SAIL Lab Application

How Stolen Bases affect Runs Scored.

library(Stat2Data)  
library(readr)  
library(rmarkdown)

## Warning: package 'rmarkdown' was built under R version 4.2.2

**Introduction:** Since the 2002 season in Major League Baseball, The United State’s “national pastime” has become more and more reliant on raw data and analytics to influence baseball operations for each team around the league. In modern day baseball, one of the biggest tasks for each organization is to bring up new talent from within their organization, for this is a much cheaper alternative then getting players from other teams. As such, the MLB draft has become the best opportunity to gain young players for each team’s organization. For hundreds of high school and collegiate athletes around the nation and even the world, the MLB Draft is an opportunity for their dreams to become reality and play for a premier baseball organization. Nevertheless, this pressure also falls heavily on each organization to make the right decision for choosing which of these athletes are going to be the future of their franchise. Thus, most organizations use a ranking system to measure the potential of these prospects. Typically, each hitter is measured in Power, Contact, Arm Strength, Defense, and Speed.

The Power and Contact categories measure the hitter’s ability to hit a baseball for power and the hitter’s ability to make contact with the baseball, respectively. Whereas the Arm Strength and Defense categories are supposed to measure the prospects ability to play defense. The last category is Speed. This category measures the speed of the prospect and can be applied to both the offensive and defensive side of the game. However, in modern analytics, we have a tendency to undervalue the measure of players speed when discussing offensive output. For instance, arguably the most important statistic for measuring runs scored in a game is the OPS or On Base Plus Slugging Statistic, for this measures how consistent the hitter makes contact with the baseball for a hit, but also accounts for the power of the hitter. Nonetheless, it doesn’t directly take into consideration how the speed of the hitter influences the offensive efficiency of the player, despite Speed being an integral part of the scouting process in measuring the potential of a player. **Therefore, when we incorporate a measure of speed on the base pads in our OPS statistics, does it better reflect the runs scored in a game?**

**OPS Calculation and Correlation to Runs Scored:** The OPS or On Base Plus Slugging statistic is one of the biggest indicators we use to measure the success of a hitter. We first use the On Base Percentage, which is the sum of walks, hits, and hit by pitches divided by the sum of walks, hits, hit by pitches, at bats, and sacrifice flies.

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

We add this to our slugging statistics, which is total bases divided at bats.

**TB = H + 2x2B + 3x3B + 4xHR => SLG = TB/AB**

As such, the sum of the OBP statistic and the SLG statistic outputs our On Base Plus Slugging or OPS statistic. The OPS of a team has a strong correlation and effect size with runs scored by the team, as depicted in the model below.

**OPS = OBP + SLG**

team\_data\_prelim <- read\_csv("C:/Users/James Jessup/Desktop/Programming Workspaces/STOR 455/mlb\_teams\_1.csv")

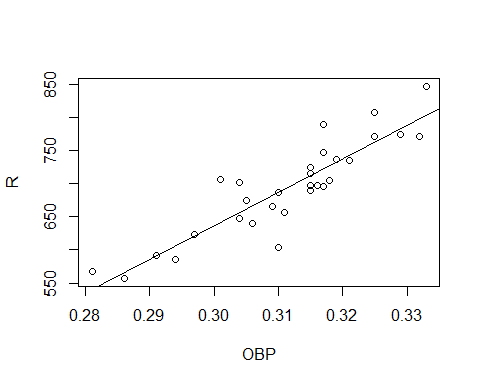
## Rows: 30 Columns: 10  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): Tm  
## dbl (9): AB, R, SB, CS, OBP, SLG, OPS, OPS+, TB  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

team\_data = subset(team\_data\_prelim, R < 1000)  
team\_data

## # A tibble: 30 × 10  
## Tm AB R SB CS OBP SLG OPS `OPS+` TB  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Arizona Diamondbacks 5351 702 104 29 0.304 0.385 0.689 95 2061  
## 2 Atlanta Braves 5509 789 87 31 0.317 0.443 0.761 111 2443  
## 3 Baltimore Orioles 5429 674 95 31 0.305 0.39 0.695 97 2119  
## 4 Boston Red Sox 5539 735 52 20 0.321 0.409 0.731 102 2268  
## 5 Chicago Cubs 5425 657 111 37 0.311 0.387 0.698 96 2097  
## 6 Chicago White Sox 5611 686 58 10 0.31 0.387 0.698 97 2172  
## 7 Cincinnati Reds 5380 648 58 33 0.304 0.372 0.676 83 2003  
## 8 Cleveland Guardians 5558 698 119 27 0.316 0.383 0.699 102 2126  
## 9 Colorado Rockies 5540 698 45 20 0.315 0.398 0.713 90 2203  
## 10 Detroit Tigers 5378 557 47 24 0.286 0.346 0.632 84 1859  
## # … with 20 more rows

*Data retrieved from* [*https://www.baseball-reference.com/leagues/majors/2022.shtml*](https://www.baseball-reference.com/leagues/majors/2022.shtml)

plot(R~OBP, data = team\_data)  
team\_mod = lm(R~OBP, data = team\_data)  
abline(team\_mod)



summary(team\_mod)

##   
## Call:  
## lm(formula = R ~ OBP, data = team\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -84.498 -20.428 -4.797 12.116 66.122   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -879.3 146.7 -5.994 1.86e-06 \*\*\*  
## OBP 5054.3 470.9 10.732 1.98e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.37 on 28 degrees of freedom  
## Multiple R-squared: 0.8044, Adjusted R-squared: 0.7975   
## F-statistic: 115.2 on 1 and 28 DF, p-value: 1.977e-11

Since the correlation between teams was .8044 and the p-value for this model is <.001, we can conclude a strong relation between OPS and Runs Scored.

**Incorporating Stolen Bases into OPS:** When taking a look at the offensive efficiency of a hitter, what is the difference between a double and single followed by that hitter stealing second base? At the end of the day, there is now a runner on second base. Of course, pitcher and runner approach and mentality may be different between hitting a double compared to a single then a stolen base; nevertheless, there are the same amount of total bases.

As such, we are going to incorporate Stolen Bases and the amount of times a Runner is caught stealing into our Total Bases, and then calculate the OPS from that. Thus, our slugging statistic becomes the sum of total bases and stolen bases, then subtracting times caught stealing and dividing that by at bats. We will call this statistic On Base plus Slugging plus Stealing or OBSS.

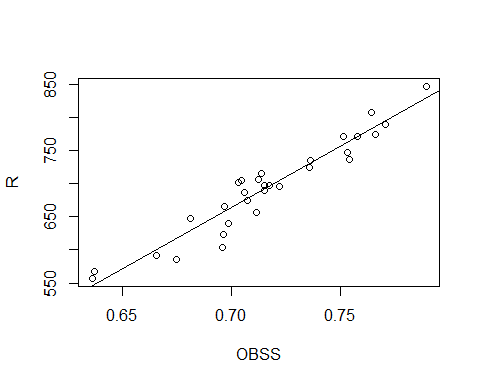
**OBSS = OBP + (TB + SB - CS)/AB**

Now we will calculate this aggregate statistic for every team in the MLB using our current data.

team\_data$OBSS = team\_data$OBP +   
 (team\_data$TB + team\_data$SB - team\_data$CS)/team\_data$AB  
team\_data

## # A tibble: 30 × 11  
## Tm AB R SB CS OBP SLG OPS `OPS+` TB OBSS  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Arizona Diamond… 5351 702 104 29 0.304 0.385 0.689 95 2061 0.703  
## 2 Atlanta Braves 5509 789 87 31 0.317 0.443 0.761 111 2443 0.771  
## 3 Baltimore Oriol… 5429 674 95 31 0.305 0.39 0.695 97 2119 0.707  
## 4 Boston Red Sox 5539 735 52 20 0.321 0.409 0.731 102 2268 0.736  
## 5 Chicago Cubs 5425 657 111 37 0.311 0.387 0.698 96 2097 0.711  
## 6 Chicago White S… 5611 686 58 10 0.31 0.387 0.698 97 2172 0.706  
## 7 Cincinnati Reds 5380 648 58 33 0.304 0.372 0.676 83 2003 0.681  
## 8 Cleveland Guard… 5558 698 119 27 0.316 0.383 0.699 102 2126 0.715  
## 9 Colorado Rockies 5540 698 45 20 0.315 0.398 0.713 90 2203 0.717  
## 10 Detroit Tigers 5378 557 47 24 0.286 0.346 0.632 84 1859 0.636  
## # … with 20 more rows

plot(R~OBSS, data = team\_data)  
new\_team\_mod = lm(R~OBSS, data = team\_data)  
abline(new\_team\_mod)



summary(new\_team\_mod)

##   
## Call:  
## lm(formula = R ~ OBSS, data = team\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -52.627 -10.447 1.607 16.693 32.557   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -632.90 77.26 -8.192 6.44e-09 \*\*\*  
## OBSS 1852.56 107.73 17.196 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.53 on 28 degrees of freedom  
## Multiple R-squared: 0.9135, Adjusted R-squared: 0.9104   
## F-statistic: 295.7 on 1 and 28 DF, p-value: < 2.2e-16

When you try to predict Runs scored using the OBSS statistic, we see how there is a R Squared value of .9135 and a p-value of <.001. As such, our model using the OBSS explains about 10.91% more variance than the model using OBP as the predictor for Runs. As such, this is a significant jump in explaining how Runs are scored in MLB. Therefore, merely looking at preliminary data, we can conclude that its possible when we take stolen bases into consideration for offensive efficiency, we can make a better estimate of how many runs a team should score in a game.