---
title: "BaseballSuccessPredictors"
author: "Hailey Yabroudy"
date: "3/22/19"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

****Research question: Can quantitative measures of a team's hitting and/or pitching success be used as predictors for the number of wins a team will have in a season?****

**Hypothesis: As a team's batting average increases, their number of season wins will also increase. As a team's ERA (earned run average) decreases, their number of season wins will increase.**

**Method: We will be using Lahman's Baseball Database. We will begin by building a linear regression model using data from the Red Sox, beginning with the single best predictor, and testing additional predictors until the final model is constructed. We will test this model on data from other teams, and compare the residuals with the residuals of the blueprint team, the Red Sox.**

First, select only the necessary rows and columns and to make a more compact and manageable data frame.

```{r message=FALSE, warning=FALSE}
library(mosaic)
library(Lahman)
View(Teams)
```

```{r}
# select the years 1960 to 2016
year<-filter(Teams, yearID %in% (1960:2016))
# select the team Red Sox from the years 1960 to 2016
name<-filter(year, name == "Boston Red Sox")
# select the variables: year, wins, losses, earned run average, at-bat, hits
Sox<-select(name, yearID, W, L, ERA, ER, AB, H)
```

Next, create a few more variables that will be used in the model: batting average and win/loss ratio.

```{r}
#create the variable batting average, called BA, which is hits divided by at-bats
Sox<-mutate(Sox, BA = Sox$H/Sox$AB)
#create the variable win/loss ratio, called ratio, which is wins divided by losses
Sox<-mutate(Sox, ratio = Sox$W/Sox$L)
head(Sox)
```

**Correlations of potential Xs and Ys**

```{r}
#create a matrix to eliminate year, losses, and at-bat variables
matrix<-Sox[, -c(6, 3, 1)]
#correlations
cor(matrix)
```

**Hits and number of wins have the highest correlation at 0.65. Interestingly, batting average has a lower correlation to wins than does hits, but this relationship flips when correlated to win/loss ratio.**

**ERA, the only measure of pitching success that we looked at, has a slightly higher correlation to number of wins than to win/loss ratio, -0.29.**

**ERA has weak correlations to hits and batting average, so multicollinearity is not a concern.**

**Hits and batting average are highly correlated, but this will not be a problem, as it would not make sense to use multiple measures of batting success as explanatory variables.**

```{r}
#scatter plots for Hits and ERA against Number of Wins
plot(W~H, data=Sox, pch=19, col="red", xlab="Total Hits in a Season", ylab = "Number of Wins in a Season", main="Hits per season against Wins per season")
```

```
plot(W~ERA, data=Sox, pch=19, col="blue", xlab="Earned Run Average in a Season", ylab = "Number of Wins in a Season", main="Earned Run Average per season against Wins per season")
```

**Hits has a moderate positive correlation to number of wins per season. There are two points in the lower left corner separate from the rest of the cluster, but they appear to be roughly in line with the rest of the data. ERA has a weakly negative correlation to number of wins per season.**

```{r}
#checking symmetry/normality of variables' distributions
histogram(Sox$W)
histogram(Sox$H)
histogram(Sox$ERA)
```

**Number of wins per season is slightly skewed to the left. Hits and ERA appear mostly symmetrical.**

**Now to build the first model**
```{r}
m1<- lm(W~H, data=Sox)
plot(W~H, data=Sox, pch=19, col="red", xlab="Total Hits in a Season", ylab = "Number of Wins in a Season", main="Hits per season against Wins per season")
abline(m1, lwd=2)
summary(m1)
```

**The equation for this model is: Predicted Wins in a Season = 0.05(Total Hits in a Season) + 5.97.**

**As the Red Sox’s total hit number increases by one, their predicted wins in a season increases by 0.05.**
**In terms of whole numbers, the slope indicates that the Red Sox number of wins in a season would increase by one for every 20 hits their batters make.**

**The r squared value indicates that 43% of the variability in Number of Wins can be explained by the number of hits. This means that 57% of the variability in number of wins is caused by variables other than number of hits.**

**R gives the slope a p-value of less than 0.0001. The model is highly significant.**

Correlation of both Xs to Y

Investigating/ruling out multicollinearity

| | W | ERA | H | BA | ratio |
|---|---|---|---|---|---|
| W | 1.0000000 | -0.292578235 | 0.658006249 | 0.4548225 | 0.86613696 |
| L | -0.5282073 | 0.172307826 | -0.007105759 | -0.4988245 | -0.86744558 |
| ERA | -0.2925782 | 1.000000000 | 0.223851076 | 0.3762476 | -0.26063439 |
| H | 0.6580062 | 0.223851076 | 1.000000000 | 0.7013183 | 0.37366086 |
| BA | 0.4548225 | 0.376247590 | 0.701318300 | 1.0000000 | 0.53861697 |
| ratio | 0.8661370 | -0.260634388 | 0.373660858 | 0.5386170 | 1.00000000 |

| ## | W | ERA | ER | H | BA | ratio |
|---|---|---|---|---|---|---|
| ## W | 1.00000000 | -0.2925782 | 0.04751206 | 0.6580062 | 0.4548225 | 0.8661370 |
| ## ERA | -0.29257823 | 1.0000000 | 0.80741578 | 0.2238511 | 0.3762476 | -0.2606344 |
| ## ER | 0.04751206 | 0.8074158 | 1.00000000 | 0.6181853 | 0.3363600 | -0.2047731 |
| ## H | 0.65800625 | 0.2238511 | 0.61818525 | 1.0000000 | 0.7013183 | 0.3736609 |
| ## BA | 0.45482249 | 0.3762476 | 0.33636003 | 0.7013183 | 1.0000000 | 0.5386170 |
| ## ratio | 0.86613696 | -0.2606344 | -0.20477312 | 0.3736609 | 0.5386170 | 1.0000000 |

Model with best predictor

```
Call:
lm(formula = W ~ H, data = Sox)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1988  -5.6876   0.3057   5.9020  13.9214

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.971690  12.118980   0.493    0.624
H            0.052985   0.008176   6.481 2.67e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.73 on 55 degrees of freedom
Multiple R-squared:  0.433,   Adjusted R-squared:  0.4227
F-statistic:    42 on 1 and 55 DF,  p-value: 2.668e-08
```

## **Question: What team has the most postseason wins?**

This project demonstrates my ability to take specific information from huge databases and work with it in order to problem solve. This is a very useful skill if you were to feed announcers stats to talk about during a game.

```
View(SeriesPost)
```

**Filter out all of the years before 1960.**

```
year<-filter(SeriesPost, yearID %in% (1960:2016))
```

**Create a table with the variables I will be working with.**

```
win1 <- select(year, yearID, teamIDwinner, teamIDloser, wins, losses)
```

**Find the total number of wins they have.**

```
winsum <- tally(group_by(win1, teamIDwinner), wins)
a <- arrange(winsum, desc(n))
losers <- tally(group_by(win1, teamIDloser), losses)
```

**Renamed the column in the losers table to “teamIDwinner” so it will be easier to join later.**

```
colnames(losers) <- c("teamIDwinner", "n")
```

**Used a join to combine the tables together so there is one row for each team. There will still be two columns for the games they won when they won the series overall and when they lost overall.**

```
test3<- full_join(winsum, losers, by = c("teamIDwinner"))
```

**Added together the columns of the games won and arranged them so the team with the most wins is on top.**

```
sum <- mutate(test3, total = n.x + n.y)
a <- arrange(sum, desc(total))
View(a)
```

**Made a barchart to show the data visually.**

```
barchart(total~teamIDwinner, data = a, main = "Postseason Wins By
Team", xlab = "Team ID", ylab = "Number of Wins")
```

**According to the data, the team with the most wins is NYA with 109. The second highest is SLN with 78. In third, it's BOS with 49 wins.**

**The team with the most postseason wins is NYA (just based on my knowledge of the MLB and a quick search, I know that this is probably supposed to be NYY, or the Yankees). I tried to look in the Teams Franchise table to see what NYA stood for and it wasn't there. SLN wasn't there either. But, going off of what I think it's supposed to mean, the Yankees have the most with 109, the Cardinals are second with 78, and the Red Sox are third with 49.**

## **Last Question: What is the average age of players who are awarded MVP?**

In this problem, I had to go through the AwardsPlayers database and filter out all of the award that weren't MVP. I then had to use the players birth years and subtract it from the year they won the award and then find the average of all of those numbers.
This was a pretty straight forward problem, but it was interesting to find out nonetheless.

```
View(AwardsPlayers)
year<- filter(AwardsPlayers, yearID %in% (1960:2016))
mvp <- filter(year, awardID == "Most Valuable Player")
```

**I chose these dates because I figured they would be at least playing by 1960.**

```
age<- filter(Master, birthYear %in% (1929:1996))
birth <- select(age, playerID, birthYear, nameFirst, nameLast)
yos<- full_join(mvp, birth, by = c("playerID"))
yos <- na.omit(yos)
agediff <- mutate(yos, total = yearID - birthYear)
agediff <- select(agediff, nameFirst, nameLast, total)
mean(agediff$total)
```

**The mean value for a player to receive the MVP award is 28.9.**