Assignment 4

Joseph Walker

2017-07-24

Contents

Introduction 1
Import the Data 1
RBI Summary Statistics 2
Regression Analysis 3
Research & Commentary 3

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/

http://www.beyondtheboxscore.com/2014/5/26/5743956/sabermetrics-stats-offense-learn-sabermetrics

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)

Import the Data

See the associated dataset for the case, "DataScience_7_Case_Posey.xls". 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.

In the code below, I imported the dataset from the web using the `read.xls()` function from the `gdata` package. Essentially, the `read.xls` function uses perl programming to convert the xls file to a csv file and reads it into the R environment. Additionally, the `read.xls()` contains all the arguments that `read.csv()` does. Downloading the file directly from the source is efficient and reproducible.

```
#Load required packages
library(tidyverse)
library(gdata)

#get URL with the dataset
url <- "http://www.stephansorger.com/content/DataScience_7_Case_Posey.xls"
```

```r
#Read in the dataset using the URL provided
posey_data <-  read.xls(url, perl = "C:/Perl/bin/perl.exe", skip = 20)

#convert column names to lower case
colnames(posey_data) <- tolower(colnames(posey_data))

#clean up the column names with rename()
posey_data <- rename(posey_data, wrc = wrc., double = x2b, triple = x3b)

#review the data set
str(posey_data)
```

```
## 'data.frame':    5 obs. of  13 variables:
##  $ year  : int  2009 2010 2011 2012 2013
##  $ wrc   : int  -51 134 116 163 133
##  $ g     : int  7 108 45 148 148
##  $ avg   : num  0.118 0.305 0.284 0.336 0.294
##  $ ab    : int  17 406 162 530 520
##  $ r     : int  1 58 17 78 61
##  $ h     : int  2 124 46 178 153
##  $ double: int  0 23 5 39 34
##  $ triple: int  0 2 0 1 1
##  $ hr    : int  0 18 4 24 15
##  $ rbi   : int  0 67 21 103 72
##  $ sb    : int  0 0 3 1 1
##  $ so    : int  4 55 30 96 70
```

RBI Summary Statistics

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.

```r
#create vector rbi
rbi <- posey_data$rbi

#create a list of the stats of the rbi vector
rbi_stats <- list(mean = mean(rbi), median = median(rbi), range = range(rbi))

#print out the stats
rbi_stats
```

```
## $mean
## [1] 52.6
##
## $median
## [1] 67
##
## $range
## [1]   0 103
```

Regression Analysis

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.

```
#create linear model using lm() function
wrc_model <- lm(data = posey_data, formula = wrc ~ r + h + rbi)

#view the regression model
summary(wrc_model)
```

```
##
## Call:
## lm(formula = wrc ~ r + h + rbi, data = posey_data)
##
## 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
```

Looking at the summary of the regression model above, it is clear that runs, hits, and rbi's are not indicative of the weighted runs created plus (wrc+) statistic. Our null hypothesis in this case is that these coefficients have no relationship or effect on the wrc+ stat. If we were to rerun this experiment (analyze the relationship between wrc and r,h,rbis for another player) the probability, represented in the model as the p-value, of seeing a difference as or more extreme as the observed data is high for each coefficient. Generally, the accepted cut-off to reject the null hypothesis is a p-value of < 0.05. Because none of the coefficients are less than 0.05, we fail to reject the null hypothesis. Also of interest is the R^2 value equal to 0.72, indicating that our model does a fairly good job of explaining the variation in the data. However, this seems odd considering the high p-values. It may be possible that the model is overfitted, meaning there are too many independent variables being used forcing an inflation of the R^2 value. Or there is the possibility that the terms are non-linear leading to unexpected results.

Research & 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?

Baseball is a world full of statistics where no single metric reigns supreme. For a long period of baseball's history, the fundamental stats - at-bats, hits, runs, rbis, etc... - were the only stats that people had to go off

of. It wasn't until the 1950's that the Dodgers executive Branch Rickey and statistician Allan Roth invented on-base percentage (OBP) and even then, this useful measurement of offensive performance did not become an official MLB stat until 1984.[1] Combined with slugging percentage, a metric used to measure a player's power, these two metrics evolved into what we now know as on-base plus slugging (OBS).

$$OBP = (H + BB + HBP)/(AB + BB + HBP + SF)$$

$$SLG = (1B + 2 * 2B + 3 * 3B + 4 * HR)/AB$$

OBS is a fairly accurate representation of a hitter's overall productivity because it accounts for a player's ability to get on base (OBP) and their ability to hit for extra bases. However, these two factors are not equal in terms of their overall contribution to offense. Because OBP is about twice as valuable than slugging, this metric tends to overrate power hitters and undervalue high frequency on-base players.[2] Despite this minor shortcoming, OBS has and continues to be a simple measurement of a player's offensive performance at any level of play.

Another metric, batting average on balls in play (BABIP), measures a player's batting average based on plate appearances that result in the ball entering the field of play (home runs, walks, strike out, hit by pitch not included). While BABIP is a common place metric in fantasy baseball circles, it is often misinterpreted because it doesn't account for the defensive factor of a ball in play or the sheer luck of a player's performance (which can also be tied to a variety of uncontrollable factors).[3]

In 2006, three baseball analysts Tom Tango, Mitchel Lichtman, And Andrew Dolphin published the book The Book: Playing the Percentages in Baseball.[4] In it, they revealed a new metric, an evolved version of OBS called weighted on-base average or wOBA (pronounced whoa-buh) which applied a linear weighting system to properly value the individual outcomes of an at-bat in porportion to their overall impact to scoring runs.[5]

$$wOBA = (.693*BB+.722*HBP+.876*1B+1.231*2B+1.551*3B+1.978*HR)/(AB+BB-IBB+SF+HBP)$$

The weights change on a yearly basis and the formula above reflects the values for the 2017 season.[6] And because wOBA is a rate stat that normalizes a player's performance to the total number of plate appearances, it has the advantage of allowing one to compare players across the league. Since it's inception wOBA has generally been regarded as the go-to catch-all metric for evaluating a hitter's performance. Yet still, even it has it's drawbacks.

Taking wOBA one step further, we arrive back where we started with weighted runs created plus (wRC+). The major advantage of wRC+ is that it include constants for the ballpark and era to account for the style of play of a particular player. Furthermore, it is easy to understand as the value of a player as a percentage is relative to the league average of 100.[7] That way, you can compare any player to another regardless of the season they played in.

Whether you are a professional baseball analyst or an amateur fantasy fanatic, there is no single metric that will provide you with the best possible indicator of a player's batting performance. Like anything in life, never put all your eggs in one basket. MLB, ESPN, Bleacher Report, you name it; they all rely on a variety of metrics to make the best decision on a player's performance. At a professional level, it is common place to see these stats plugged into computer models to make projections. In a 2015 article on fantasy baseball projections, ESPN discusses 7 different projection systems or models used, one of the most popular being

[1] http://m.mlb.com/glossary/standard-stats/on-base-percentage
[2] http://www.fangraphs.com/library/offense/ops/
[3] http://www.espn.com/fantasy/baseball/flb/story?page=mlbdk2k11babipprimer
[4] https://en.wikipedia.org/wiki/Tom_Tango
[5] http://www.hardballtimes.com/why-woba-works/
[6] http://www.fangraphs.com/guts.aspx?type=cn
[7] https://www.lookoutlanding.com/2017/3/7/14783982/an-idiots-guide-to-advanced-statistics-woba-and-wrc-sabermetrics

Marcel developed by Tom Tango.[8]. And likewise at the fantasy level, sites like FanDuel, Fantasy Labs, Roto World, and FanGraphs encourage making use of a variety of data to make the best decision about a player.

[8]http://www.espn.com/fantasy/baseball/story/_/page/mlbdk2k15_projectionstalk/how-fantasy-baseball-projections-calculated-how-best-use-th