Correlation continued and simple linear regression

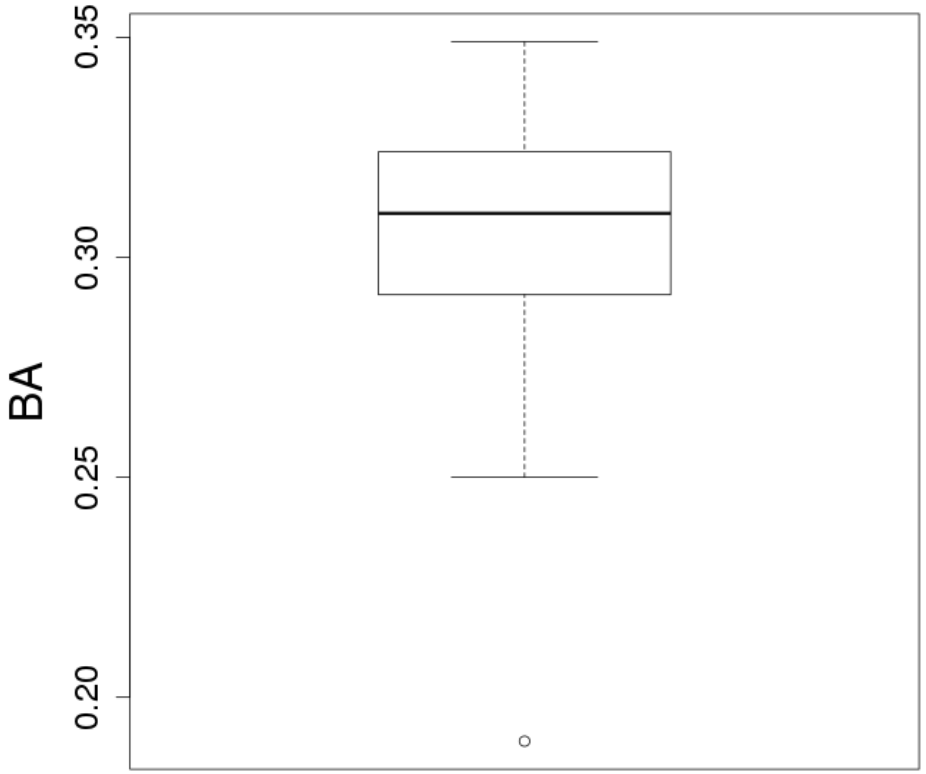
Outline for today

Better know a player: Mark Prior

Review and continuation of correlation

Simple linear regression!

Worksheet 3: Jeter BA boxplot



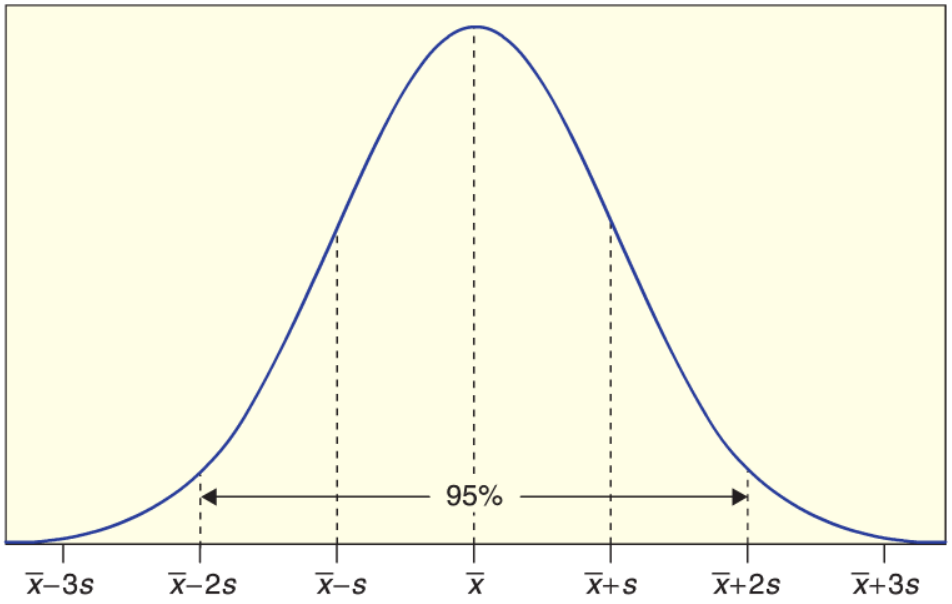
What is the reason for the outlier?

Review

Worksheet 3: interpreting z-scores

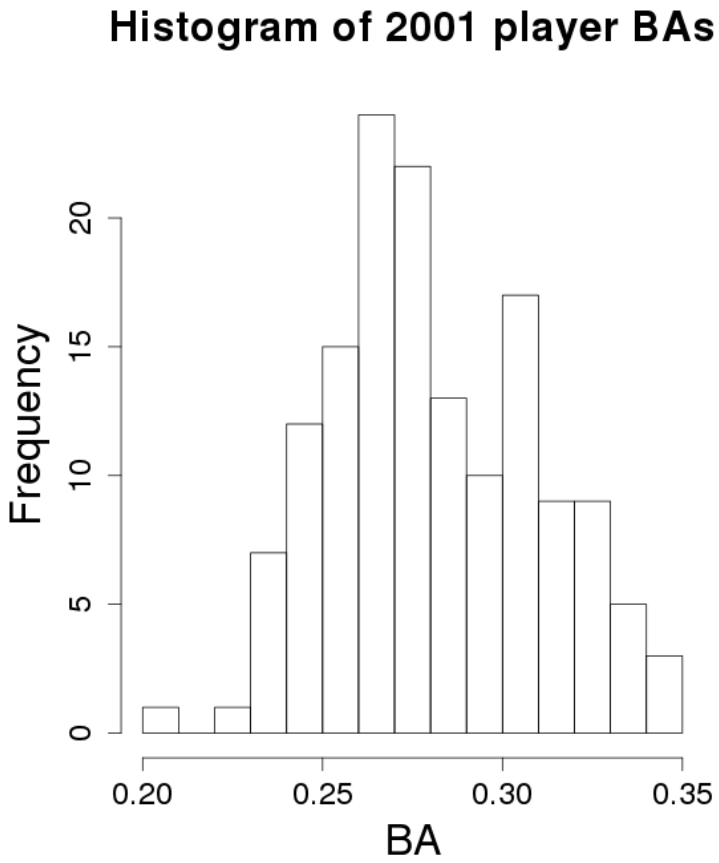
BA z-scores

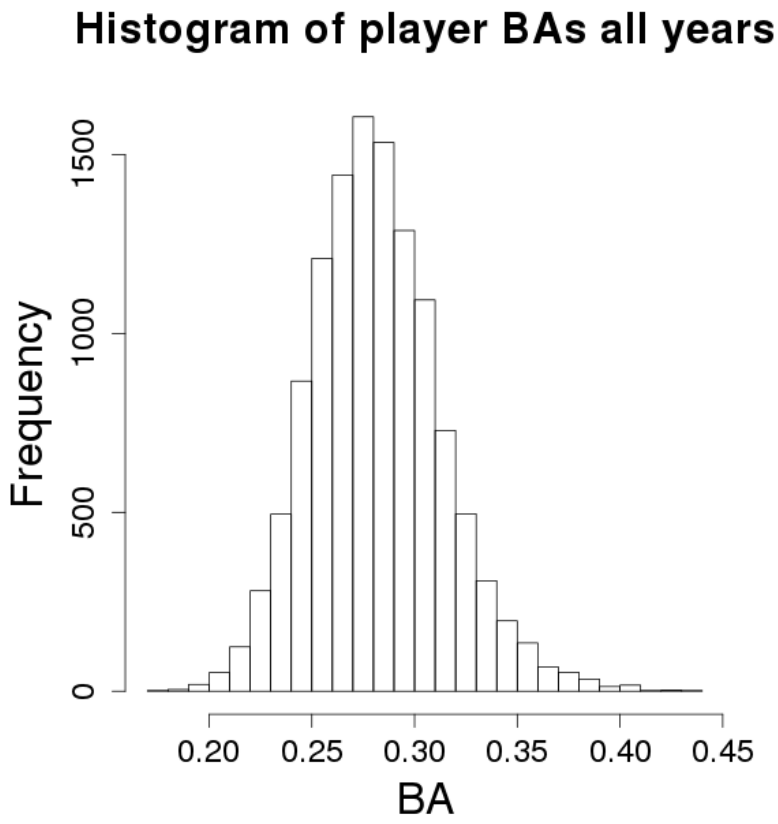
Jeter:	1.02
Ruth:	2.168
Gehrig:	-0.389
Mantle:	2.70
Mattingly:	1.46



How do we interpret what a good z-score is?

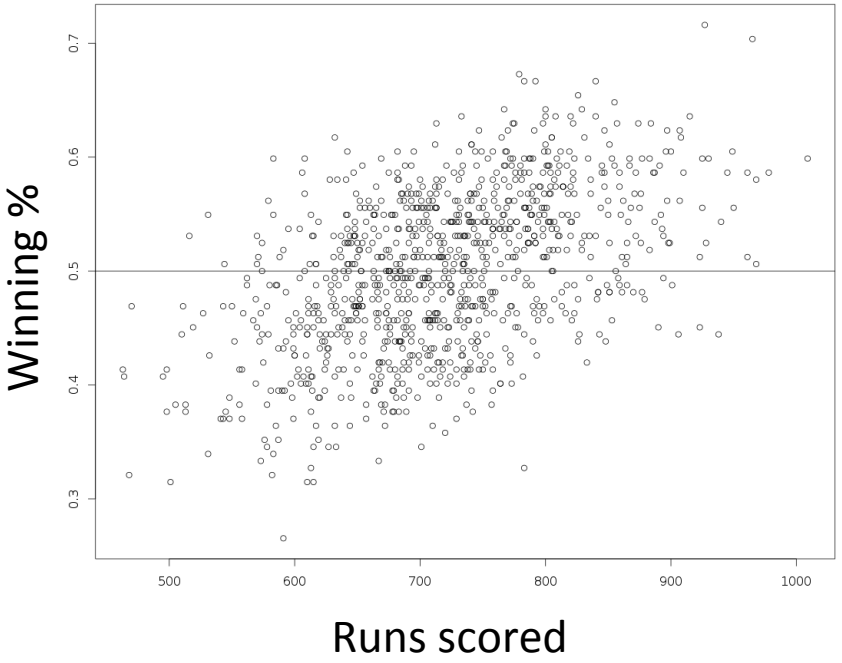
Are BAs normally distributed?





Scatter plots

A **scatterplot** graphs the relationship between two variables



If there is an explanatory and response variable, then the explanatory variable is put on the x-axis and the response variable is put on the y-axis

```
R: plot(x, y)
```

Correlation

The **correlation** is measure of the strength and direction of a linear association between two variables.

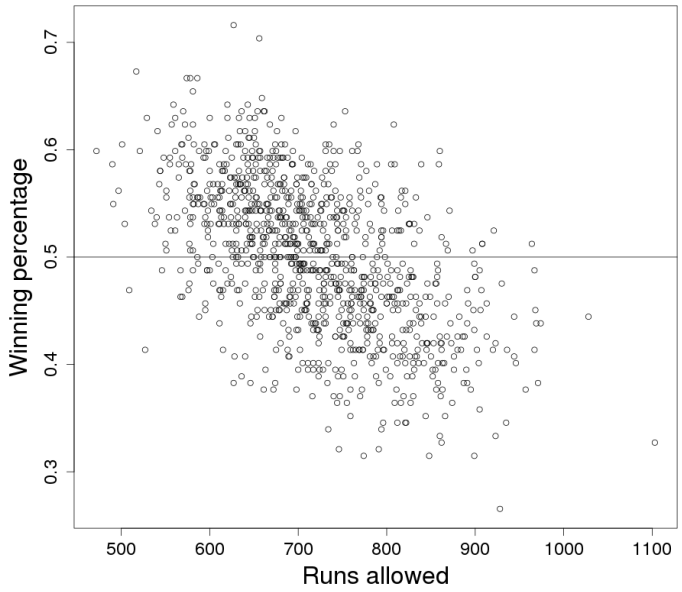
$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

```
R: cor(x, y)
```

Correlation examples

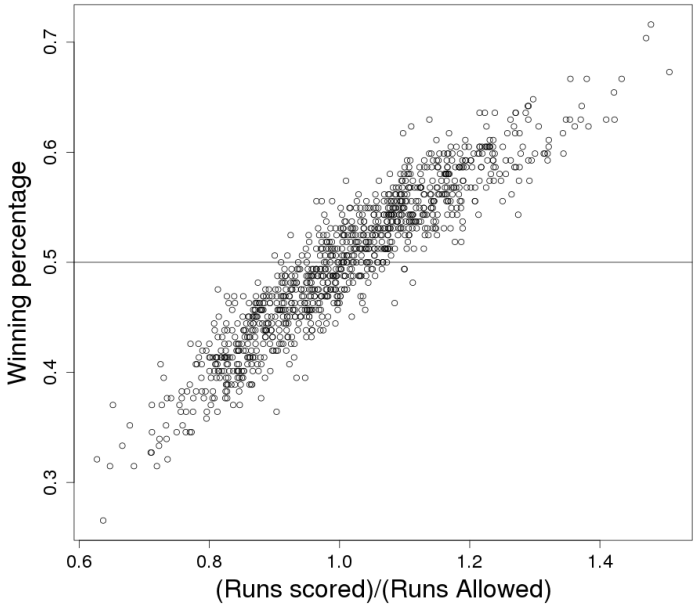
Correlation Examples

Runs allowed and wins



r = -.55

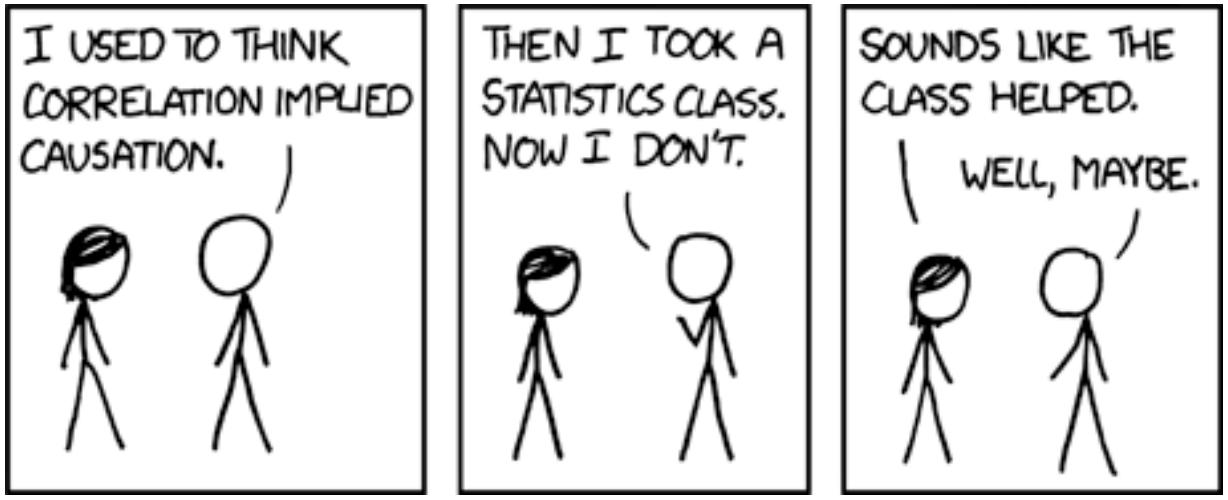
(runs scored)/(runs allowed) and wins



r = .93

Correlation cautions

1. A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables



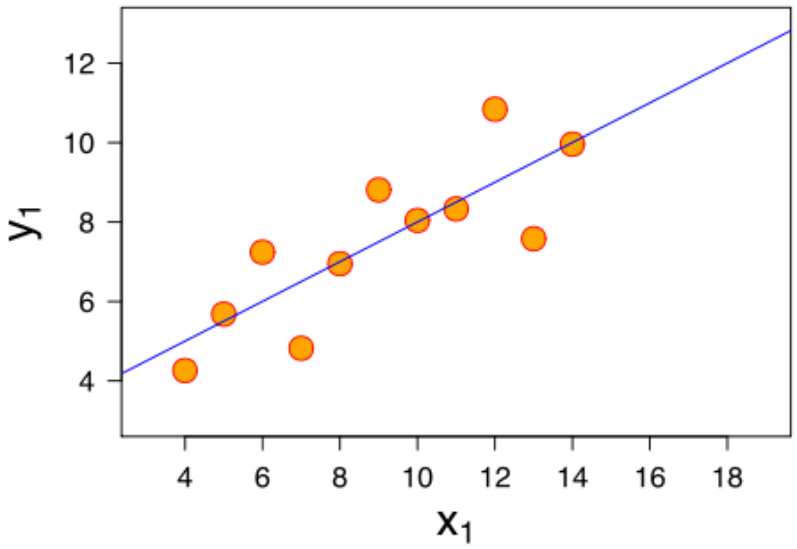
Correlation cautions

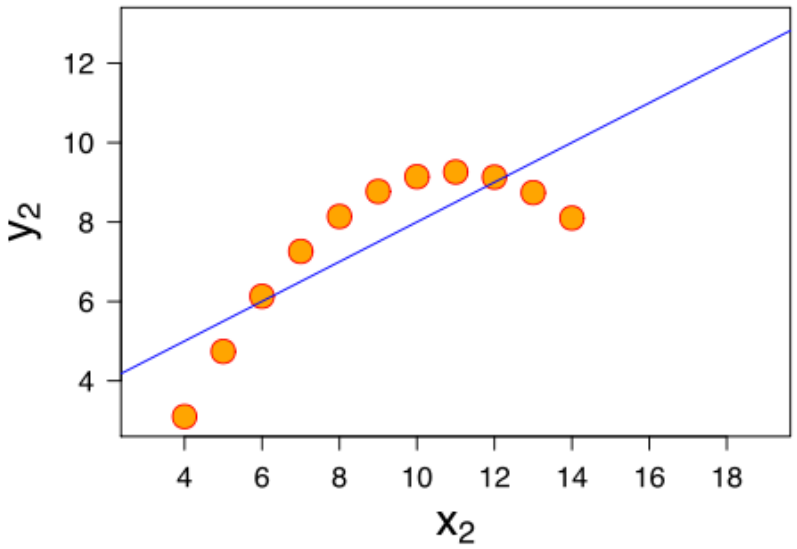
1. A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables

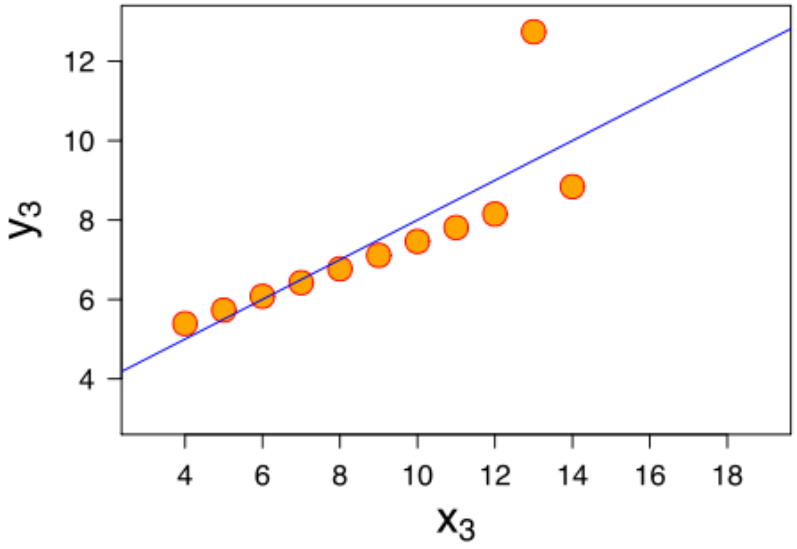
2. A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship.

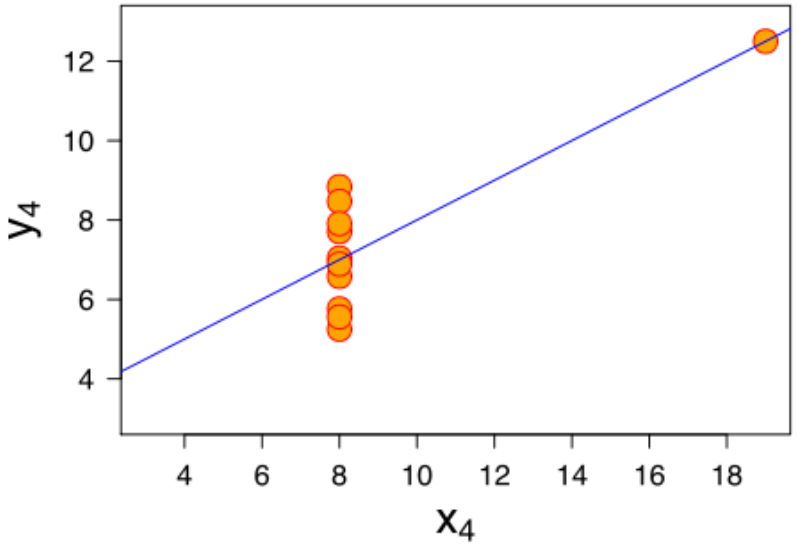
3. Correlation can be heavily influences by outliers. Always plot your data!

Anscombe's quartet (r = 0.81)









Offensive statistics

What do the following abbreviations stand for?

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

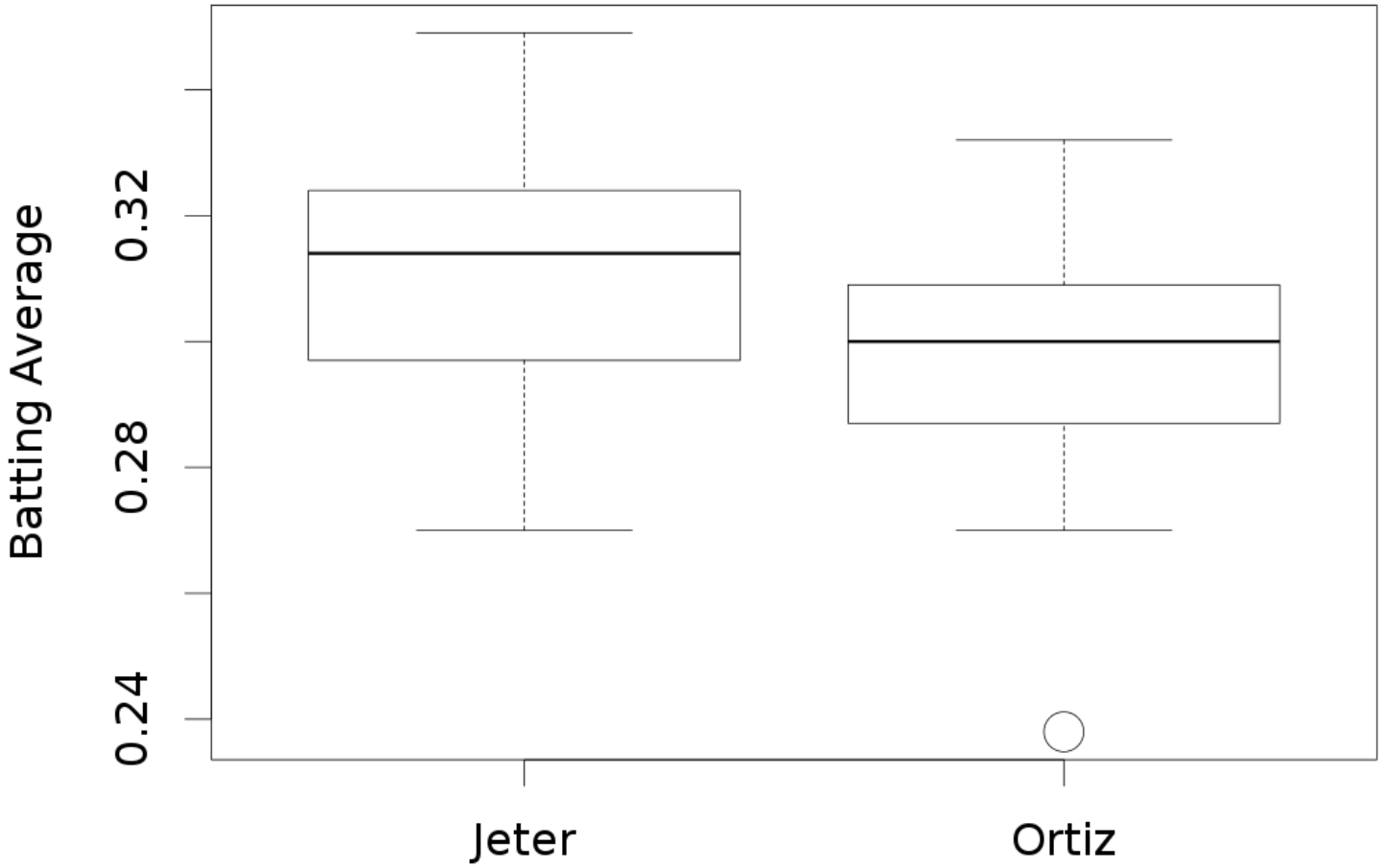
Who is a better hitter: Derek Jeter or David Ortiz?

Derek Jeter

David Ortiz

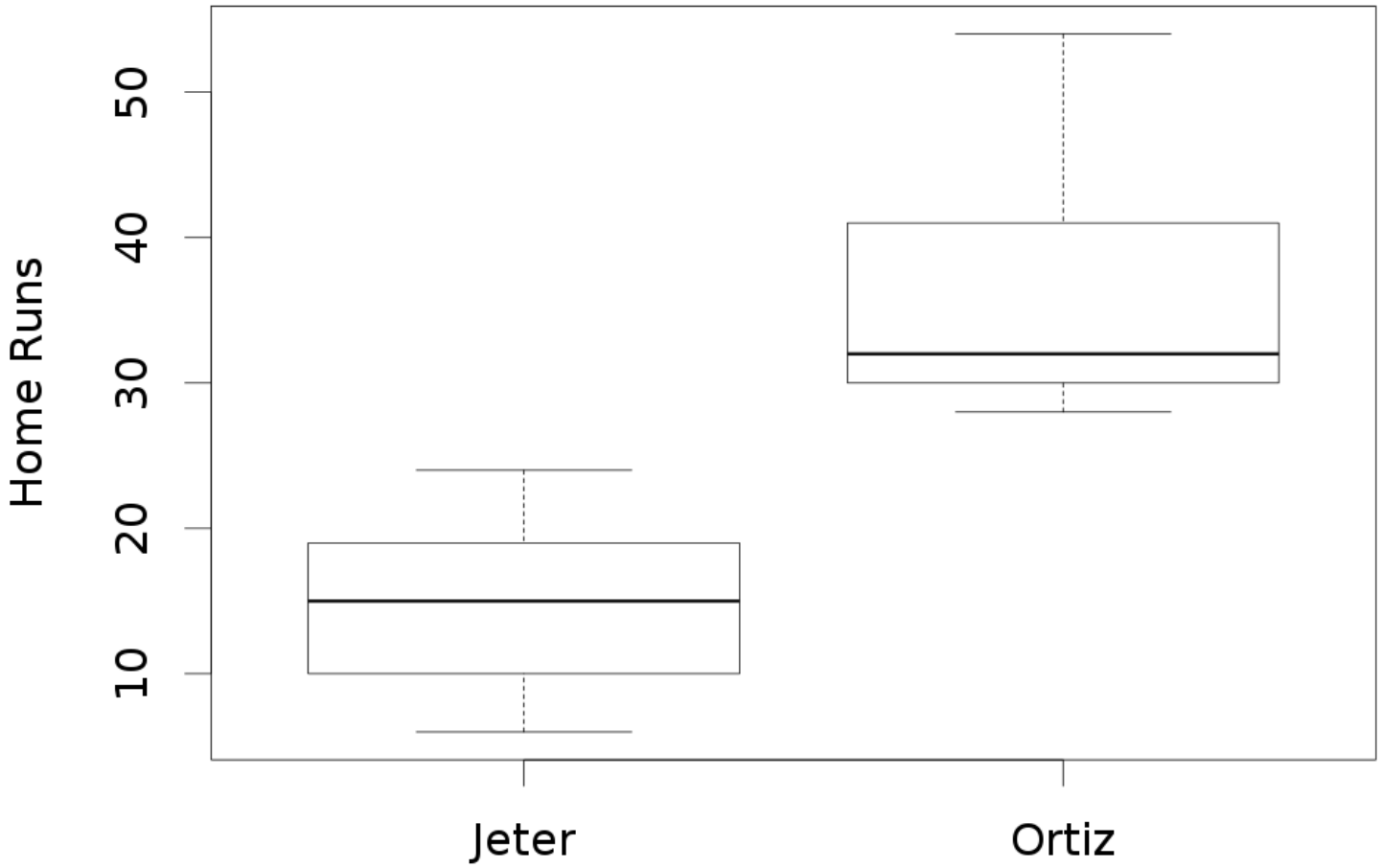
Who would you rather have on your team?

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?



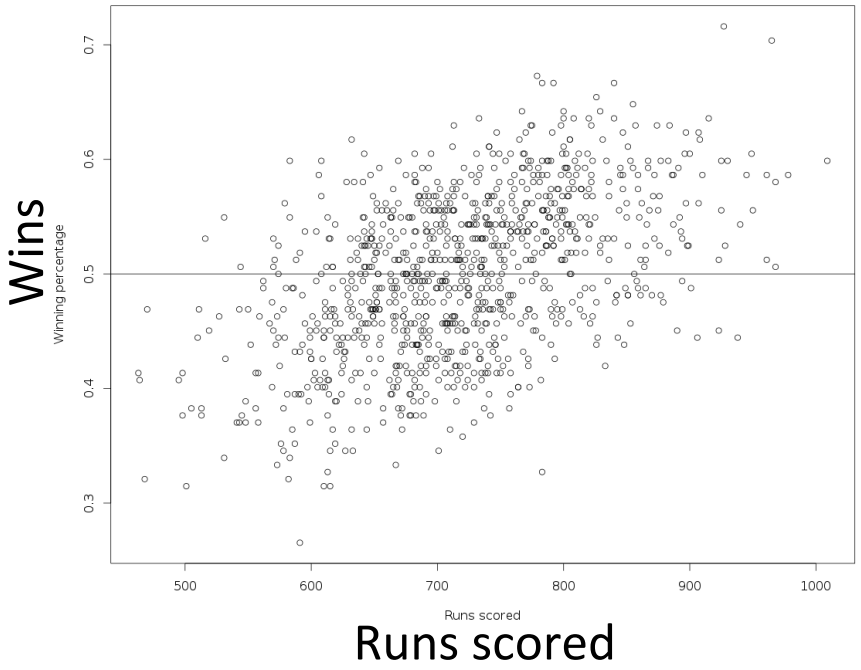
Ortiz hits more home runs

Is power or batting average more important?

Compare them based on the “best” statistic

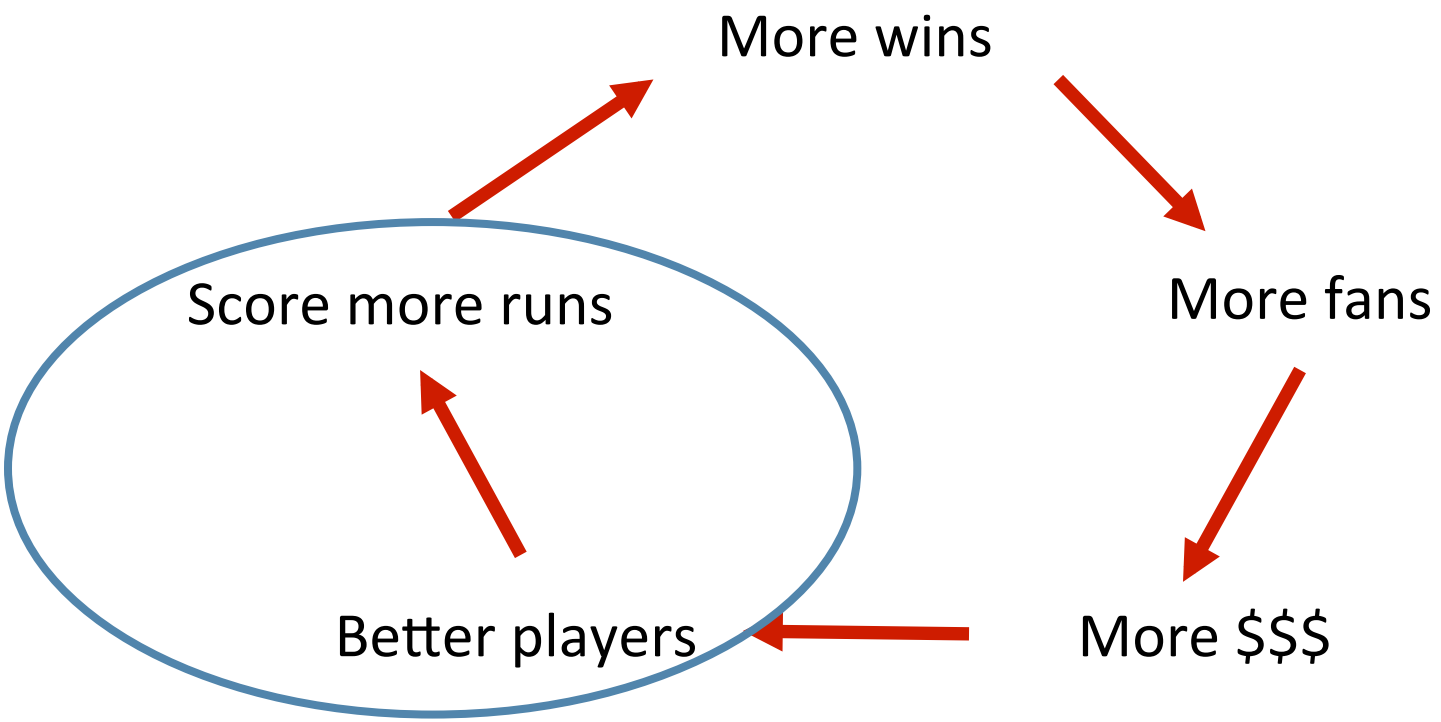
How do we determine which statistic is best?

Runs scored and wins



Runs scored

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

We will use a data set that has season total statistics going back to 1961

These statistics include the total runs a team scored, total team HR, total team BA, etc.

```
load('/home/shared/baseball_stats_2007/
data/team_batting_stats.Rda')
```

What is the best statistic to use?

Statistics to compare:

1. Home runs (HR)
2. Batting average (BA)
3. On-base percentage (OBP)
4. Slugging percentage (Slug)
5. On-base percentage + slugging percentage (OPS)

For each of these 5 statistics:

- Create a scatter plot between the statistic and runs (R)
- Calculate the correlation between the statistic and runs (R)

Once you have found the statistic that is more correlated with runs, create a side-by-side box plot to compare Derek Jetter and David Ortiz's data on this statistic

You can get the team yearly total statistics run:

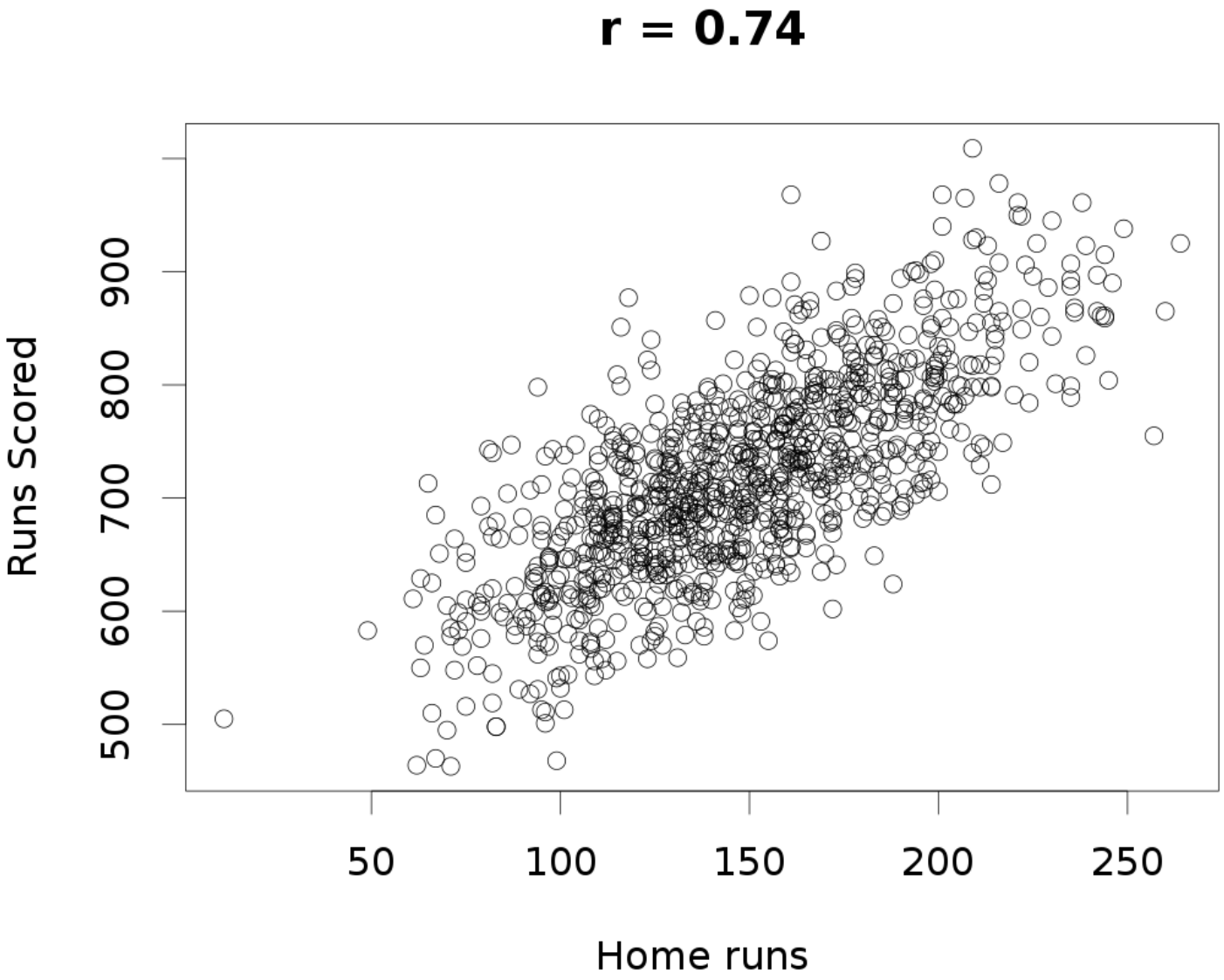
```
load('/home/shared/baseball_stats_2017/team_batting_stats.Rda')
```

Useful functions:

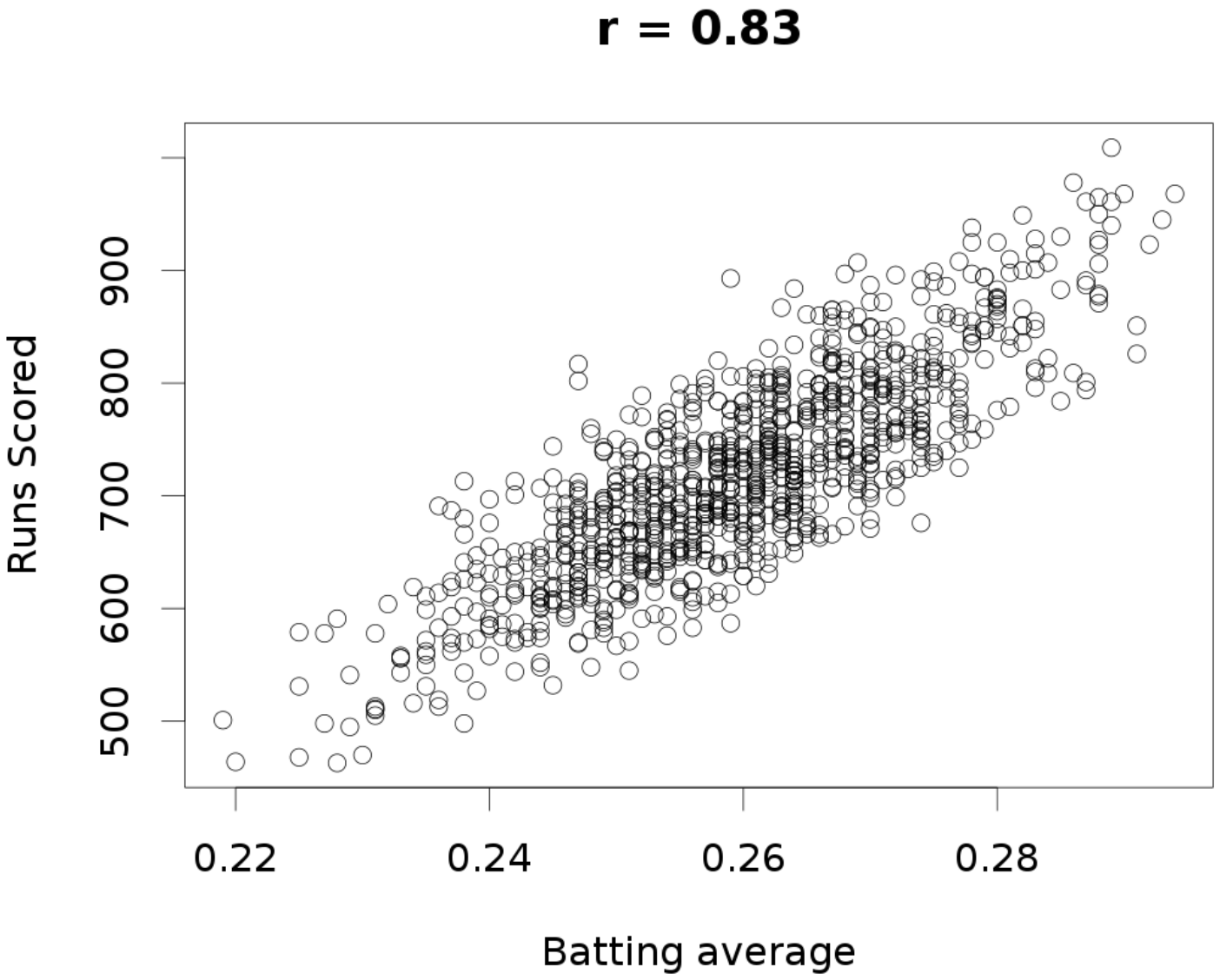
- plot(x, y) # create a scatter plot of different statistics and runs
- cor(x, y) # calculate the correlation between different statistics and runs
- boxplot(v1, v2, names = c('Derek', 'David')) # compare players on this 'best' statistic

Results…

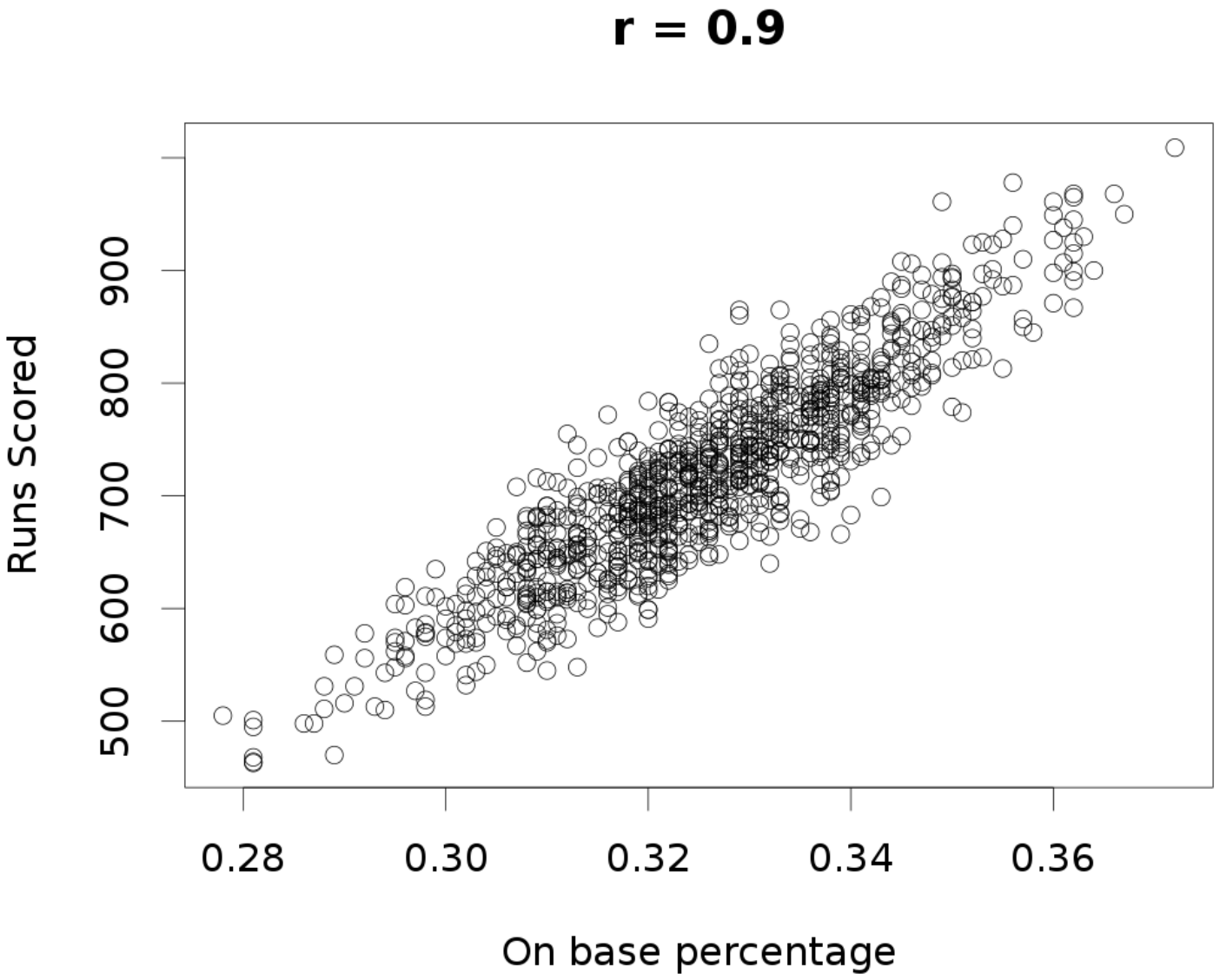
Correlation between HR and runs



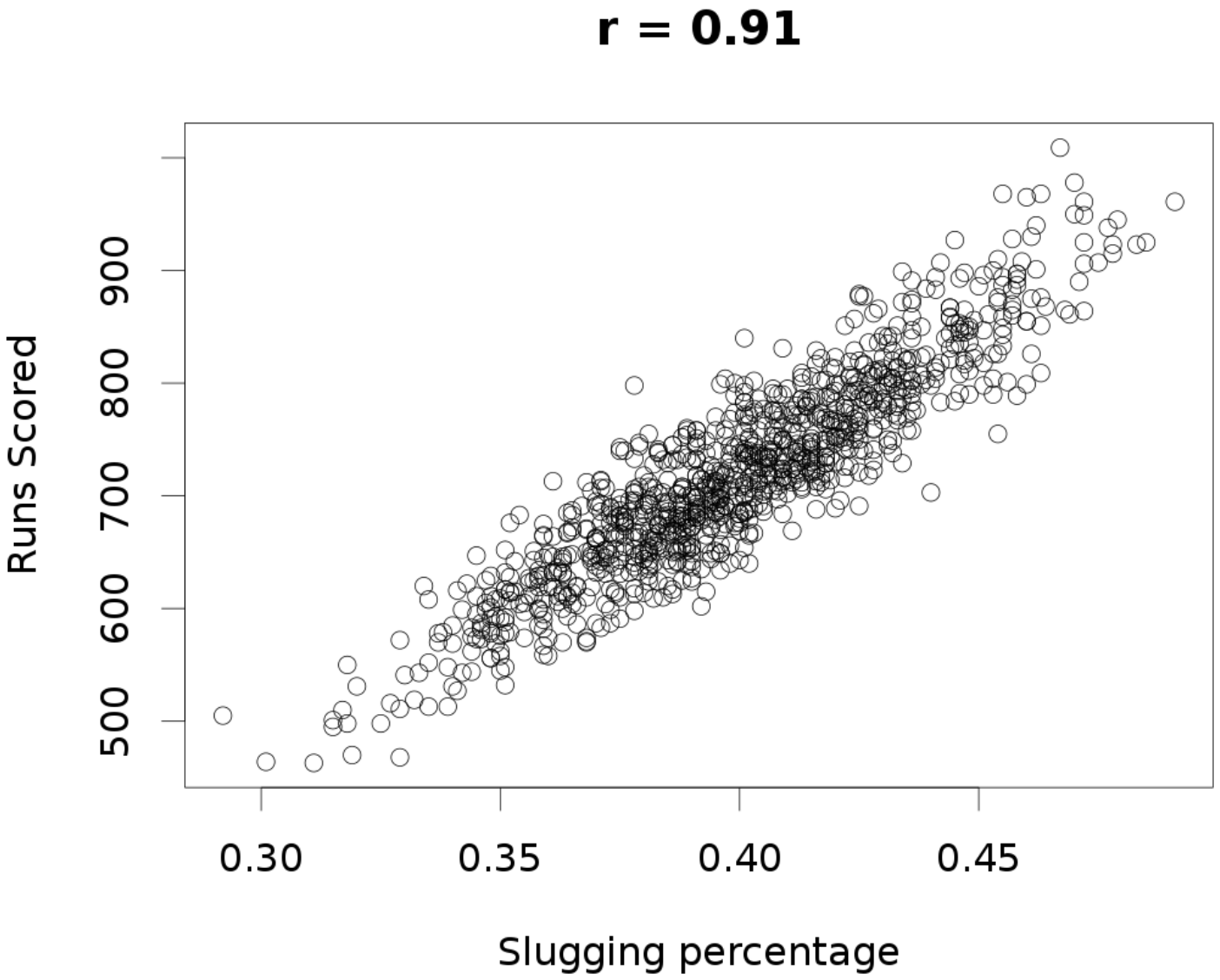
Correlation between BA and runs



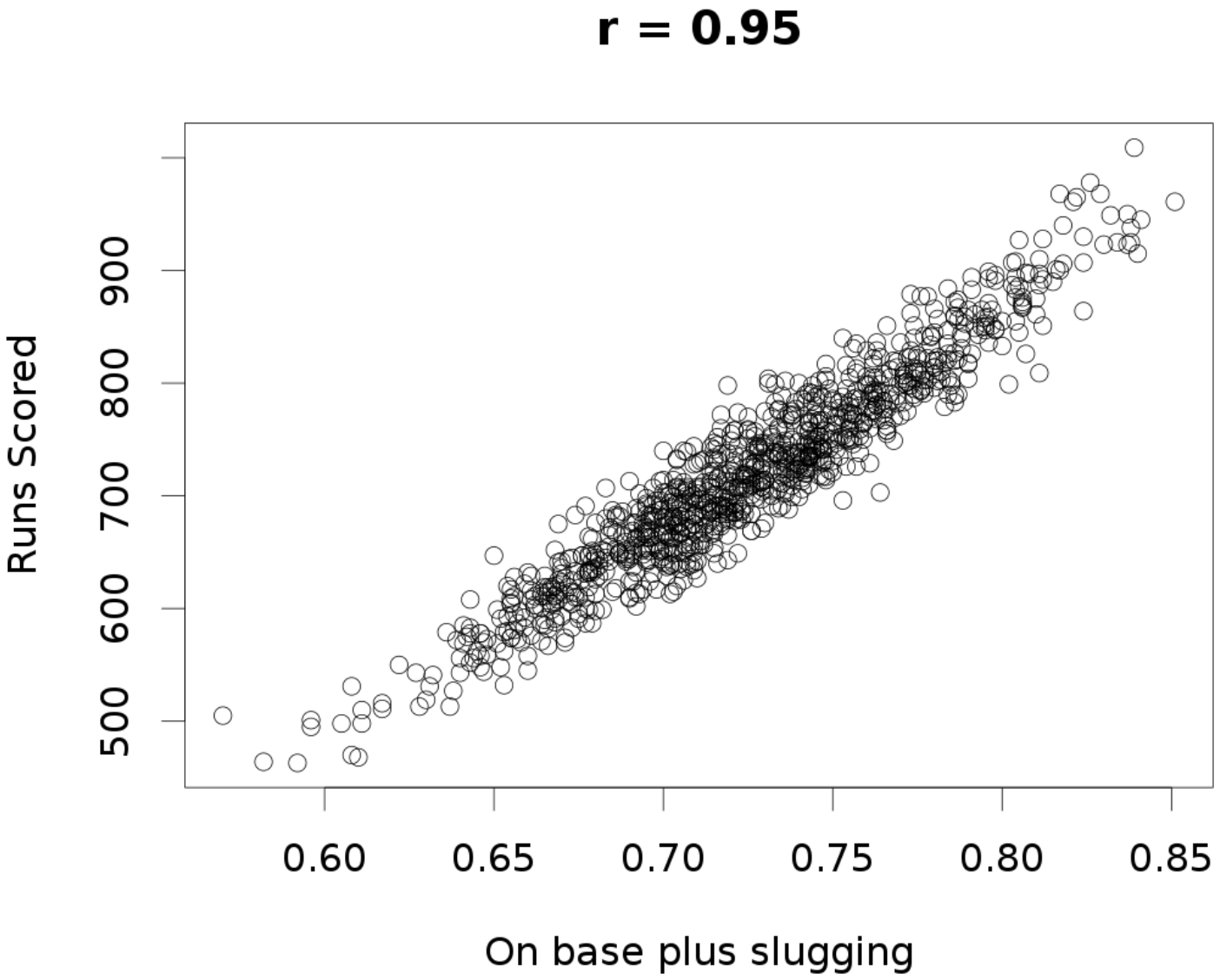
Correlation between OBP and runs



Correlation between Slug and runs



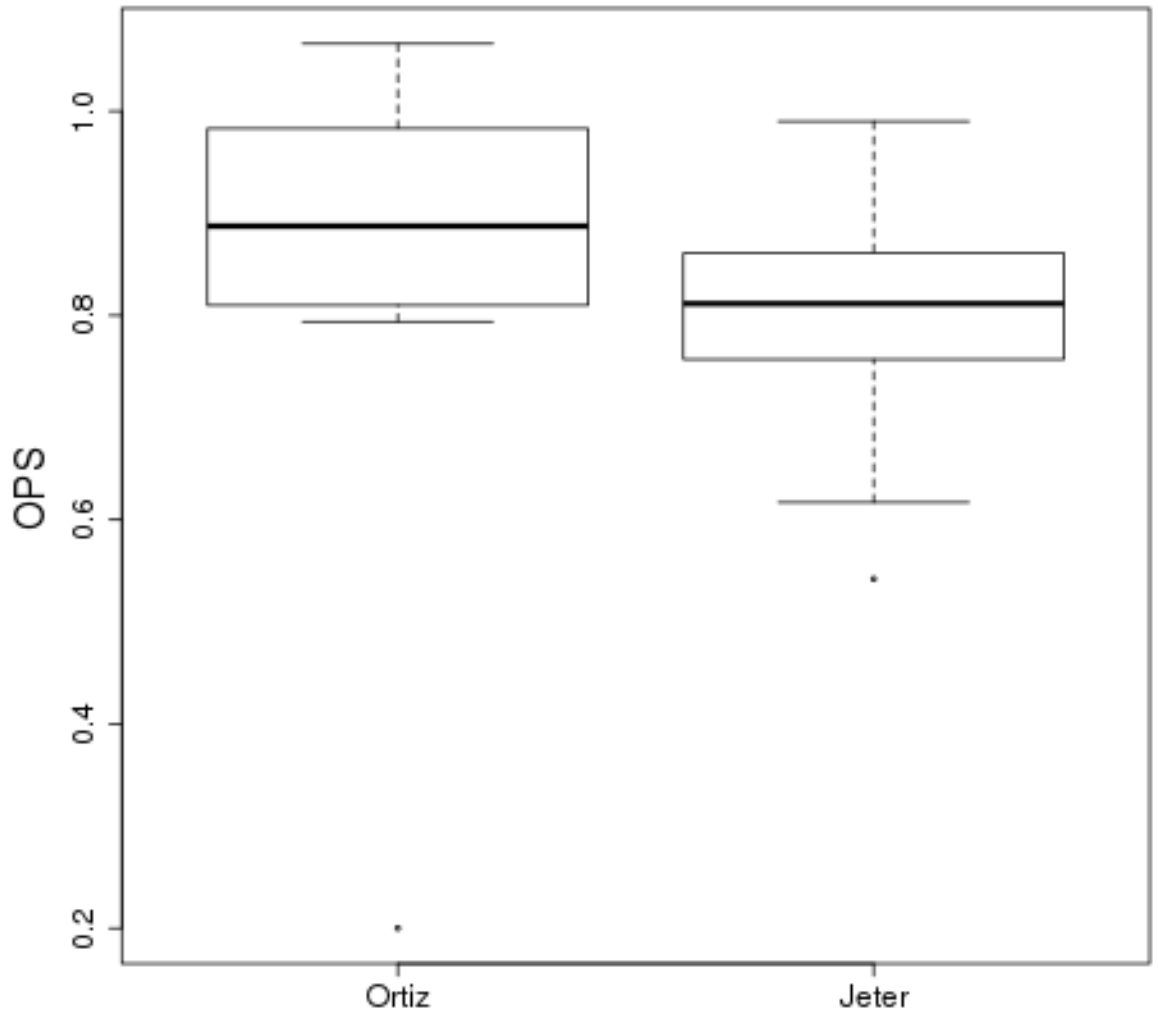
Correlation between OPS and runs



The winner…

On-base plus slugging seems like the best statistic to use!

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



Ortiz has a better on-base plus slugging!

Better know a player: Derek Jeter

[Onion infographic]

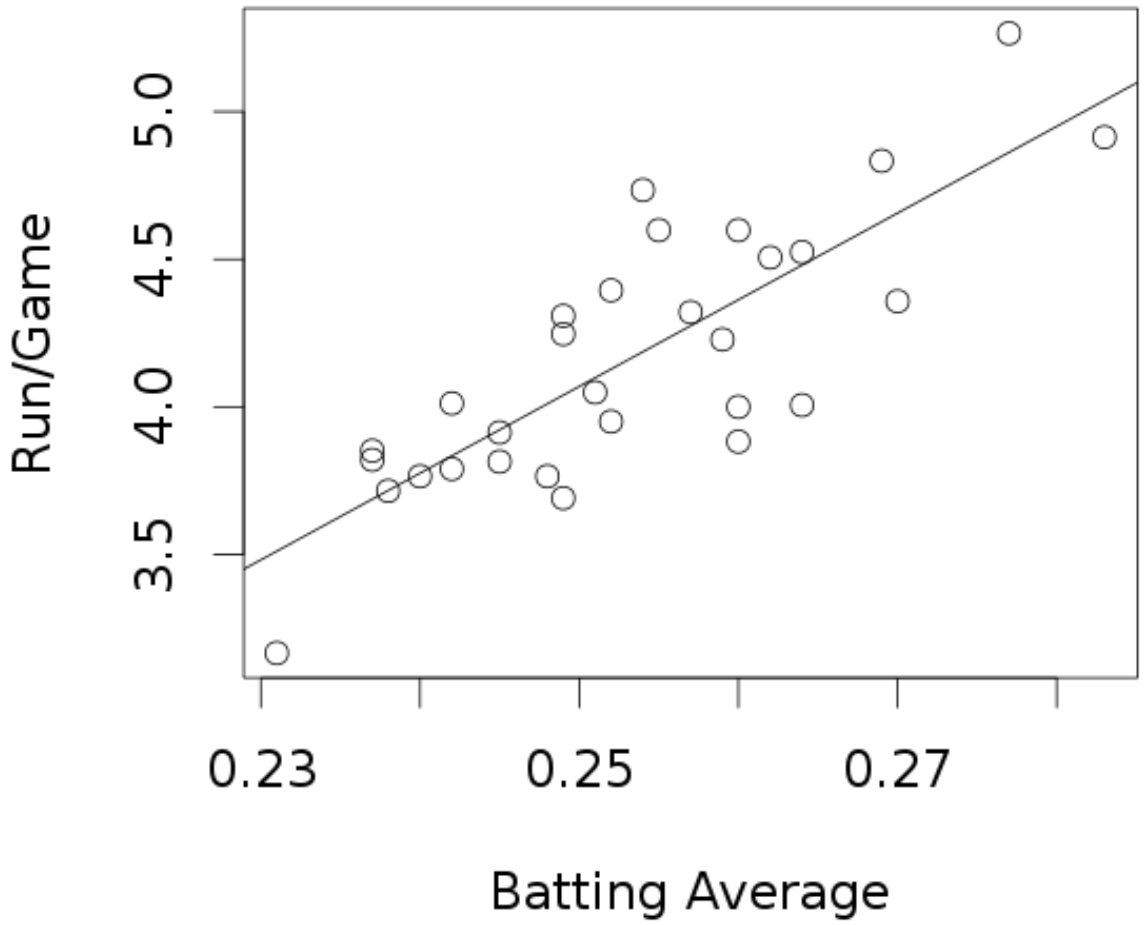
[Other Onion articles]

Regression

Regression is method of using one variable to predict the value of a second variable

In **linear regression** we fit a line to the data, called the **regression line**.

Regression line: runs/game as a function of team batting average (2013)

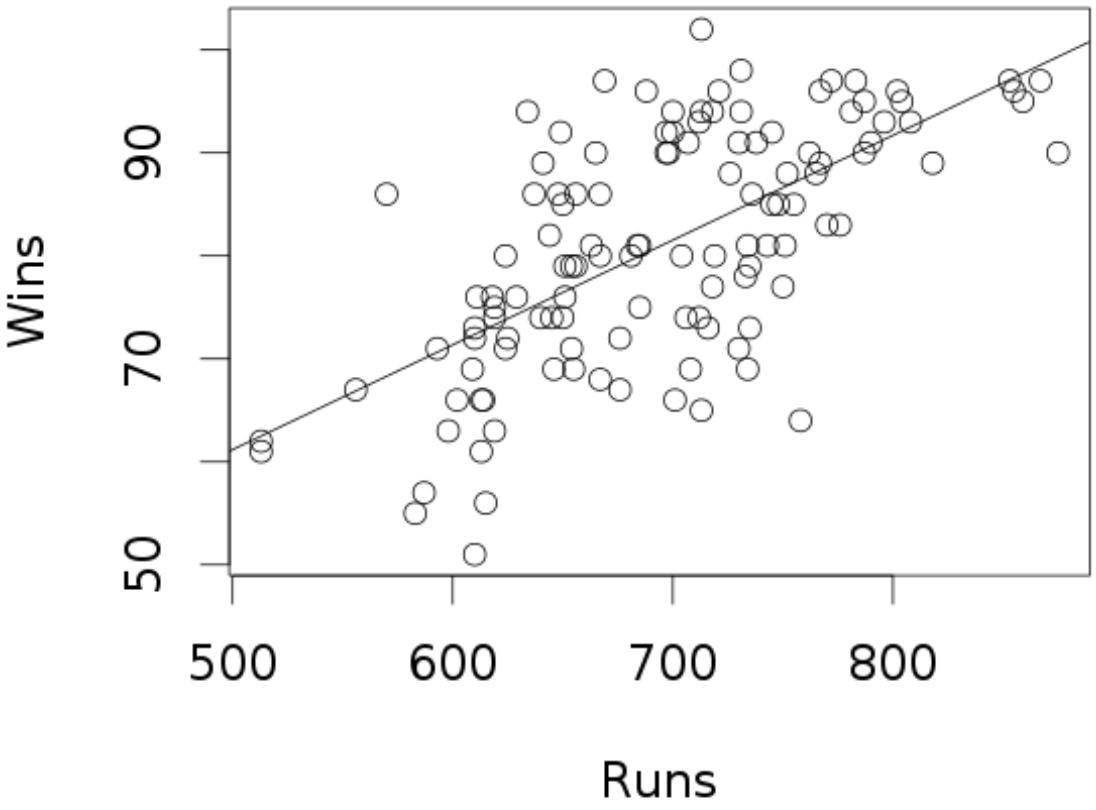


Equation for a line

$$\hat{y} = a + b \cdot x$$

$$Response = a + b \cdot Explanatory$$

Wins runs regression



$\hat{y} = a + b \cdot x$

a = *14.47*

b = *.088*

$\hat{w} = 14.47 + .088 \cdot runs$

```
R: lm(y ~ x)
```

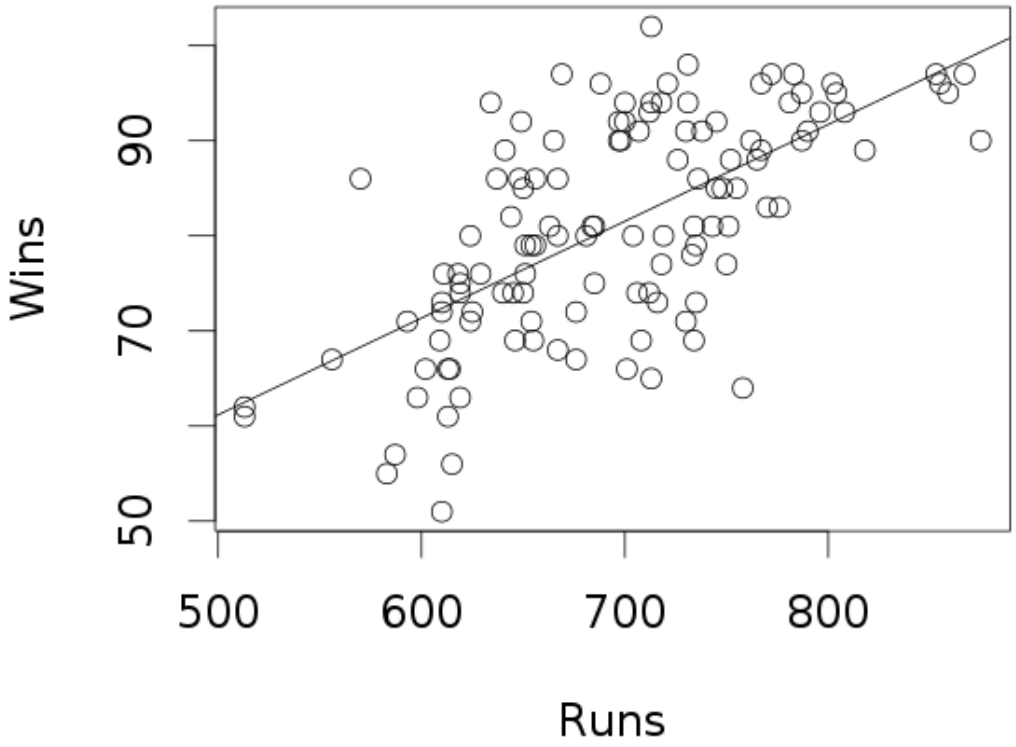
Interpreting the slope and intercept

$\hat{y} = a + b \cdot x$

The slope *b* represents the predicted change in the response variable *y* given a one unit change in the explanatory variable *x*

The intercept a represented the predicted value of the response variable y if the explanatory variable x were 0

Using the regression line to make predictions



$\hat{y} = a + b \cdot x$

a = *14.47*

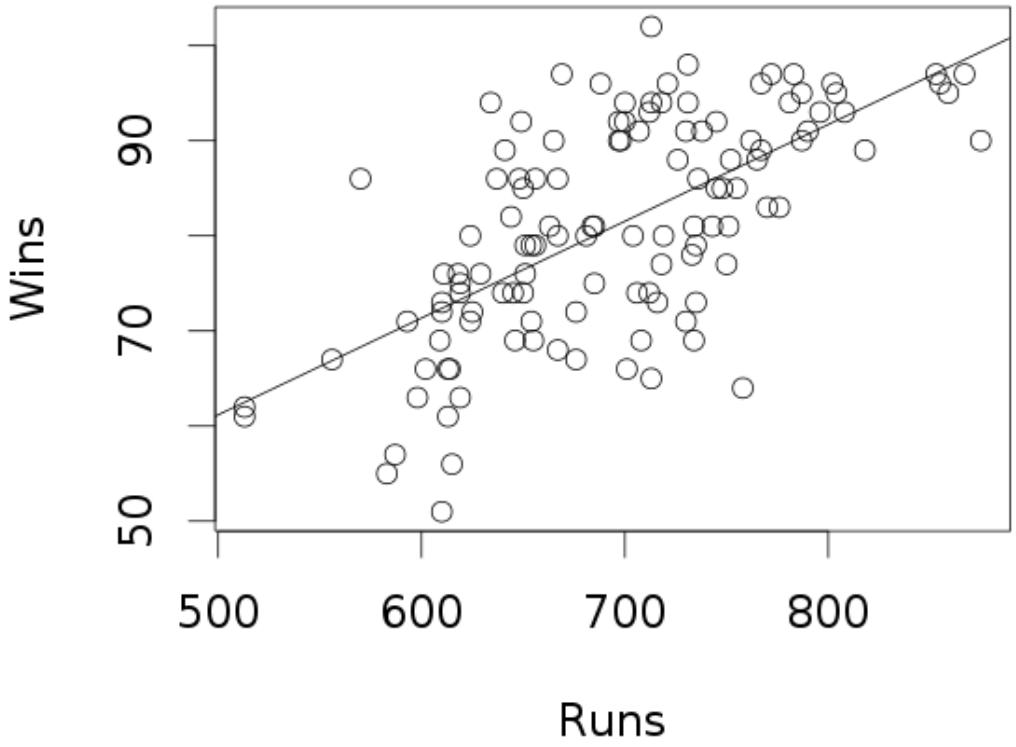
b = *.088*

$\hat{w} = 14.47 + .088 \cdot Runs$

1. Approximately how many additional runs do you need to score for an additional win?

2. How many wins will you have if you score 0 runs all season?

Wins runs regression



$\hat{y} = a + b \cdot x$

a = *14.47*

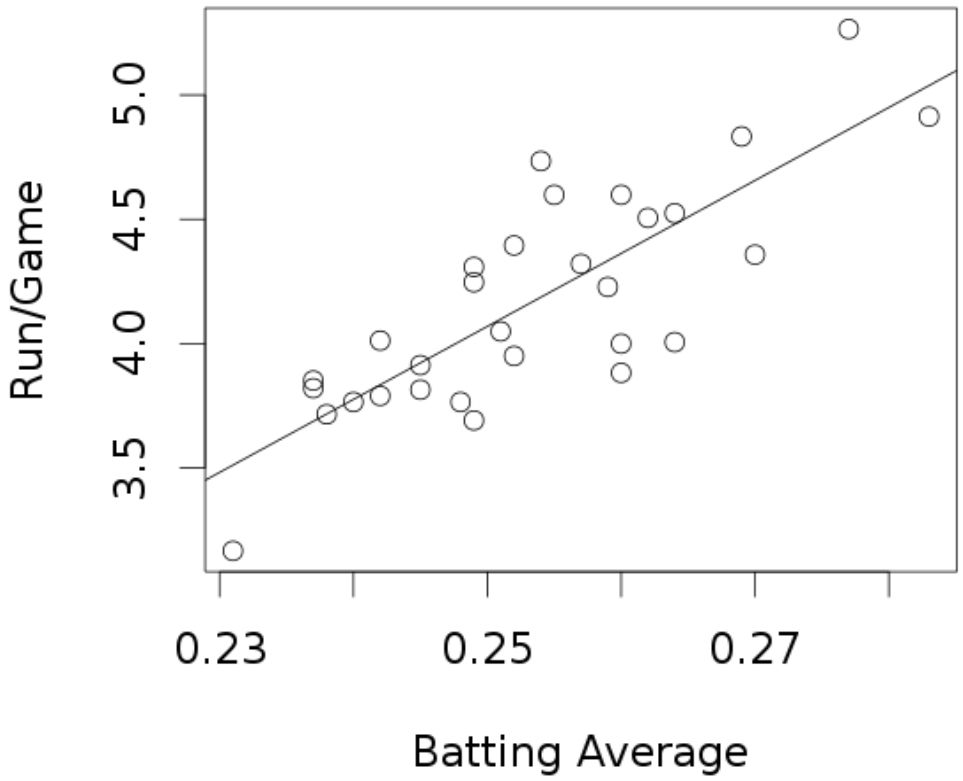
b = *.088*

$\hat{w} = 14.47 + .088 \cdot Runs$

1. An additional win for ~11 additional runs scored
2. There will be 14.47 wins if you score 0 runs all season

Example 2: Using the regression line to make predictions

1. If a team had a batting average of 0.270, how many runs would you expect in a game?



$\hat{y} = a + b \cdot x$

a = -3.27

b = 29.36

If a team had a batting average of 0.270, how many runs would you expect in a game?

$(R/G)_{expected} = 29.35 * BA - 3.27$

$\hat{y} = a + b \cdot x$

$(R/G)_{expected} = 29.35 * .270 - 3.27$

$a = -3.27$

$R/G)_{expected} = 4.6572$

$b = 29.36$

How about if a team batting .250?