 17  46  5  0  4  21  3 30
## 4 2012  163 148 0.336 530 78 178 39  1 24 103  1 96
## 5 2013  133 148 0.294 520 61 153 34  1 15  72  1 70
```

4. Question 3

Use regression analysis to study the relationship between wRC+ and the common batting statistics Runs (R), Hits (H), and Runs Batted In (RBI). Designate wRC+ as the dependent variable. You will need to study only a subset of the entire dataset (just the variables discussed in this question). Find the y-intercept and coefficients for the three possible explanatory variables. Add your own assessment. Present the answers in an Adobe PDF or Microsoft Word document. Apply effective R coding practices, including comments embedded in the code. Include screenshots of your work in R.

```
## Fit a linear model of WRC+ on R, H and RBI
## use `` on the colname of wRC+ as it contains special character "+"
lm.wrc = lm(formula = `wRC+`~R+H+RBI,data = df)

## print out the result
summary(lm.wrc)
```

```
##
## Call:
## lm(formula = `wRC+` ~ R + H + RBI, data = df)
##
## Residuals:
##      1      2      3      4      5
## -52.56  19.00  66.95  -8.86 -24.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.31380   77.93368  -0.004    0.997
## R           -1.38072   18.39135  -0.075    0.952
## H            1.62502    5.43062   0.299    0.815
## RBI         -0.09111   10.76176  -0.008    0.995
##
## Residual standard error: 91.02 on 1 degrees of freedom
## Multiple R-squared:  0.7169, Adjusted R-squared:  -0.1324
## F-statistic: 0.8441 on 3 and 1 DF,  p-value: 0.644
```

```
## Now print out y-intercept and coefficients for the three possible explanatory variables.

cat("Y-intercept is: ",lm.wrc$coefficients[1],"\n")
```

```
## Y-intercept is:  -0.3137964
```

```
cat("Coefficient of 'R' is: ",lm.wrc$coefficients[2],"\n")
```

```
## Coefficient of 'R' is:  -1.380725
```

```r
cat("Coefficient of 'H' is: ",lm.wrc$coefficients[3],"\n")
```

```
## Coefficient of 'H' is:  1.625022
```

```r
cat("Coefficient of 'RBI' is: ",lm.wrc$coefficients[4],"\n")
```

```
## Coefficient of 'RBI' is:  -0.09110842
```

4.1. My assessment

There is a good fitness of wRC+ on R, H and RBI with R-squared of 0.7169, however the p-value is not signficant within given dataset (pvalue = 0.644)

5. Commentary

Commentary: What would be a better way to capture batter performance in a single metric?

Include research: What methodology or metric is used for Fantasy Baseball activities? How does it compare to the method outlined in the case study?

5.1 My commentary

In the Fantasy Baseball activities, wOBA (Weighted On-Base Average) is usally used as a metric to capture batter performance. It combines all the different aspects of hitting into one metric, weighting each of them in proportion to their actual run value. While batting average, on-base percentage, and slugging percentage fall short in accuracy and scope, wOBA measures and captures offensive value more accurately and comprehensively.

wOBA vs wRC+

- wRC+ puts everybody on a more even playing field. It actually strips away the park effects. For example, you can't compare Rockies and Giants by wOBA as easily as wRC+.
- wRC+ is graded on a scale where 100 is average