Exploring Some Ideas from Bill James

Dr. Ekstrom, Abrielle Agron

Purpose: Each activity illustrates a little sports analytics, a little statistics and a little *R*.

Sports Analytics: Try to explain how runs are scored in Baseball, that is, to try to find a statistic that explains as much run scoring as possible.

Statistics: regression

R: Introduction to *R*, including: reading a csv, using a notebook, linear modeling, and plotting.

Demonstration - Runs Scored and Homeruns: In this demonstration, we are going to examine the relationship between the number of homeruns hit and the number of runs scored. Using function terms, we will consider homeruns as the independent variable and runs as the dependent variable. Using statistical terms, homeruns is called the explanatory variable and runs is the response variable.[1]

The data given in `BaseballRuns.csv` shows offensive statistics for MLB teams from 2000-2005. Import the data `BaseballRuns.csv` into *R* using

```
bbr=data=read.csv("BaseballRuns.csv")
```

The table of data is now in a data frame[2] called `bbr`.

We can get a summary of what is in this data frame:

```
summary(bbr)
```

```
##       year           lg                team                G
##  Min.   :2000   Length:180         Length:180         Min.   :161.0
##  1st Qu.:2001   Class :character   Class :character   1st Qu.:162.0
##  Median :2002   Mode  :character   Mode  :character   Median :162.0
##  Mean   :2002                                         Mean   :161.9
##  3rd Qu.:2004                                         3rd Qu.:162.0
##  Max.   :2005                                         Max.   :163.0
##      Ghome             W                L                R
##  Min.   :77.00   Min.   : 43.00   Min.   : 46.00   Min.   :574.0
##  1st Qu.:81.00   1st Qu.: 71.00   1st Qu.: 71.00   1st Qu.:714.0
##  Median :81.00   Median : 82.00   Median : 80.00   Median :768.0
##  Mean   :80.88   Mean   : 80.93   Mean   : 80.93   Mean   :773.6
```

[1] It may seem arbitrary that we picked runs to be the response variable, but it is not. Ultimately, the goal of offense in baseball is to score as many runs as possible. We are trying to understand how runs are scored - or what statistic best explains how runs are scored - is it homeruns? batting average? or stolen bases? or something else?

[2] Think of a data frame as an array on steriods - where the entries of the array can be objects more complicated than just numbers or strings

```
##  3rd Qu.:81.00   3rd Qu.: 91.00   3rd Qu.: 91.00   3rd Qu.:828.5
##  Max.   :82.00   Max.   :116.00   Max.   :119.00   Max.   :978.0
##        AB             H             X1B              X2B             X3B
##  Min.   :5330   Min.   :1300   Min.   : 850.0   Min.   :201.0   Min.   :12.00
##  1st Qu.:5497   1st Qu.:1422   1st Qu.: 933.0   1st Qu.:277.8   1st Qu.:25.00
##  Median :5545   Median :1470   Median : 966.5   Median :293.0   Median :30.00
##  Mean   :5553   Mean   :1472   Mean   : 969.6   Mean   :294.6   Mean   :30.67
##  3rd Qu.:5610   3rd Qu.:1516   3rd Qu.:1000.5   3rd Qu.:310.2   3rd Qu.:36.00
##  Max.   :5769   Max.   :1667   Max.   :1186.0   Max.   :373.0   Max.   :61.00
##        HR             BB              SO              SB
##  Min.   :116.0   Min.   :363.0   Min.   : 805.0   Min.   : 31.00
##  1st Qu.:152.0   1st Qu.:489.0   1st Qu.: 984.2   1st Qu.: 69.00
##  Median :173.5   Median :536.5   Median :1040.5   Median : 88.50
##  Mean   :177.1   Mean   :542.3   Mean   :1046.8   Mean   : 91.68
##  3rd Qu.:200.0   3rd Qu.:593.5   3rd Qu.:1099.0   3rd Qu.:109.25
##  Max.   :260.0   Max.   :775.0   Max.   :1399.0   Max.   :177.00
##        CS             HBP             SF
##  Min.   :12.00   Min.   :29.00   Min.   :25.00
##  1st Qu.:32.75   1st Qu.:51.00   1st Qu.:40.00
##  Median :40.00   Median :58.00   Median :46.00
##  Mean   :40.63   Mean   :59.46   Mean   :46.39
##  3rd Qu.:48.00   3rd Qu.:67.00   3rd Qu.:52.00
##  Max.   :74.00   Max.   :95.00   Max.   :75.00
```

The variables are:

`lg:` league - AL for American League and NL for National League

`team:` city of team (three letter abbreviation)

`G:` number of games played (162 is normal)

`Ghome:` number of home games played (81 is normal)

`W:` number of wins

`L:` number of losses

`R:` number of runs scored

`AB:` number of at bats

`X1B:` number of singles

`X2B:` number of doubles

`X3B:` number of triples

`HR:` number of homeruns

`BB:` number of walks

`SO:` number of strike outs

`SB:` number of stolen bases

`CS:` number of times caught stealing

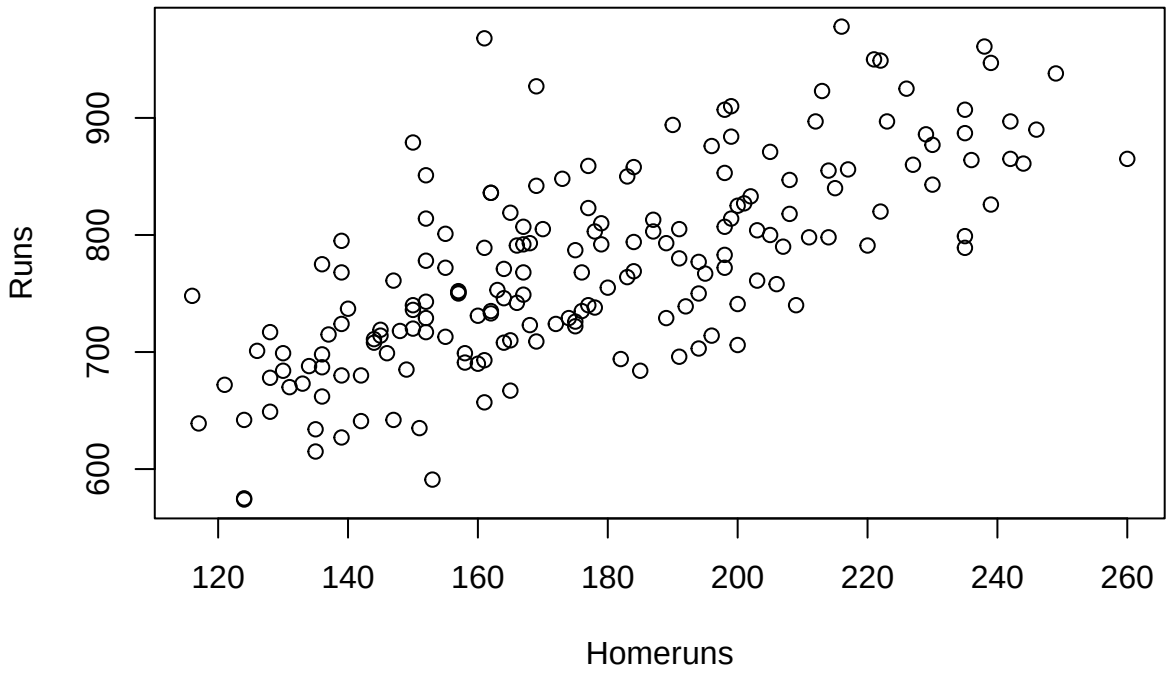
`HBP:` number of batters hit by pitch

`SF:` number of sacrifice flies

Creating and Plotting a Linear Model:

To plot R versus HR

```
plot(bbr$HR,bbr$R,xlab="Homeruns",ylab="Runs")
```



To create the "best-fitting line'' for this data:

```
g=lm(formula=R~HR, data = bbr)
```

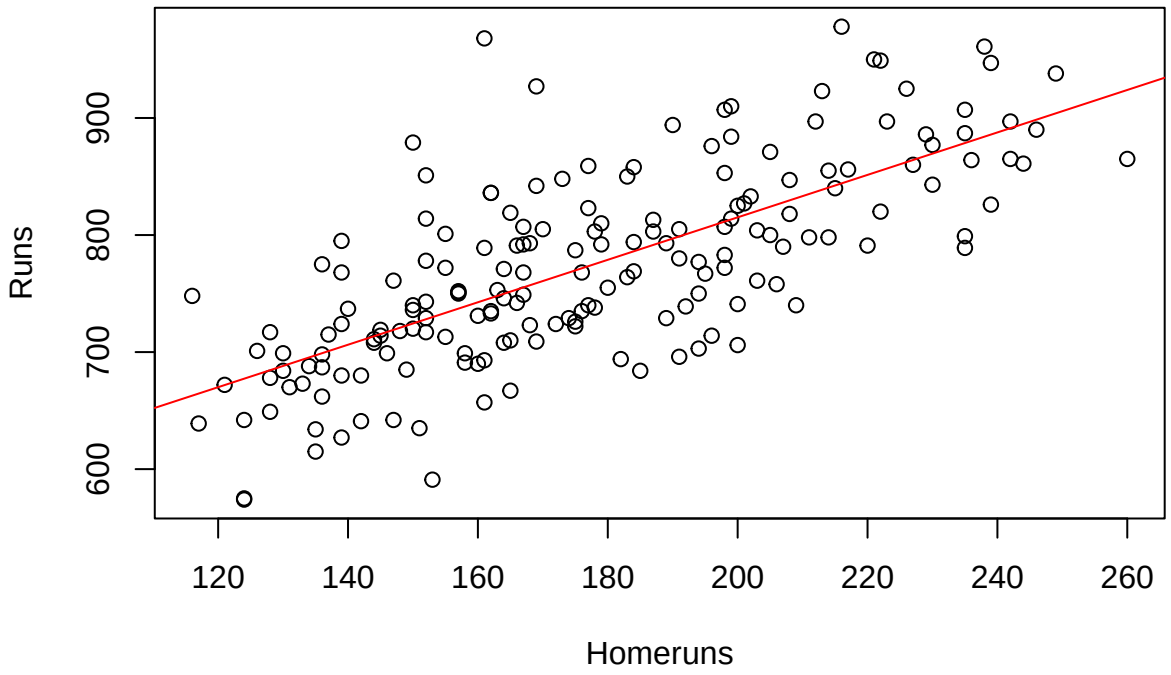
which stores this line - and much information about this line - in the variable g. To see the intercept and slope of this line, simply type:

```
g
```

```
##
## Call:
## lm(formula = R ~ HR, data = bbr)
##
## Coefficients:
## (Intercept)           HR
##     452.358        1.814
```

To add this line to the plot:

```
plot(bbr$HR,bbr$R,xlab="Homeruns",ylab="Runs")
abline(g,col="red")
```


Runs
600
700
800
900
120
140
160
180
200
220
240
260
Homeruns

What Do We Mean by "Best-Fitting Line"?

The equation for this best fitting line is

$$\texttt{R} = 452.358 + 1.815 \cdot \texttt{HR}$$

We can use this equation (model) to predict how many runs a team scored based on how many homeruns they hit. For example, the 2000 Anaheim Angels (`ANA`) hit 236 homeruns, so our model predicts they scored 880.698 runs. They actually scored 864 runs. The difference between the predicted number of runs and the actual number of runs is called a residual. In this case, the residual is 16.698. We can use *R* to calculate the predicted number of runs scored for each team

```
Pred=452.358+1.814*bbr$HR
```

and the residual for each team

```
Resid=Pred-bbr$R
```

A "best-fitting'' line should somehow minimize these residuals - and our line actually minimizes the (sum of the) squares of the residuals - which is why it is often called the least-squares line.[3] It is also called the regression line.

Important Info Regarding This Linear Model:

To get more details about our best-fitting line:

```
summary(g)
```

```
##
## Call:
## lm(formula = R ~ HR, data = bbr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -138.867  -37.312   -3.702   33.233  223.622
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 452.3584    23.6450   19.13   <2e-16 ***
## HR            1.8138     0.1312   13.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

[3] You can calculate the sum of the squares of the residuals

```
sum(Resid*Resid)
```

Now try a different line, such as

$$\texttt{R} = 410 + 2 \cdot \texttt{HR}$$

and calculate the predicted values for runs and their residuals

```
Pred2=410+2.0*HR
Resid2=Pred2-R
sum(Resid2*Resid2)
```

The sum of the squares for these residuals will be more than the sum of the squares of the residuals for our original model.

```
## Residual standard error: 58.18 on 178 degrees of freedom
## Multiple R-squared:  0.5177, Adjusted R-squared:  0.515
## F-statistic: 191.1 on 1 and 178 DF,  p-value: < 2.2e-16
```

A couple of things to note in this summary: the R-squared value (given in the output as `Multiple R-squared`) of 0.5177. This means that 51.77% of the variation in runs can be explained by the number of homeruns. The second is the Residual standard error is 58.18. This gives us an idea of the size of the average residual - for our purposes you can think of it as the standard deviation of the residuals. For example, we would expect 95% of residuals to be within the interval $(-2 \cdot 58.18, 2 \cdot 58.18)$.

Group Activity:

Consider the statistics:

`BA:` batting average - which is: $\frac{\texttt{X1B} + \texttt{X2B} + \texttt{X3B} + \texttt{HR}}{\texttt{AB}}$

`SLG:` slugging percentage - which is: $\frac{\texttt{X1B} + 2 \cdot \texttt{X2B} + 3 \cdot \texttt{X3B} + 4 \cdot \texttt{HR}}{\texttt{AB}}$

`OBP:` on-base percentage - which is: $\frac{\texttt{X1B} + \texttt{X2B} + \texttt{X3B} + \texttt{HR} + \texttt{BB} + \texttt{HBP}}{\texttt{AB} + \texttt{BB} + \texttt{HBP} + \texttt{SF}}$

`OPS:` on-base plus slugging - which is: `OBP` + `SLG`

Determine the best fit linear model for each of the following: runs versus batting average, runs versus slugging percentage, runs versus on-base percentage, and runs versus on-base plus slugging percentage. Be sure each model includes the equation that predicts the number runs scored, the R-squared value, and the Residual standard error.

Use R to compute these new statistics. For example

```
bbr$BA=(bbr$X1B+bbr$X2B+bbr$X3B+bbr$HR)/bbr$AB
```

Now you can use the variable `BA` to compute your linear model.

Turn in your R notebook in the appropriate assignment folder on Gradescope.

When You're Done

Consider the following questions:

- What variables can I make combining these statistics to give a high R-squared value?
- Download more recent data from baseballreference.com. Do the R-squared values for our statistics change dramatically?
- What is most predictive (based on R-squared) value that you can make with only 2 columns from your original dataframe? What about 3 or 4?