Measuring Batting Performance

Introduction

The purpose of this analysis is to investigate the usefulness of various bsaeball statistics to predict offensive production. After running linear regressions on several batting performance measures, we determined that BRA (Batting Run Average) showed a strong correlation to Runs scored. BRA is calculated by multiplying OBP by SLG, which essentially creates a product of ability to get on base and hitting for power. The data clearly show that teams with high BRA score a large number of runs. Using this data, we then decided to compare BRA to R in the American League in the years 2001 and 2010. The output provided interesting insight into the varying production of teams over one decade. Regression output and graphs are shown below (note that "ANA" in 2001 and "LAA" in 2010 both represent the Angels franchise).

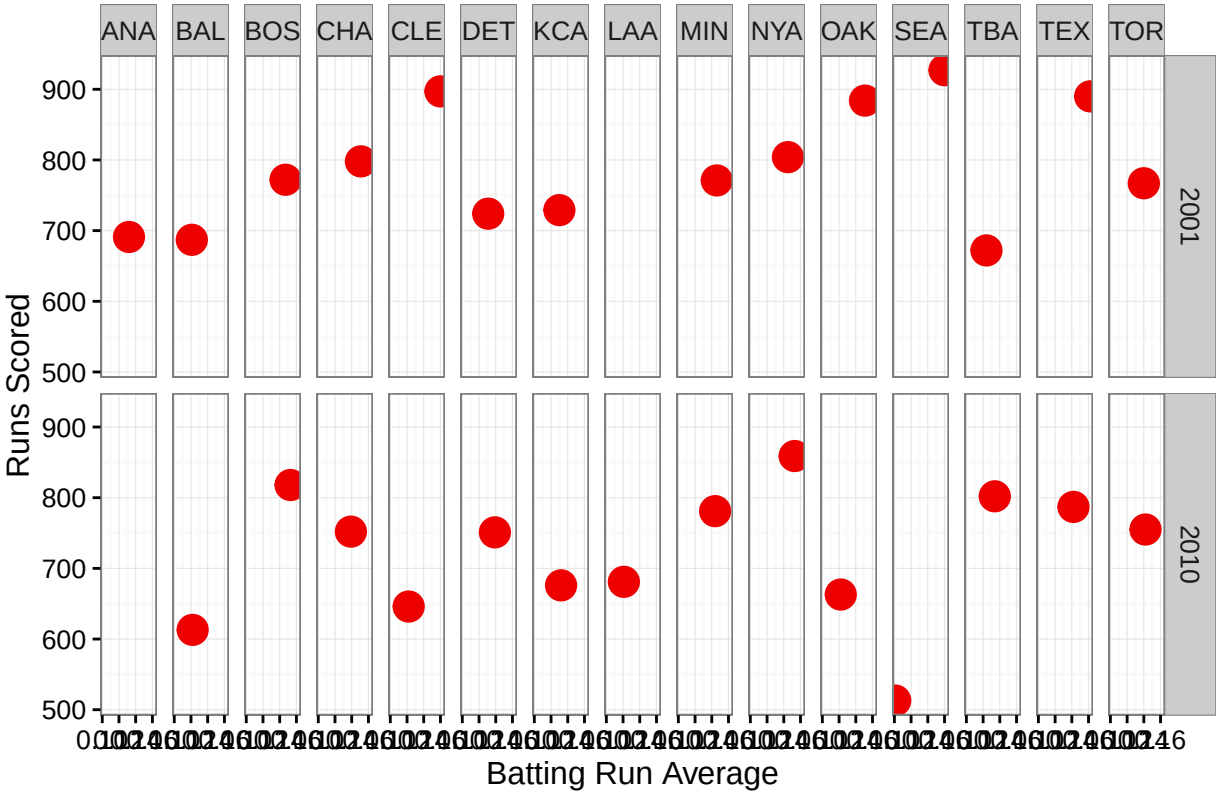
Methodology

We used data from the Lahman database to complete our analysis. R coding using dplyr and ggplot was used to extract data and create graphs.

```
data(Teams)
tm.batting <- Teams %>% select(-(Rank:WSWin),-(RA:teamIDretro)) %>%
    filter(yearID==2001|yearID==2010) %>% filter(lgID=="AL") %>%
    filter(!is.na(HBP),!is.na(SF)) %>%
    group_by(yearID,teamID) %>%
    summarize(BA = round(H/AB,3),
              PA = AB+BB+HBP+SF,
              OBP = round((H+BB+HBP)/PA,3),
              SING = H - (X2B + X3B + HR),
              TB = (SING + (2*X2B) + (3*X3B) + (4*HR)),
              SLG = round(TB/AB,3),
              OPS = OBP + SLG,
              weightedOPS = 1.5*OBP + SLG,
              ISO = SLG - BA,
              TAv = (TB + HBP + BB + SB) - CS/(AB - H) + CS,
              RC = ((H + BB - CS)*(TB + 0.55*SB))/(AB + BB),
              R,
              BRA = round(OBP*SLG,3),
              SOR = round(SO/PA,3),
              BBR = round(BB/PA,3))

ggplot(tm.batting, aes(BRA, R)) + geom_point(col = "red2", cex=5) + xlab("Batting Run Average") +
  ylab("Runs Scored") + ggtitle("Figure 1.") + theme_bw() + facet_grid(yearID~teamID) + stat_smooth(meth
```

Figure 1.




```
model2 <- lm(R~BRA, data=tm.batting)
summary(model2)
```

```
##
## Call:
## lm(formula = R ~ BRA, data = tm.batting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.649 -23.665  -3.165  13.431  72.969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -90.41      54.24  -1.667    0.108
## BRA          6115.20     390.77  15.649 9.52e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.53 on 26 degrees of freedom
## Multiple R-squared:  0.904,  Adjusted R-squared:  0.9003
## F-statistic: 244.9 on 1 and 26 DF,  p-value: 9.523e-15
```

Results

Comparisons between team performance in runs as a function of batting run average shows how an given team's production changes over the course of a decade. Notable decreases in run production and batting

run average belong to the Seattle Mariners and Cleveland Indians; the top two producers in 2001, these two teams were at the top of the league in runs scored and in the top three in BRA, but were near the bottom in 2010. The Detroit Tigers and the Tampa Bay Rays made slight improvements in production between 2001 and 2010. The Toronto Blue Jays, Kansas City Royals, and Baltimore Orioles showed very little change in BRA, but all had a noticable decrease in run production. This pattern suggests that scoring levels were inconsistent among these teams despite their similar BRA statistics. As seen in the regression of BRA as a predictor of runs, Batting Run Average is a very significant statistic in seeking to predict offensive production. Through the statistical analysis and the case study of the American League in 2001 and 2010, it is safe to say that Batting Run Average is important in an analysis of offensive predictive statistics.