Data exploration and analysis

Overview

Questions about the class 19 material

Example of data exploration and analysis in baseball

Note on homework 8

If you downloaded the homework before Wednesday afternoon, adding these lines to question 4.2 will be helpful:

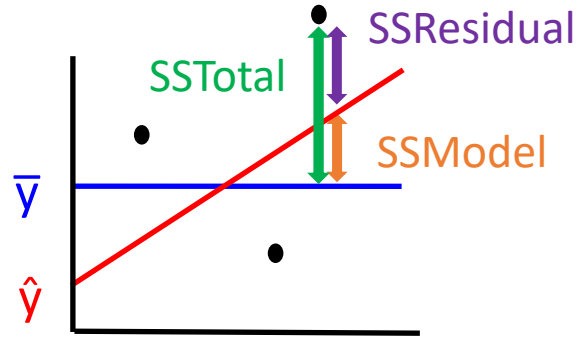
```
predict_df <- data.frame(mileage_bought = seq(0, 300000))

# this create 4 subplots which makes the plots take up less space
par(mfrow = c(2, 2))
```

Review of class 19 material

ANOVA for regression: SSTotal = SSModel + SSResidual

- R^2 = SSModel/SSTotal



Multiple linear regression:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \ldots + \hat{\beta}_k \cdot x_k$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2$$

Categorical predictors and interactions

$$\text{salary}_i \approx \beta_0 + \beta_1 \cdot \text{endowment}_i + \beta_2 \cdot x_i + \beta_3 \cdot (x_i \cdot \text{endowment}_i)$$

$$x_i = \begin{cases} 1 & \text{if Assistant Professor} \\ 0 & \text{if Full Professor} \end{cases}$$

Polynomial regression:

$$\text{salary} \approx \beta_0 + \beta_1 \cdot \text{endowment} + \beta_2 \cdot (\text{endowment})^2 + \beta_3 \cdot (\text{endowment})^3$$

Questions?

Multiple regression and data analysis continued…

Multiple regression

In multiple regression we try to predict a quantitative response variable *y* using several predictor variables $x_1, x_2, \ldots, x_k$

For multiple linear regression, the underlying model is:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots \beta_k \cdot x_k + \epsilon$$

We estimate coefficients using a data set to make predictions ŷ

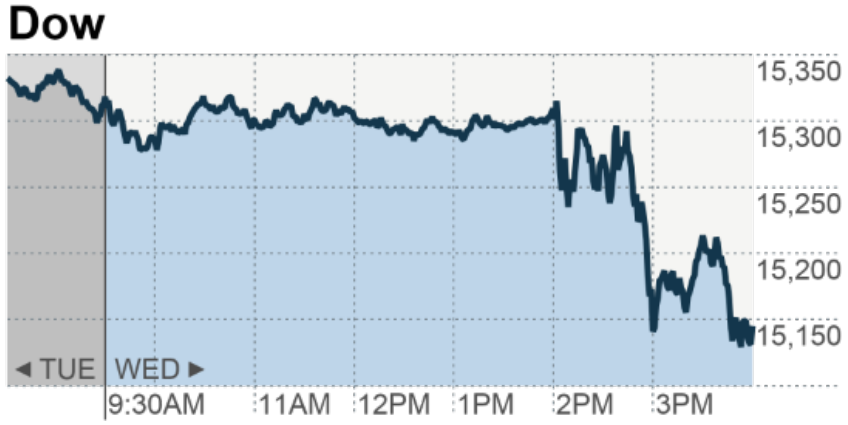
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \ldots + \hat{\beta}_k \cdot x_k$$

Multiple regression

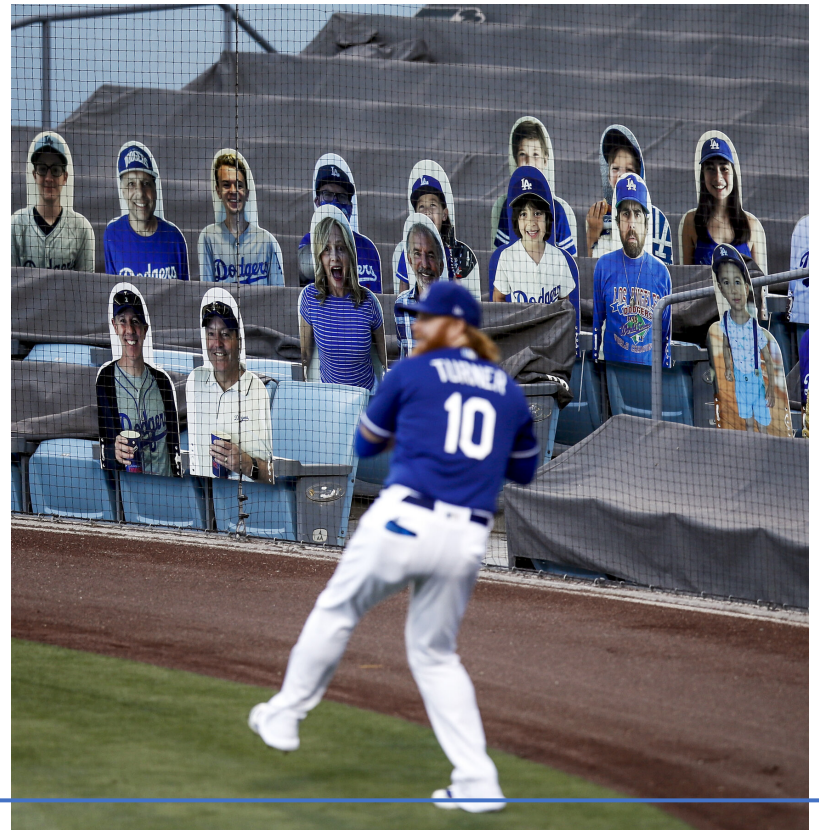
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \ldots + \hat{\beta}_k \cdot x_k$$

There are many uses for multiple regression models including:

- To make predictions as accurately as possible
- To understand which predictors (x) are related to the response variable (y)
- To create new statistics ("metrics") that give a useful numerical description of a phenomenon



What do during challenging times?



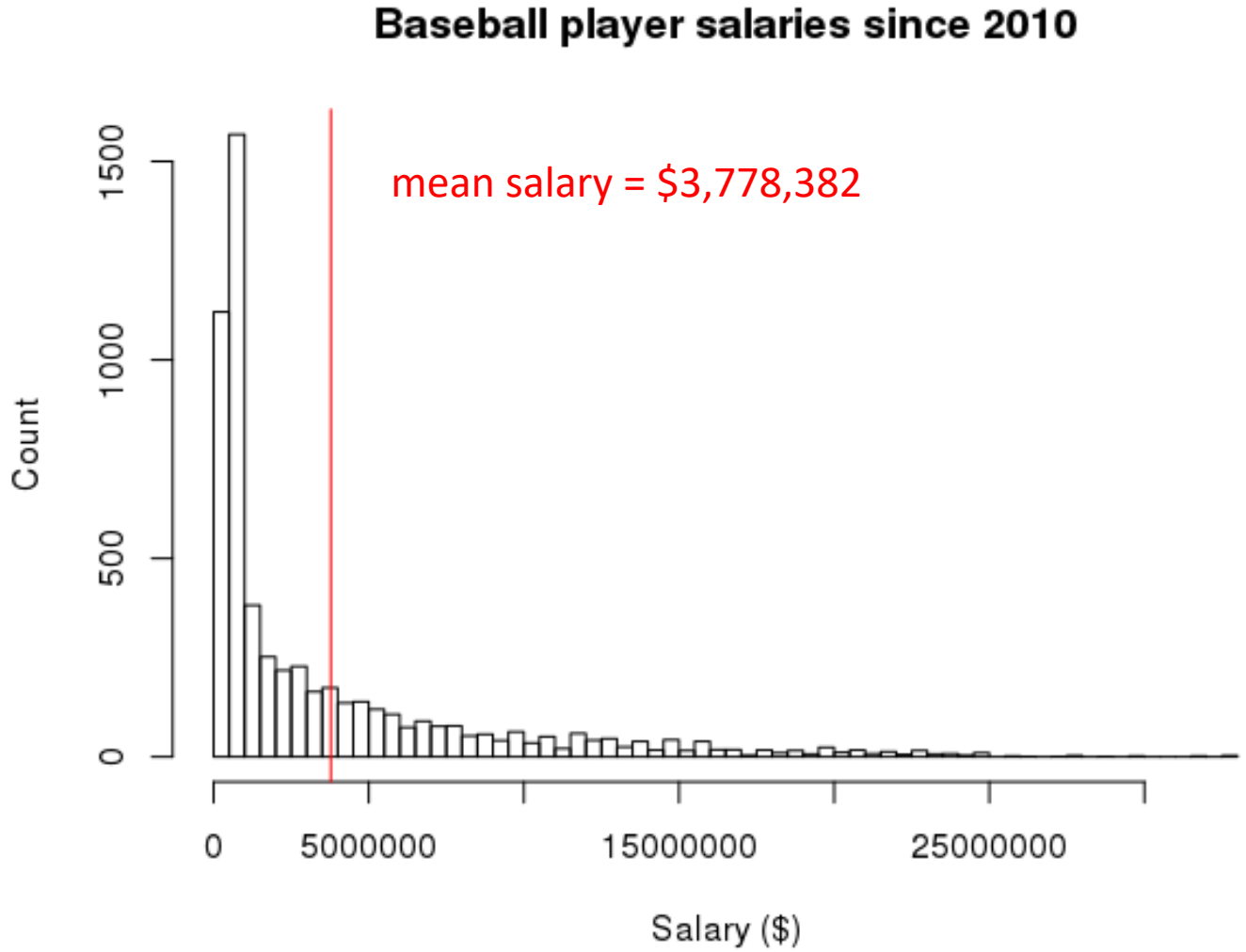
Motivation: Who is a better hitter- Derek Jeter or David Ortiz?

Derek Jeter

David Ortiz

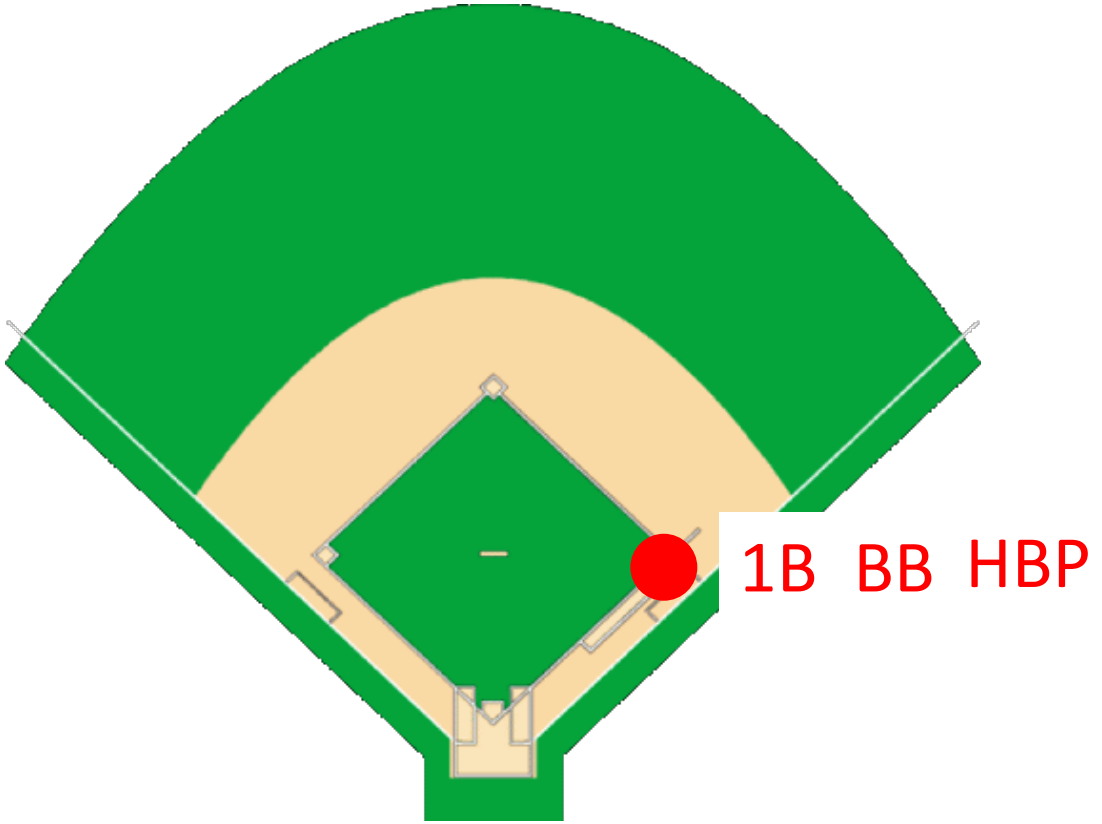
If we are going to pay these players millions of dollars, how can we assess who is best?

Baseball

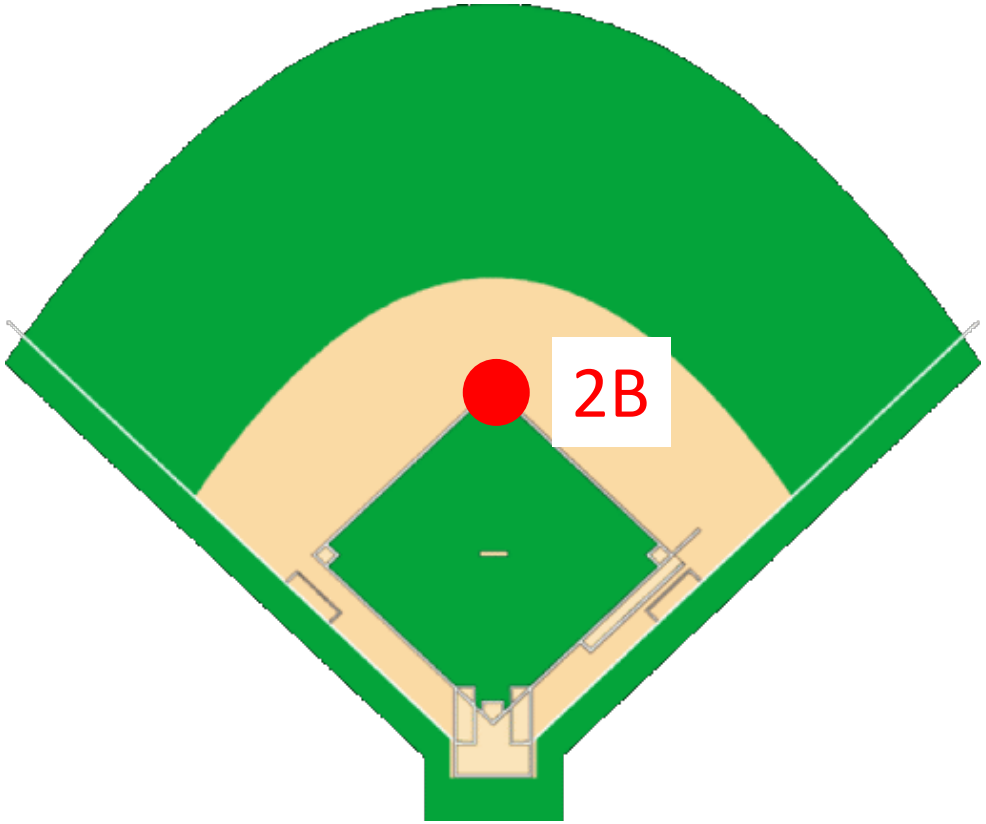

Baseball player salaries since 2010
mean salary = $3,778,382
Count
0
500
1000
1500
0
5000000
15000000
25000000
Salary ($)

Are you familiar with the rules of baseball?

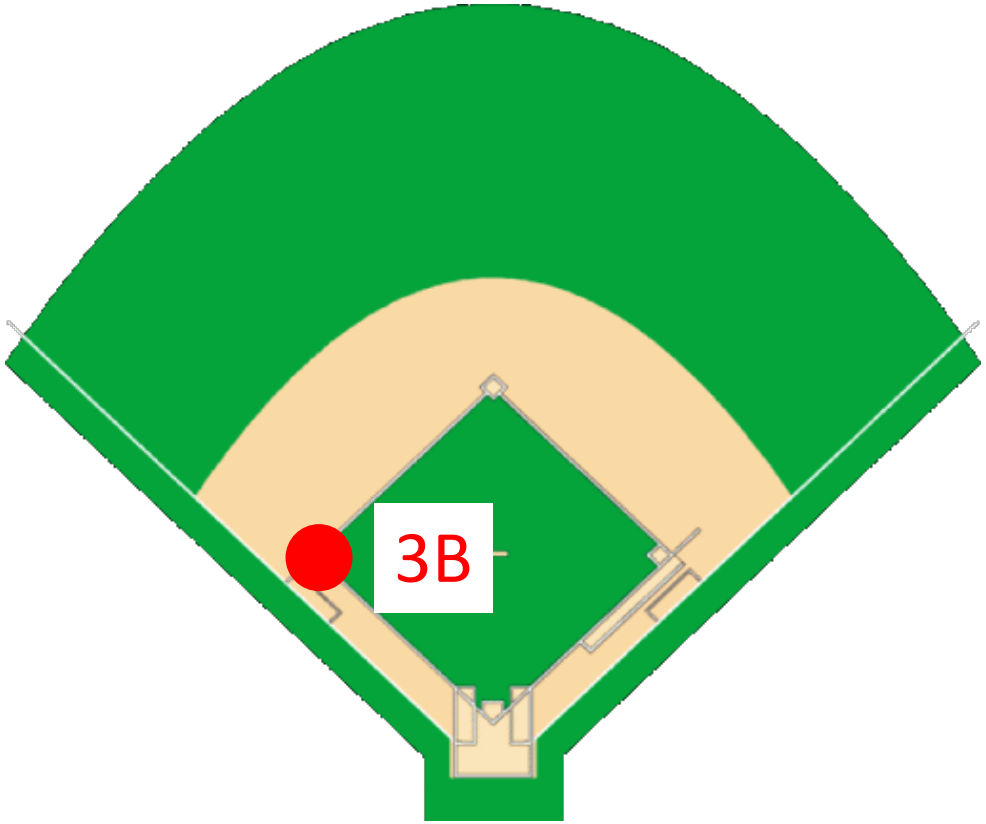
Basic rules of baseball



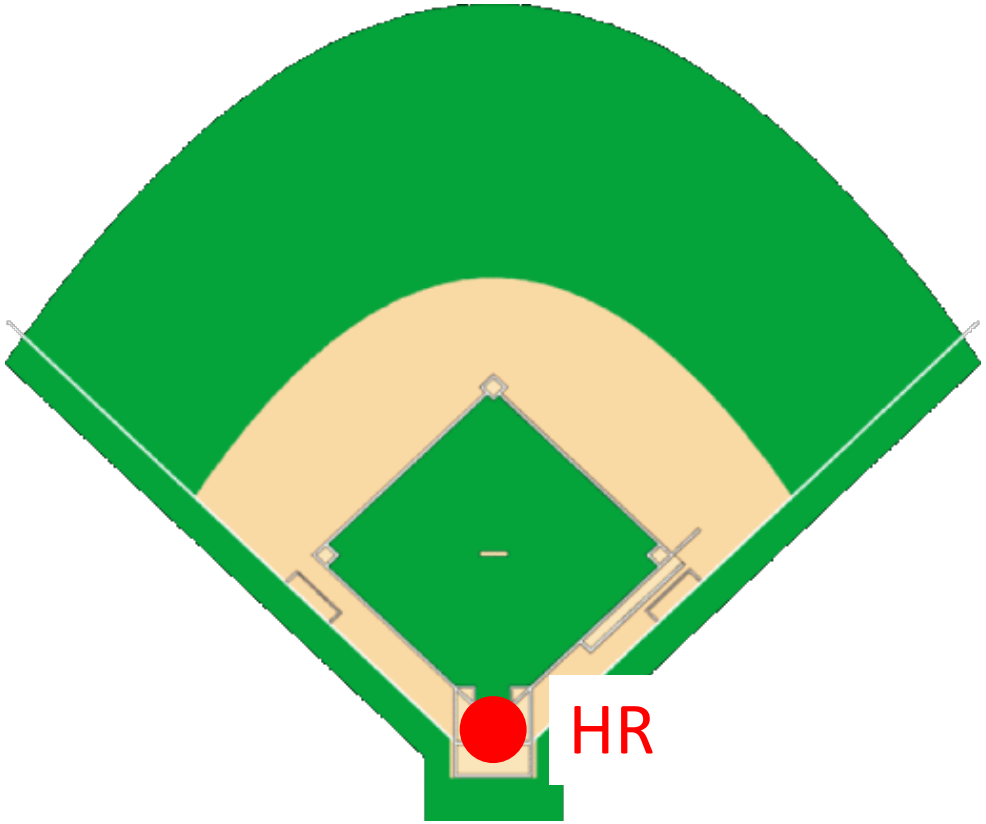
Basic rules of baseball



Basic rules of baseball



Basic rules of baseball



History of baseball statistics

Henry Chadwick (1824-1908) created the first box score in the 1859 issue of Clipper.

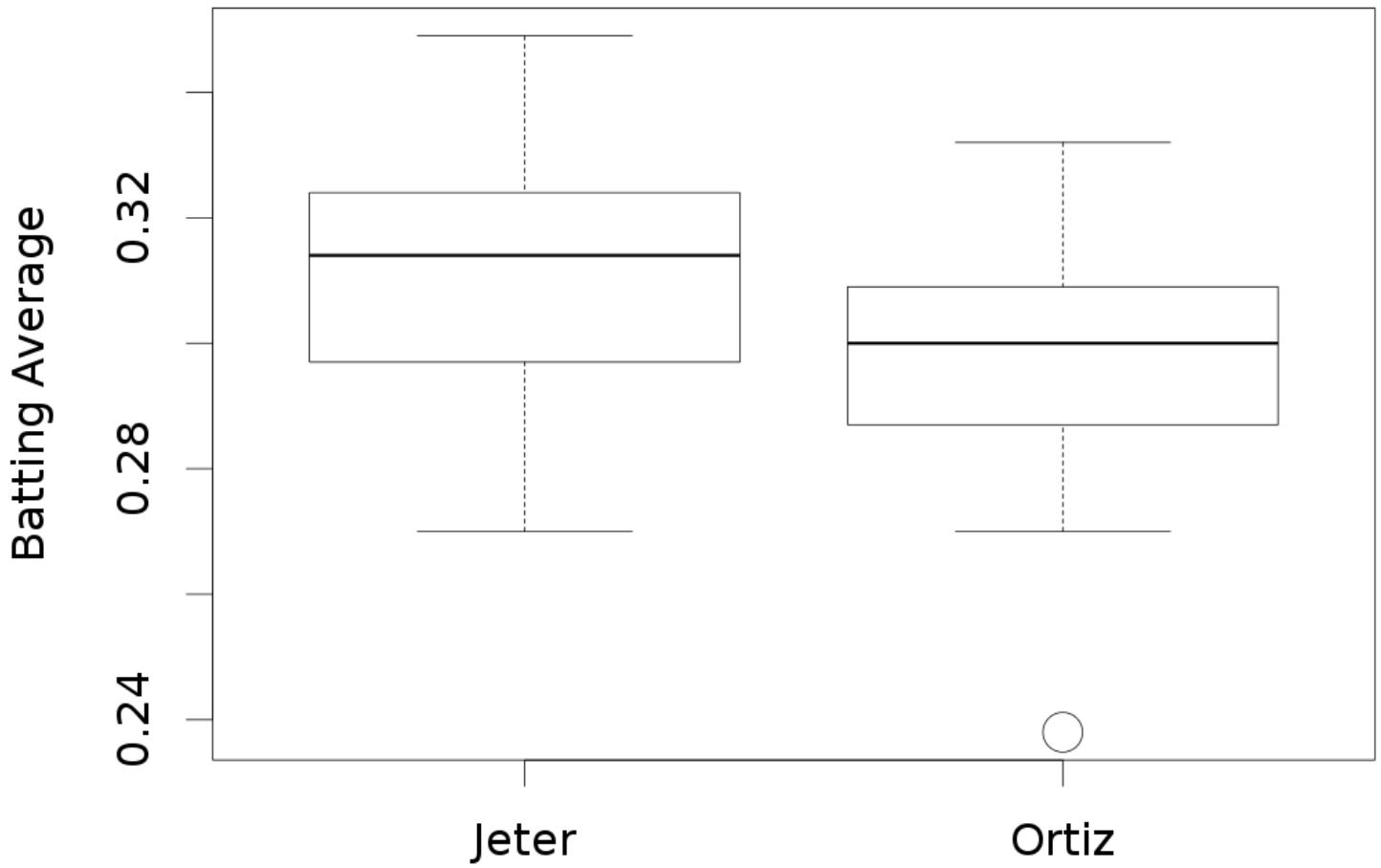
Boston.	T.	R.	Bs.	P.O.	A.	E.	Athletic.	T.	R.	Bs.	P.O.	A.	E.
G. Wright, s. s	6	4	4	1	5	2	Force, s. s.	5	1	2	1	3	2
Leonard, 2 b.	6	3	3	4	4	3	Eggler, c. f.	5	3	3	0	0	0
O'Rourke, 1b	6	2	3	9	0	1	Fisler, r. f.	5	0	1	2	0	0
Murnan, l. f.	6	1	0	3	1	0	Meyerle, 3db	5	1	2	2	3	3
Schafer, 2d b	6	3	3	3	1	2	Sutton, 1st b.	5	1	2	10	0	0
McGinley, c. f	6	0	0	0	0	1	Coons, c.	5	1	0	1	1	3
Manning, r. f.	6	0	2	2	0	0	Hall, l. f.	5	1	3	5	0	0
Morrill, c.	6	2	2	4	1	2	Fowser, 2d b.	6	1	2	6	7	5
Josephs, p.	5	4	4	1	1	2	Knight, p.	5	2	2	0	1	2
Totals.	53	19	21	27	13	13	Totals.	46	11	17	27	15	15

Boston	9	1	3	3	4	1	0	2	5—19
Athletic	1	0	0	0	3	3	2	2	0—11

Runs earned—Boston, 4; Athletic, 5. Home-run—Hall, 1. Total bases on hits—Boson, 22; Athletic, 20. First base by errors—Boston, 8; Athletic, 5. Umpire, George White of Lowell, Mass. Time 2h. 47m.

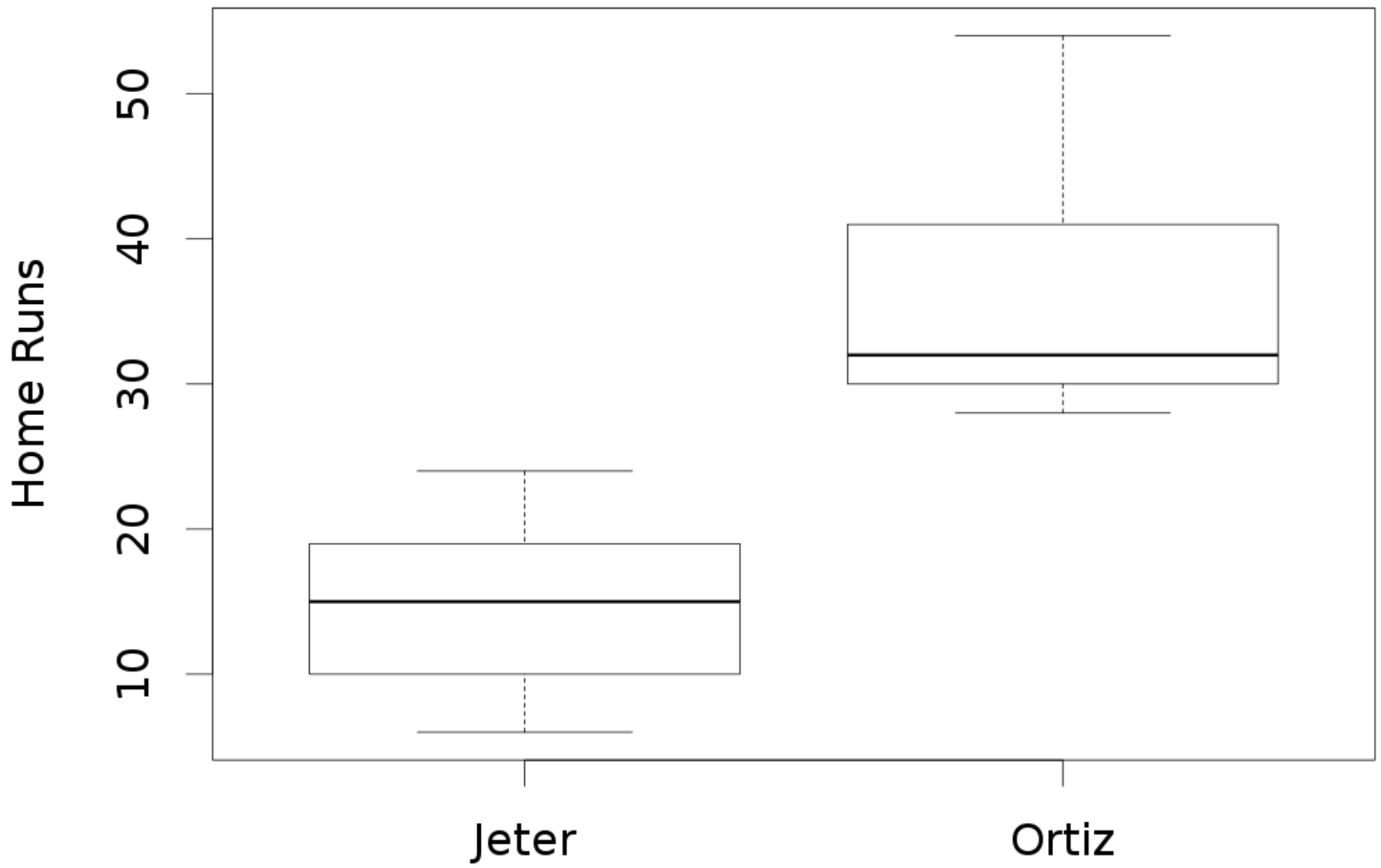
People could now decide who is best player by comparing their statistics!

Who is a better hitter: Derek Jeter or David Ortiz?



Jeter has a better batting average

Who is a better hitter: Derek Jeter or David Ortiz?



Ortiz hits more home runs

Who is a better hitter: Derek Jeter or David Ortiz?

Derek Jeter

David Ortiz

How can we decide on the most important statistic?

Sabermetrics

Sabermetrics: the empirical or mathematical/Statistical study of baseball

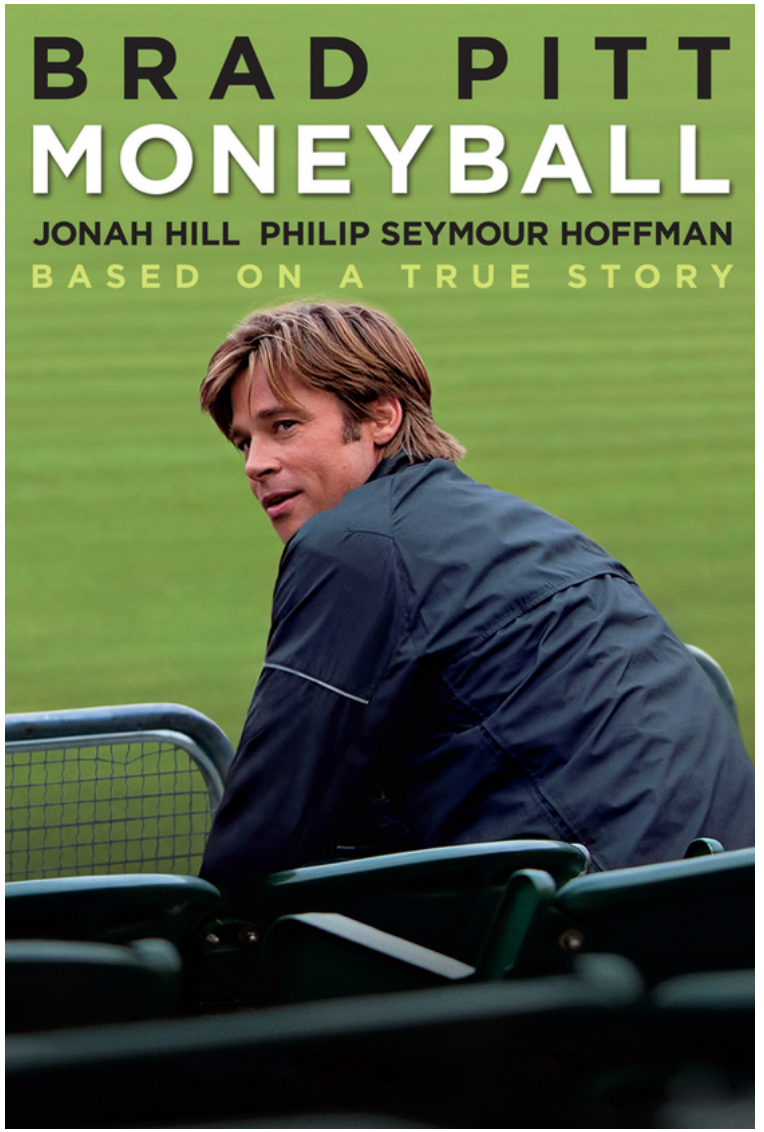
- 'Society for American Baseball Research' (SABR)

Started in the 1970's by Bill James to find more useful measures than classical statistics

- Pre-computers, had to compile all information from old box scores by hand

Billy Bean, the general manager of the A's, used these techniques to create a top ranked team in 2002

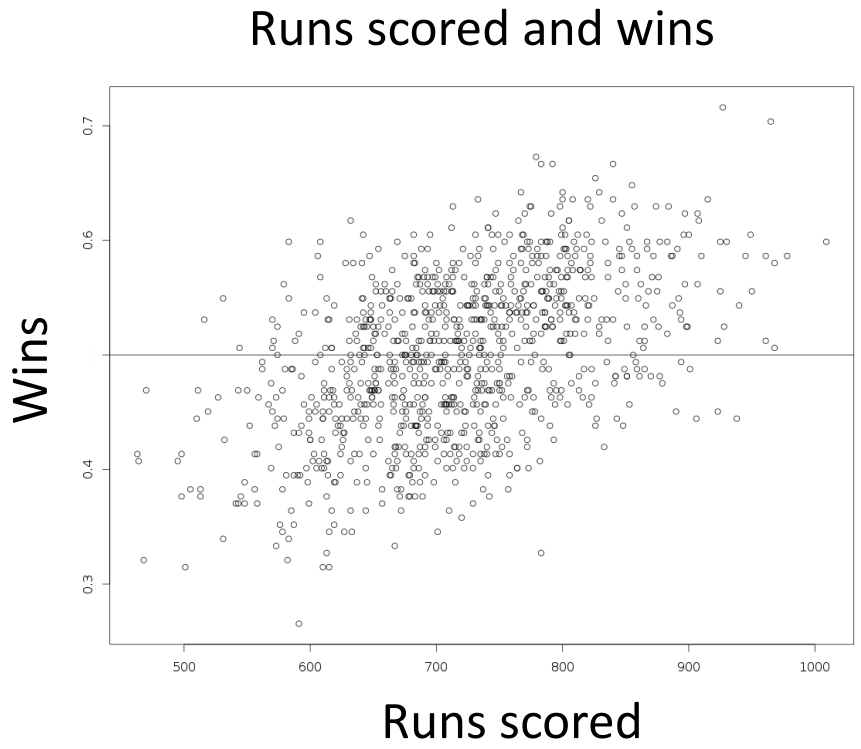
- The book moneyball had a big impact on the expansion of major league clubs doing advanced data analyses



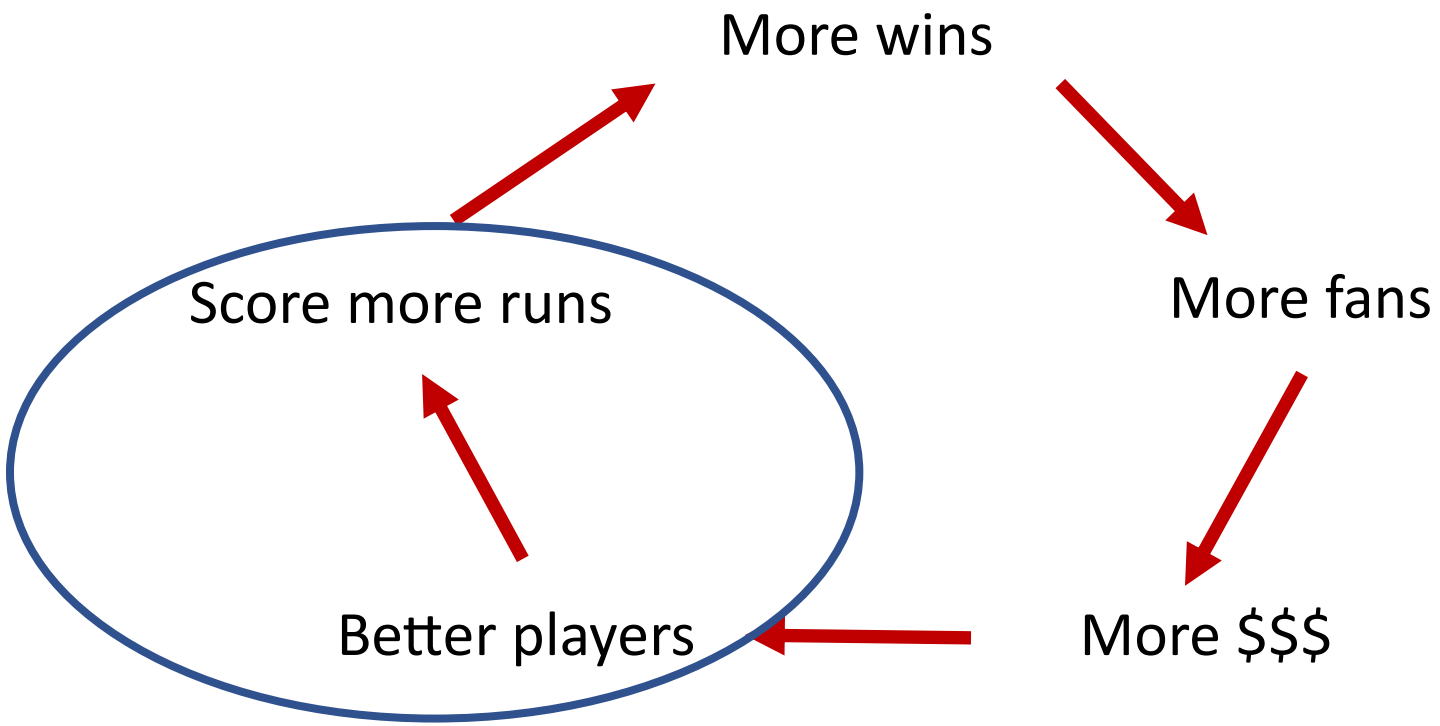
Is power or batting average more important?

It would be good to compare Jeter and Ortiz based on the “best” statistic

How do we determine which statistic is best?



The great cycle of baseball



We can evaluate how ‘good’ a statistic is based on how well it correlates with the number of runs a team scores

What is the best statistic to use?

One idea: the 'best' statistic to judge a player is the statistic that is most correlated with runs

- We can then use this to examine how good a hitter is

Common baseball descriptive statistics are:

H:	Hits: 1B + 2B + 3B + HR
BB:	Walks: 4 balls
PA:	Plate Appearances: Number of times "up"
AB:	At Bats: PA - BB
OBP:	On-Base Percentage: (H + BB)/PA
BA:	Batting Average: H/AB
SlugPct:	Slugging percentage: (1·1B + 2·2B + 3·3B + 4·HR)/AB

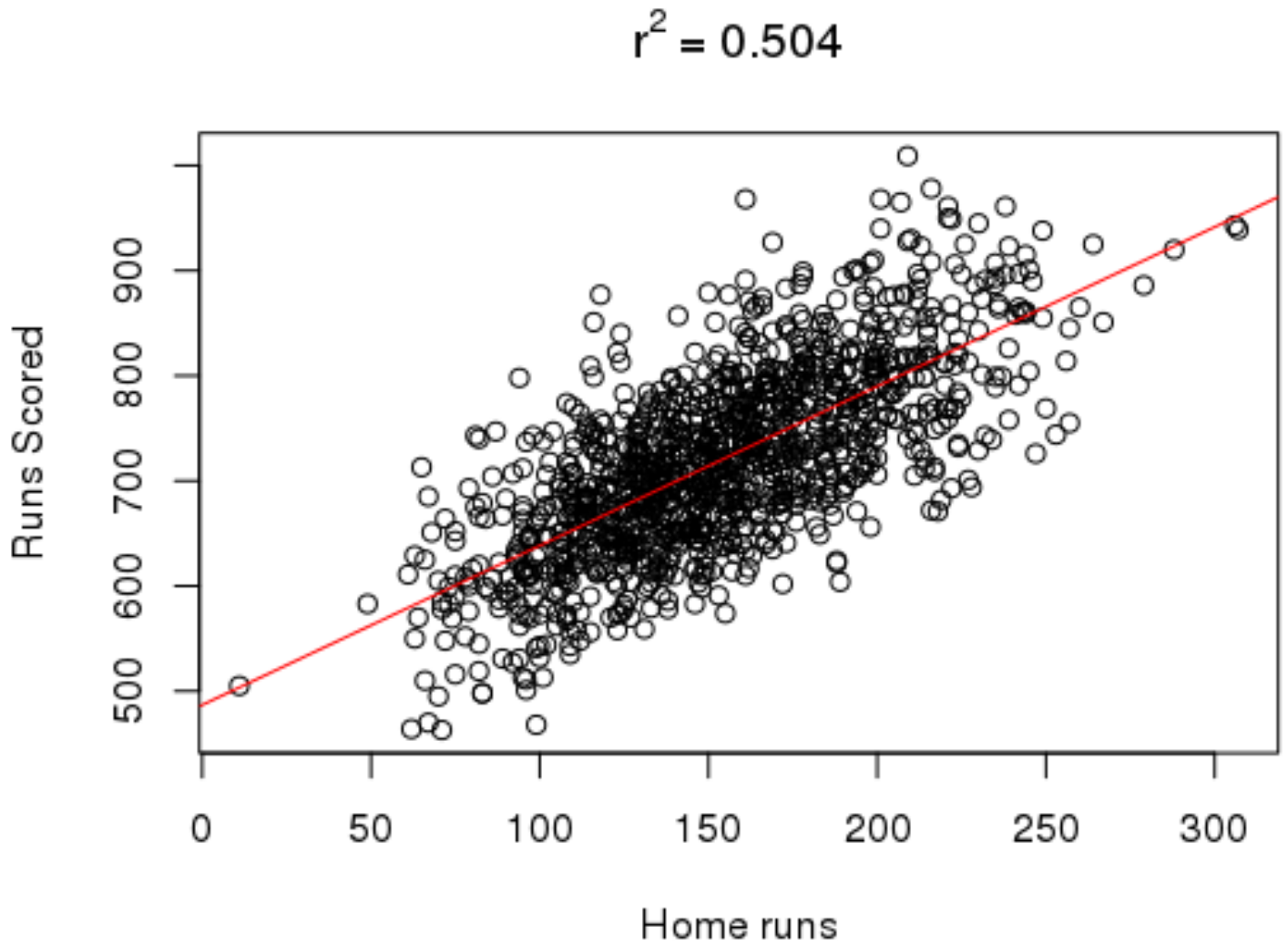
Data exploration in R…

Let's see which statistic is has the highest coefficient of determination (r^2) with how many runs a team scored

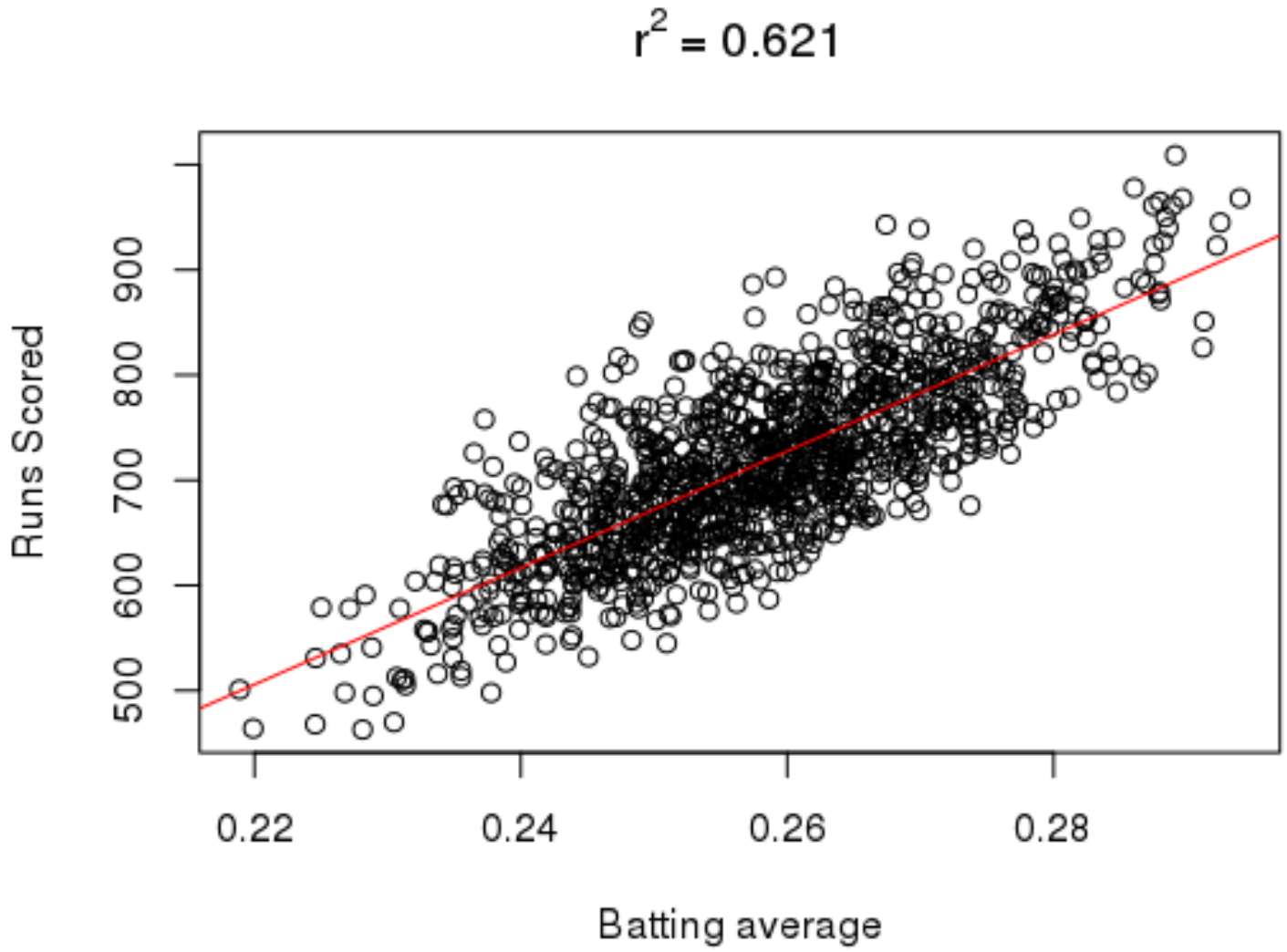
- Using team level data from all baseball seasons with 162 games (i.e., data since 1961)

Data is available in the Lahman package

Correlation between HR and runs

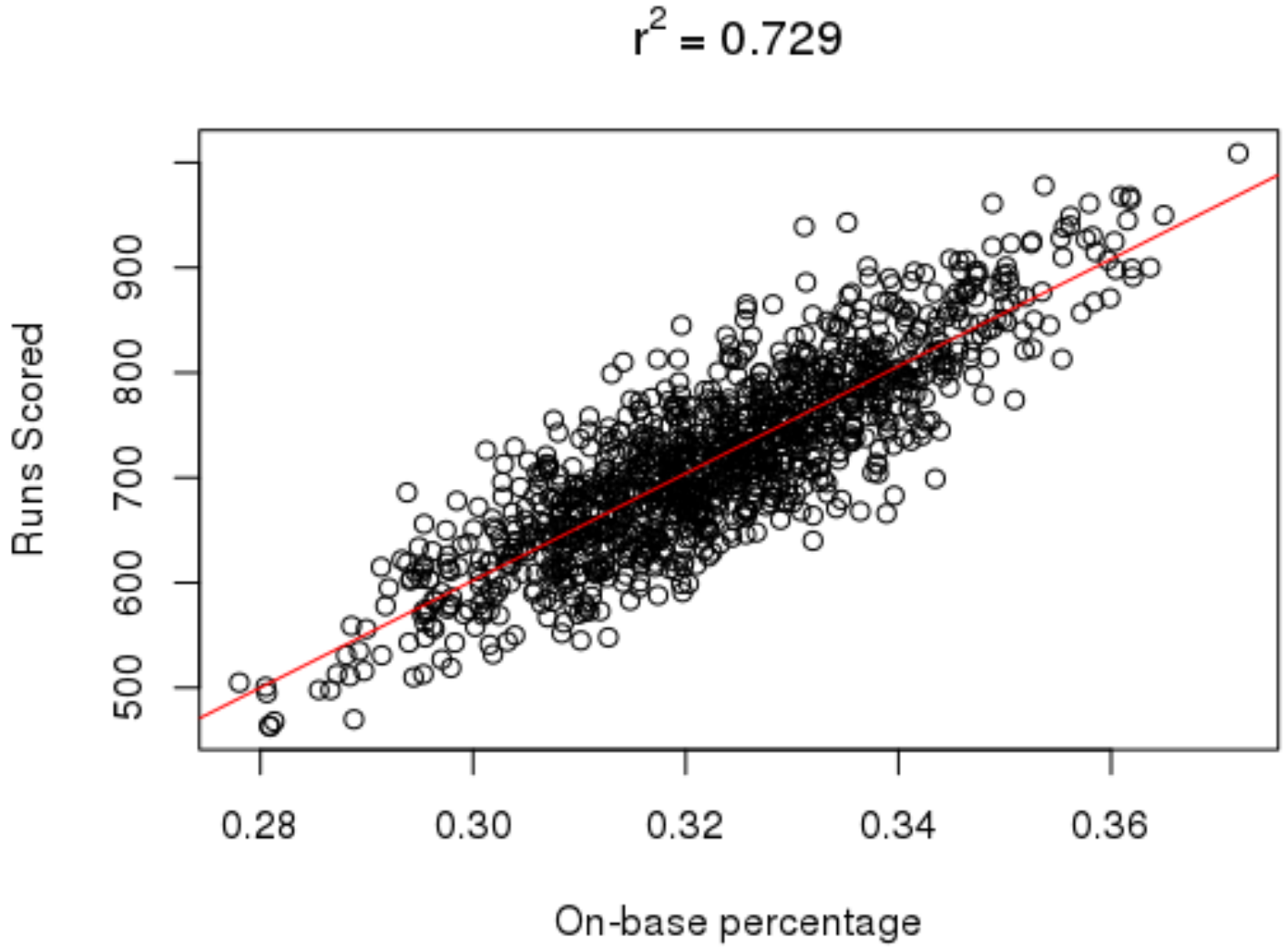


Correlation between BA and runs



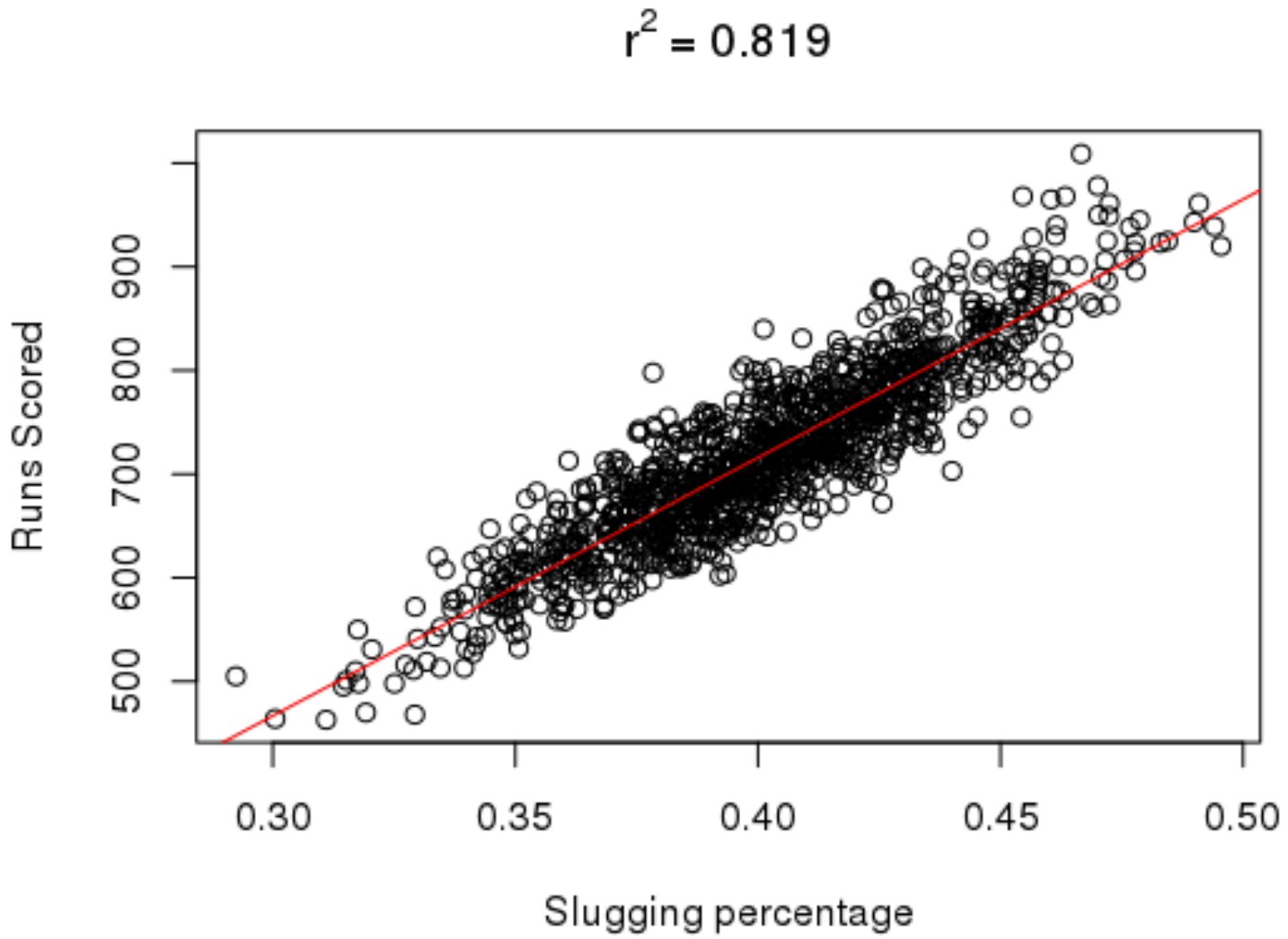
BA = (1B + 2B + 3B + HR)/AB

Correlation between OBP and runs



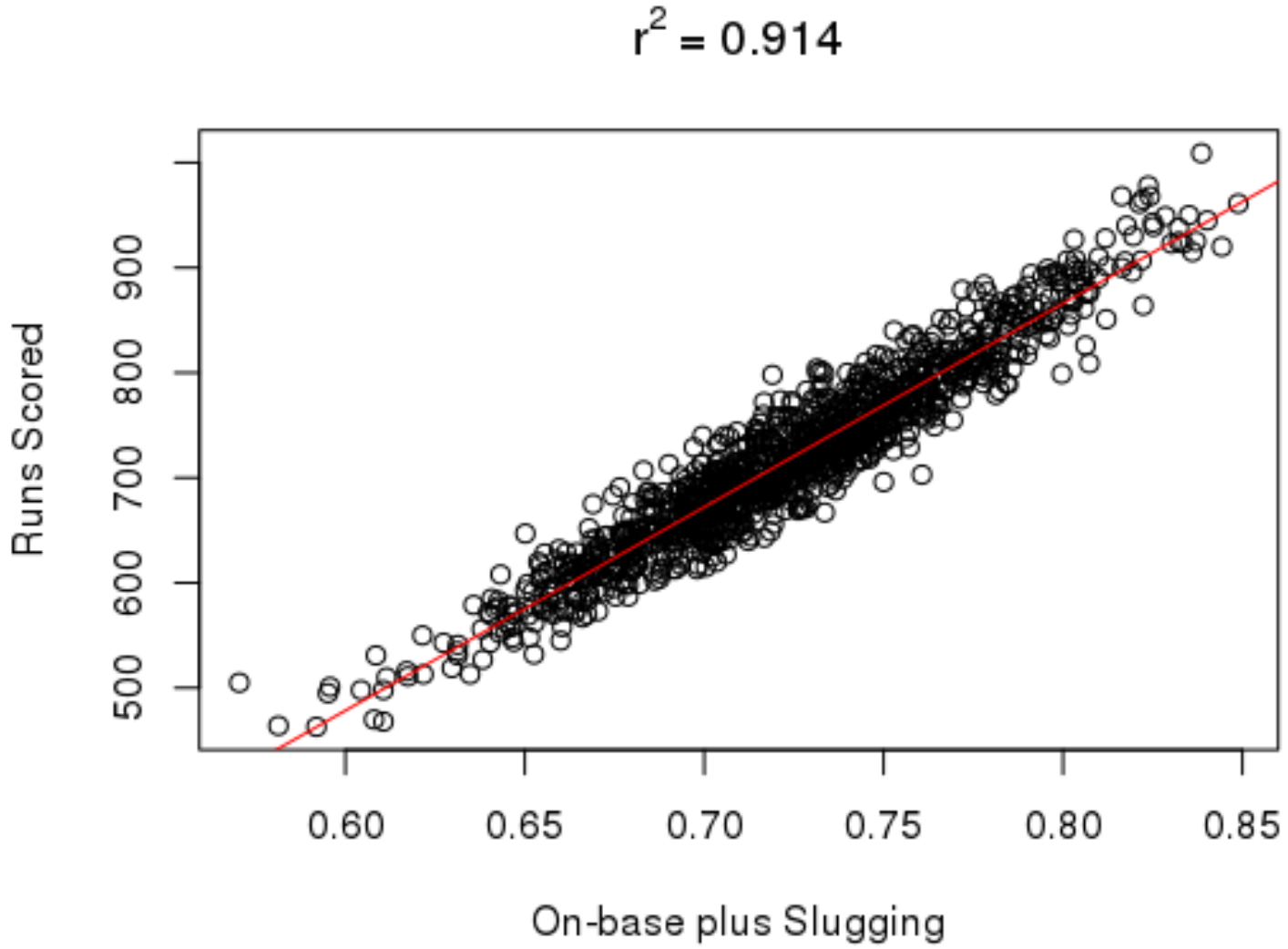
OBP = (BB + HBP + 1B + 2B + 3B + HR)/PA

Correlation between Slug and runs



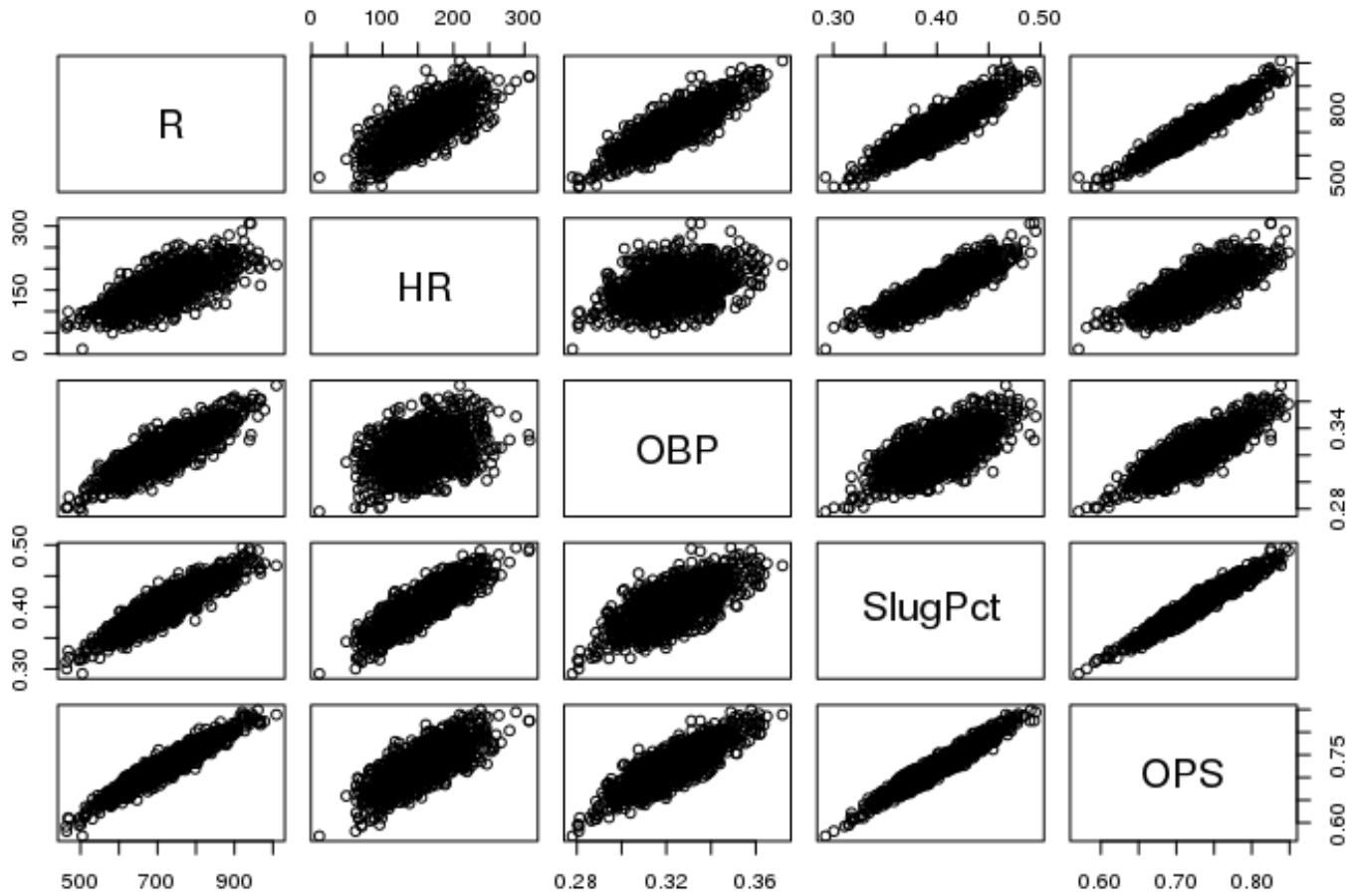
Slug = $(1 \cdot 1B + 2 \cdot 2B + 3 \cdot 3B + 4 \cdot HR)/AB$

Correlation between OPS and runs



OPS = OBP + Slug

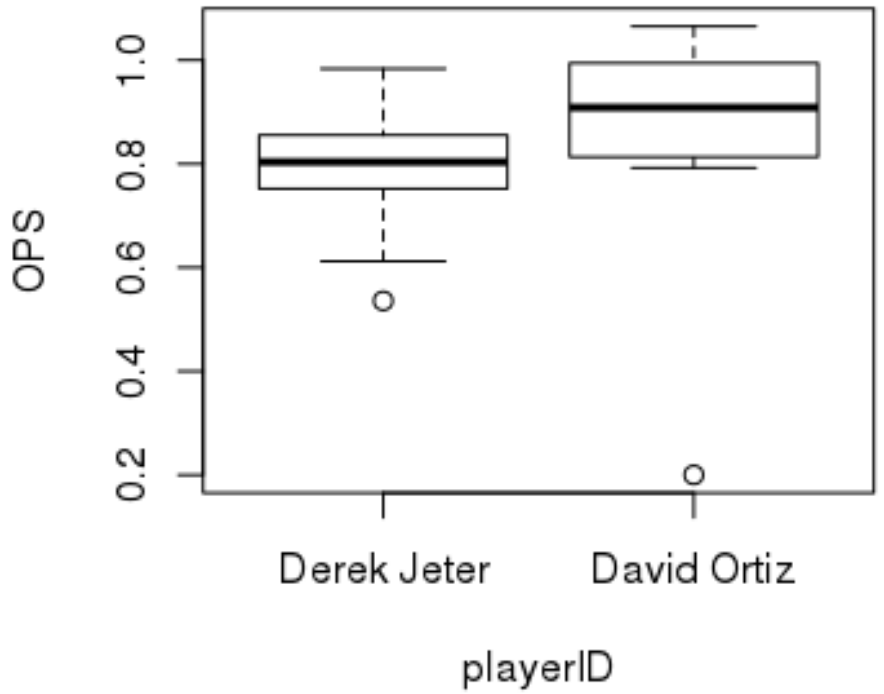
Correlation between all variables



R: pairs()

What is the best statistic to use?

It seems like the winner is on-base plus slugging percentage!



Better know a player: Derek Jeter

[Onion infographic]

[Other Onion articles]

Creating better ‘metrics’

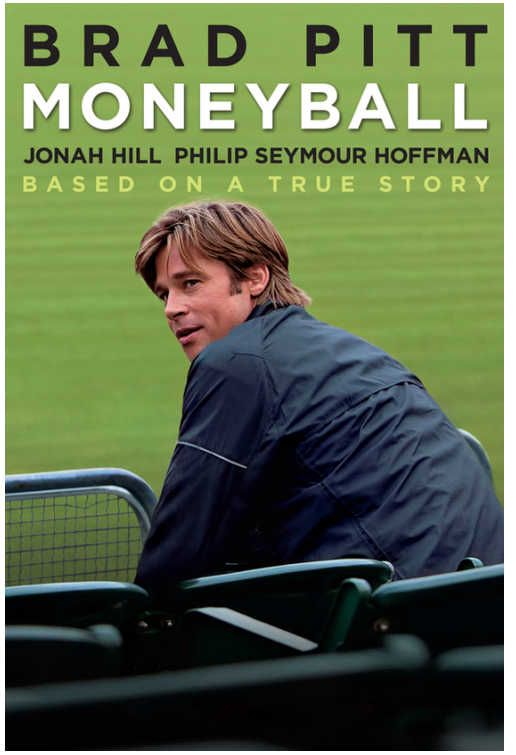
Slugging percentage seemed like the best statistics for predicting runs scored we have found so far

But who says we can’t do better!

Creating better ‘metrics’

Henry Chadwick (1824-1908)

Sabermetrics



Creating better ‘metrics’

Batting average:

$$BA = [(1) \cdot \mathbf{1B} + (1) \cdot \mathbf{2B} + (1) \cdot \mathbf{3B} + (1) \cdot \mathbf{HR}]/\mathbf{AB}$$

Slugging percentage:

$$Slug = [(1) \cdot \mathbf{1B} + (2) \cdot \mathbf{2B} + (3) \cdot \mathbf{3B} + (4) \cdot \mathbf{HR}]/\mathbf{AB}$$

On-base percentage:

$$OBP = [(1) \cdot \mathbf{BB} + (1) \cdot \mathbf{HBP} + (1) \cdot \mathbf{1B} + (1) \cdot \mathbf{2B} + (1) \cdot \mathbf{3B} + (1) \cdot \mathbf{HR}]/\mathbf{PA}$$

Optimal statistic:

$$OPT = b_1 \cdot \mathbf{BB} + b_2 \cdot \mathbf{HBP} + b_3 \cdot \mathbf{1B} + b_4 \cdot \mathbf{2B} + b_5 \cdot \mathbf{3B} + b_6 \cdot \mathbf{HR} + b_0$$

We want to find the “best” b_i‘s for predicting how many runs a team scored

What are the optimal weights?

Any ideas for the best b_i's ?

$$\mathbf{OPT} = b_1 \cdot \mathbf{BB} + b_2 \cdot \mathbf{HBP} + b_3 \cdot \mathbf{1B} + b_4 \cdot \mathbf{2B} + b_5 \cdot \mathbf{3B} + b_6 \cdot \mathbf{HR} + b_0$$

We can use multiple regression to find the b_i's that minimize sum of $(R - OPT)^2$

What are the optimal weights?

Any ideas for the best b_i's ?

$$\mathbf{OPT} = b_1 \cdot \mathbf{BB} + b_2 \cdot \mathbf{HBP} + b_3 \cdot \mathbf{1B} + b_4 \cdot \mathbf{2B} + b_5 \cdot \mathbf{3B} + b_6 \cdot \mathbf{HR} + b_0$$

Let's use multiple regression to find the b_i's that minimize sum of $(R - OPT)^2$

```
> fit <- lm(R ~ BB + HBP + H + X2B + X3B + HR, data = team_batting)

> coef(fit)
```

What are the optimal weights?

	b_i
(Intercept)	-467
HBP	0.28
BB	0.36
X1B	0.54
X2B	0.68
X3B	1.35
HR	1.44

> coef(fit)

Do these coefficients make sense?

Can you write this in the form of an equation?

What are the optimal weights?

	b_i
(Intercept)	-467
HBP	0.28
BB	0.36
X1B	0.54
X2B	0.68
X3B	1.35
HR	1.44

> coef(fit)

Do these coefficients make sense?

$$\hat{y} = .36 \cdot \text{BB} + .28 \cdot \text{HBP} + .54 \cdot \text{1B} + .68 \cdot \text{2B} + 1.35 \cdot \text{3B} + 1.44 \cdot \text{HR} - 467$$

What is the coefficient of determination for this model?

	b_i
(Intercept)	-467
HBP	0.28
BB	0.36
X1B	0.54
X2B	0.68
X3B	1.35
HR	1.44

OPT model: $R^2 = 0.929$

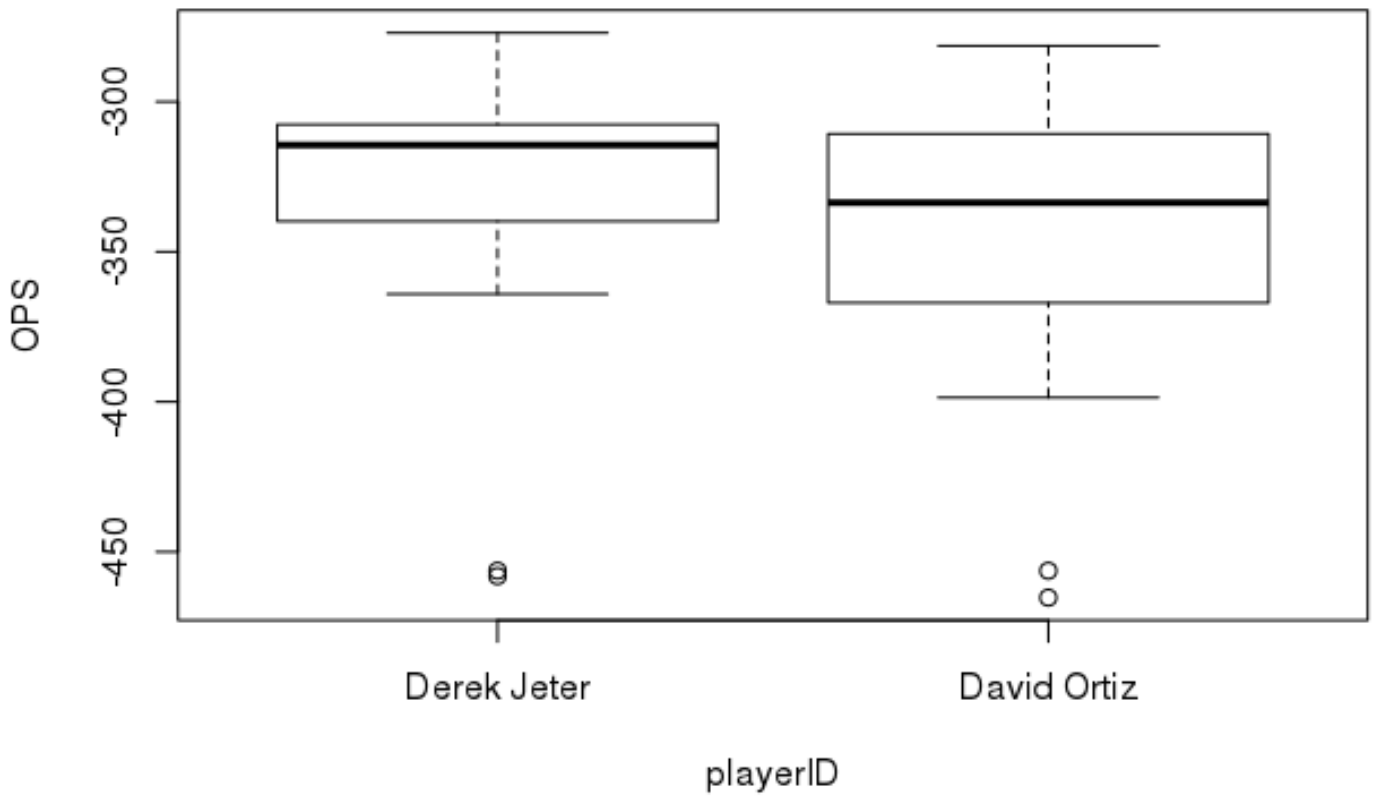
Previous best model

OPS: $R^2 = 0.921$

Our OPT statistic seems better!

$$\hat{y} = .36 \cdot BB + .28 \cdot HBP + .54 \cdot 1B + .68 \cdot 2B + 1.35 \cdot 3B + 1.44 \cdot HR - 467$$

How do Derek Jeter or David Ortiz compare on our new statistic?



Can we do even better?

Can we create additional variables in the team_batting data frame that will lead to a statistic with even higher R^2 ?

Let's try it in R!