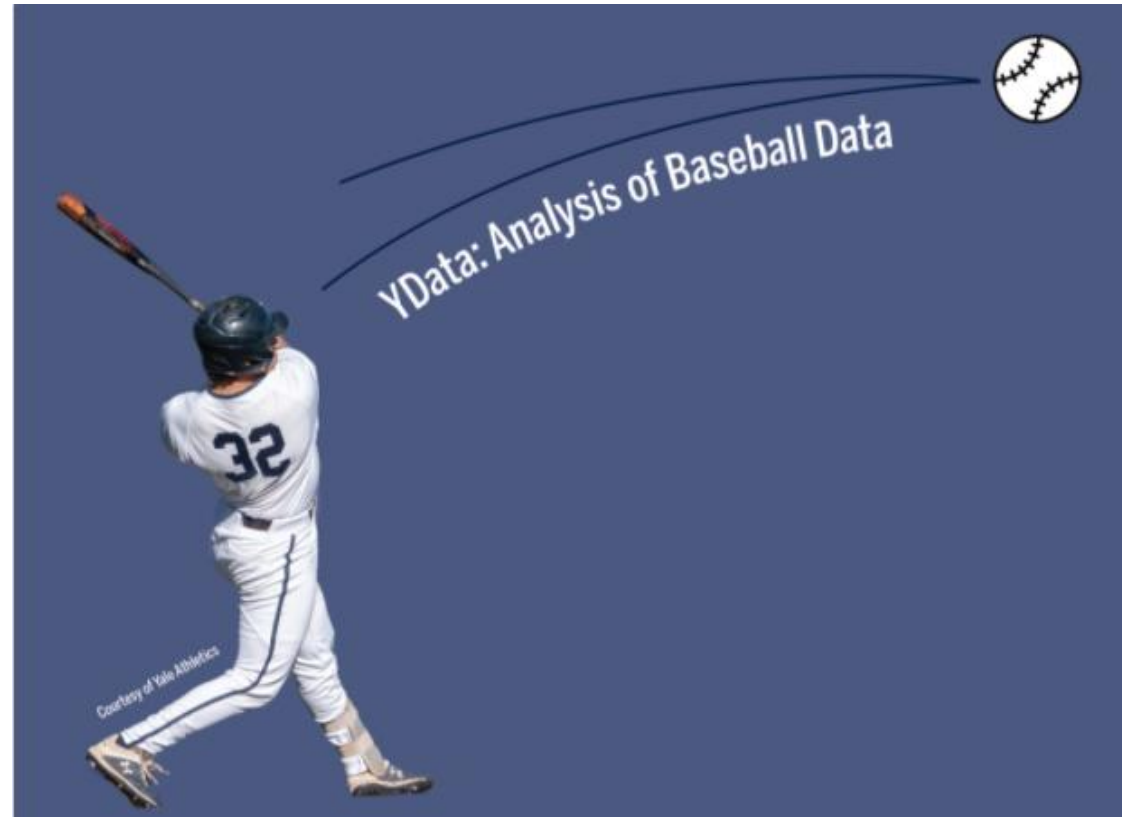


Regression



Overview

Discussion of chapter 8 of Astroball

Quick review of correlation and lab 8 answers

Simple linear regression

Multiple regression

Announcement

Statcast data is on GitHub!

```
data_url =  
'https://raw.githubusercontent.com/emeyers/SDS173/master/data/compressed_statcast_data/statcast_2019.bz2'
```






```
statcast_data = Table.from_df(pd.read_csv(data_url))
```

How is progress going on the final project?

The MLB season is in week 3...

Project idea –

How well do the first 3 weeks of the season predict a team's final record?

AL East								
Team	W	L	Pct	GB	Home	Away	L10	
 Red Sox	12	6	.667	-	6-5	6-1	7-3	
 Rays	10	8	.556	2.0	3-4	7-4	6-4	
 Orioles	8	9	.471	3.5	1-6	7-3	4-6	
 Blue Jays	7	10	.412	4.5	3-3	4-7	4-6	
 Yankees	6	10	.375	5.0	4-6	2-4	3-7	

Astroball discussion

Let's discuss the chapter for 8 minutes in breakout rooms and then have a larger conversation as a group

- Discuss your quote and reaction to chapter 8

Ben Reiter will attend class next week.

What questions should we ask him?

Astroball Chapter 8

Trade for Verlander

- Astros pay \$20 million of his salary, Tigers pay \$8 million
 - 2 years on contract



Jake Rogers

Position: Catcher
Bats: Right • **Throws:** Right
6-1, 192lb (185cm, 87kg)
Team: [Detroit Tigers](#) (minors, 40-man)

[More bio, uniform, draft, salary info ▼](#)

34

SUMMARY	WAR	AB	H	HR	BA	R	RBI	SB	OBP	SLG	OPS	OPS+
Career	-0.6	112	14	4	.125	11	8	0	.222	.259	.481	28



Daz Cameron

Position: Centerfielder
Bats: Right • **Throws:** Right
6-2, 185lb (188cm, 83kg)
Team: [Detroit Tigers](#) (minors, 40-man)


[More bio, uniform, draft, salary info ▼](#)

SUMMARY	WAR	AB	H	HR	BA	R	RBI	SB	OBP	SLG	OPS	OPS+
Career	-0.6	57	11	0	.193	4	3	1	.220	.263	.483	32

41



Franklin Perez

Position: Starting Pitcher
Bats: Right • **Throws:** Right
6-3, 197lb (190cm, 89kg)
Team: [Detroit Tigers](#) (minors, 40-man)
Born: December 6, 1997 (Age: 23-134d) in Valencia, [Venezuela](#) 

Full Name: Franklin Eduardo Perez
[View Player Info](#) from the [B-R Bullpen](#)

Astroball Chapter 8



Kate Upton  @KateUpton · Nov 16, 2016

Hey @MLB I thought I was the only person allowed to fuck
@JustinVerlander ?! What 2 writers didn't have him on their ballot?

 3.7K

 76.6K

 115.7K

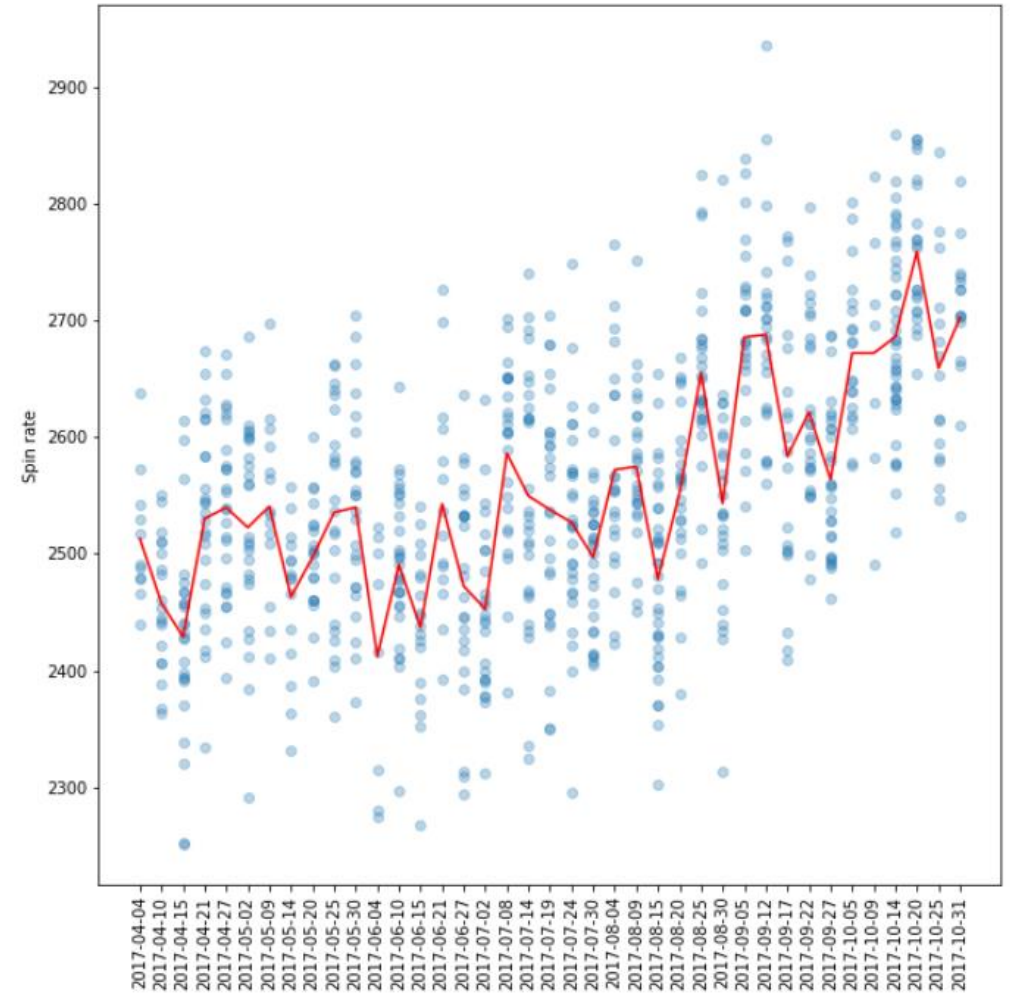


playerID	yearID	stint	teamID	lgID	W	L	G	GS	CG	SHO	SV	IPouts	H	ER	HR	BB	SO	BAOpp	ERA	IBB	WP	HBP	BK	BFP	GF	R	SH	SF	GIDP
porceri01	2016	1	BOS	AL	22	4	33	33	3	0	0	669	193	78	23	32	189	0.23	3.15	0	3	13	0	890	0	85	2	3	16
verlaju01	2016	1	DET	AL	16	9	34	34	2	0	0	683	171	77	30	57	254	0.207	3.04	1	6	8	0	903	0	81	4	7	8

Astroball Chapter 8

He began tinkering. He gripped his slider farther back in his hand, and slightly adjusted the position of his wrist, to impart more downward movement on it. On June 21, he struck out 11 Seattle Mariners in just 5 $\frac{2}{3}$ innings. "I was giddy about it," he said. "I came home and told Kate right away: 'I found it.'"

When Verlander arrived in Houston, the members of the Nerd Cave – especially Mike Fast – had been assiduously analyzing each of his starts, via video and Statcast, for months. They were dying to ask him one question, though they didn't want to come on too strong. "By the way," they asked, "did you change your grip, or something?". pg 192



Astroball Chapter 8

```
data_url =  
'https://raw.githubusercontent.com/emeyers/SDS173/master/data/compressed_statcast_  
data/statcast_2017.bz2'  
  
statcast_data = Table.from_df(pd.read_csv(data_url))  
  
verlander_statcast = statcast_data.where('pitcher', 434378)  
  
verlander_slider = verlander_statcast.where('pitch_type', 'SL').where('release_spin_rate',  
are.above(2250)).sort('game_date')  
  
plt.scatter(verlander_slider['game_date'], verlander_slider['release_spin_rate'], alpha = .3);
```

Quick review of correlation

Do power hitters strike out more often?

Chris Davis in 2013:

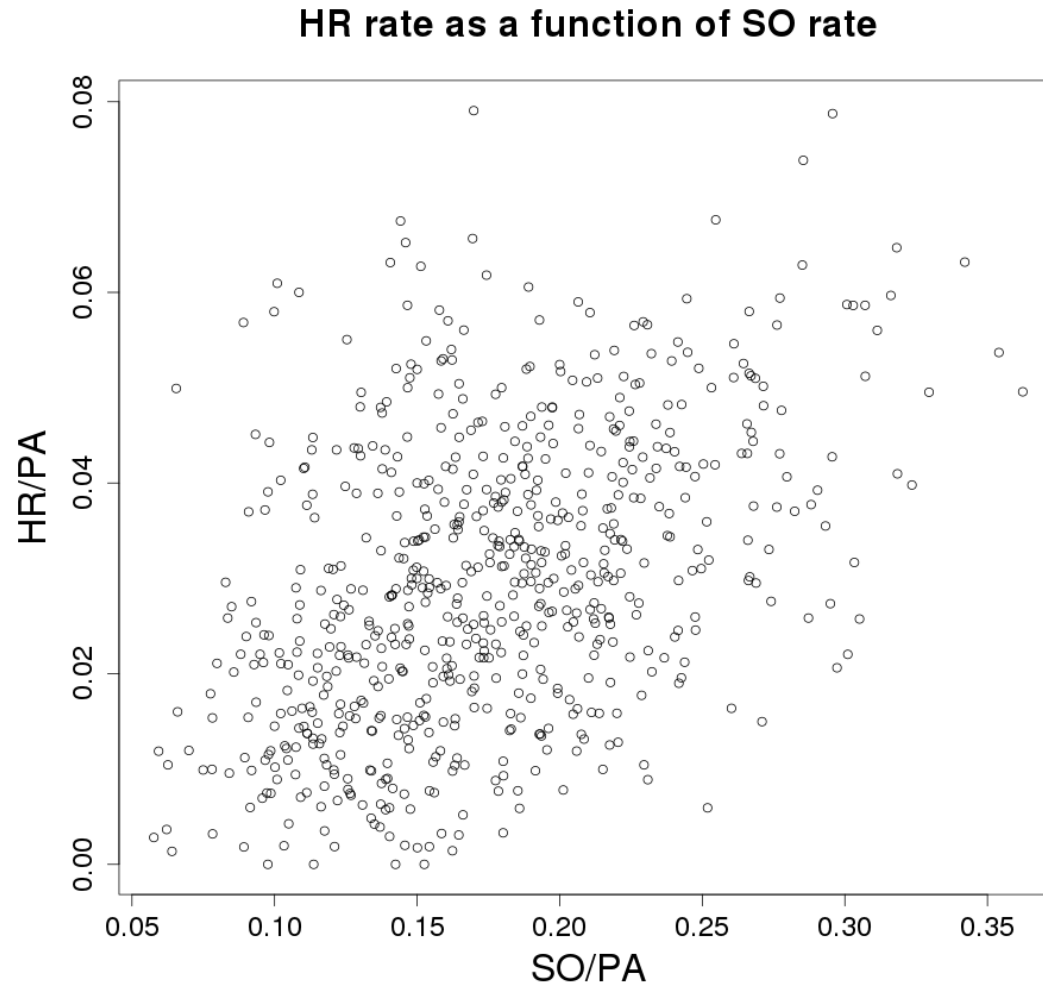


53 home runs



199 strike outs

Scatter plots: 2010-2014 home run rate as a function of strike out rate



```
tb.scatter('x_col', 'y_col')
```

Correlation

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

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

- Correlation is always between -1 and 1: $-1 \leq r \leq 1$
- The sign of r indicates the direction of the association
- Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship
- Correlation is symmetric: $r = \text{cor}(x, y) = \text{cor}(y, x)$

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



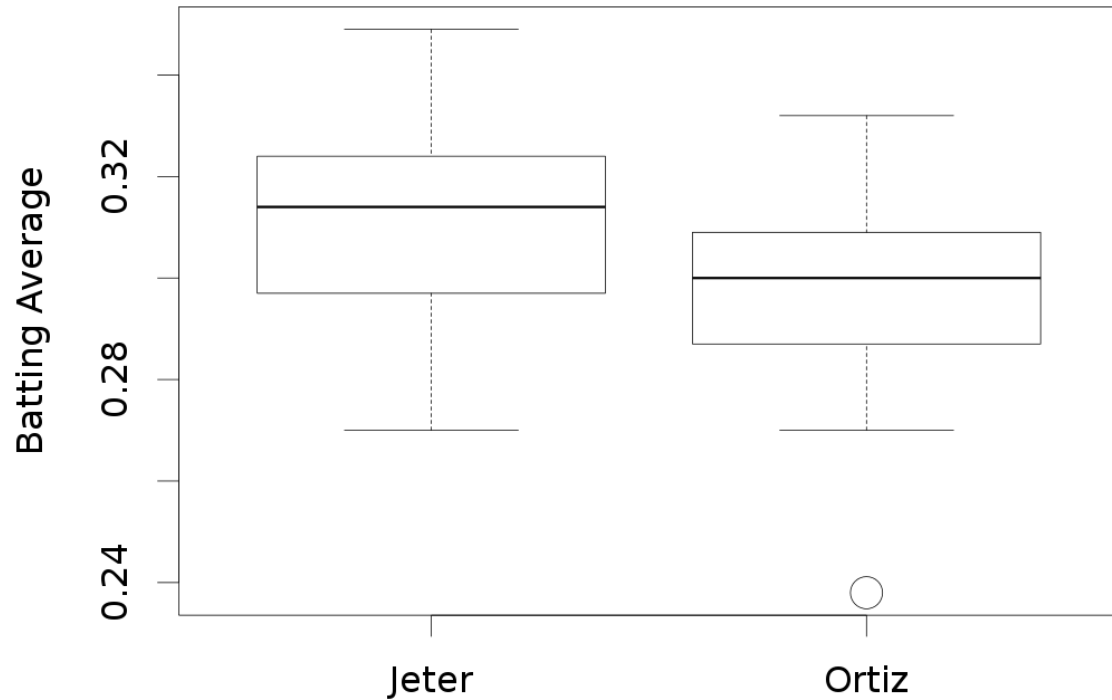
Derek Jeter



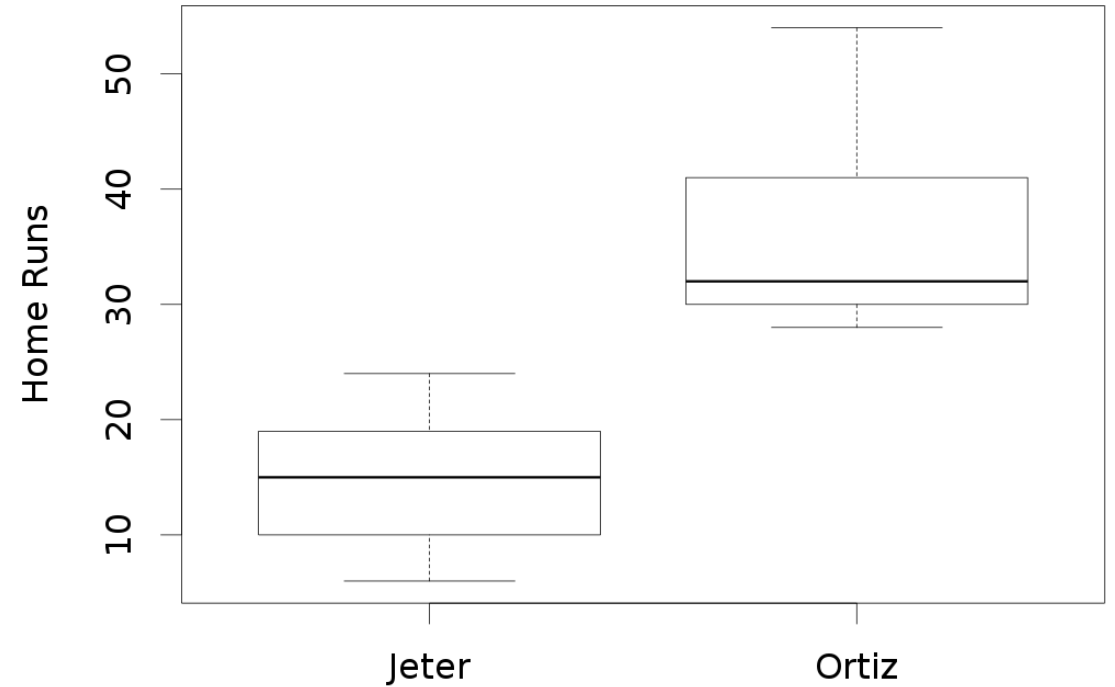
David Ortiz

How can I prove to my misguided Yankee fan friends that Ortiz is better?

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



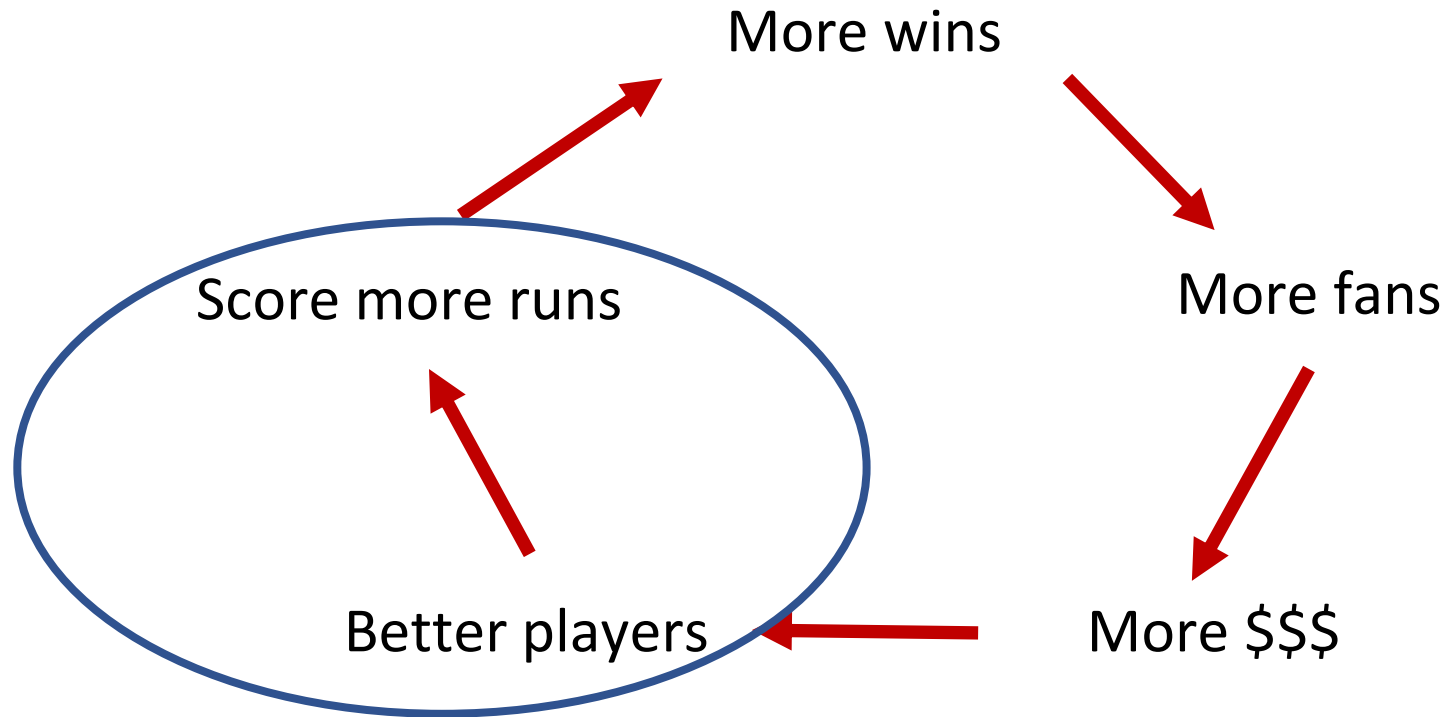
Jeter has a better batting average



Ortiz hits more home runs

Is power or batting average more important?

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

Descriptive statistics find the correlation with runs:

HR: Home runs

OBP: On-Base Percentage: $(H + BB)/PA$

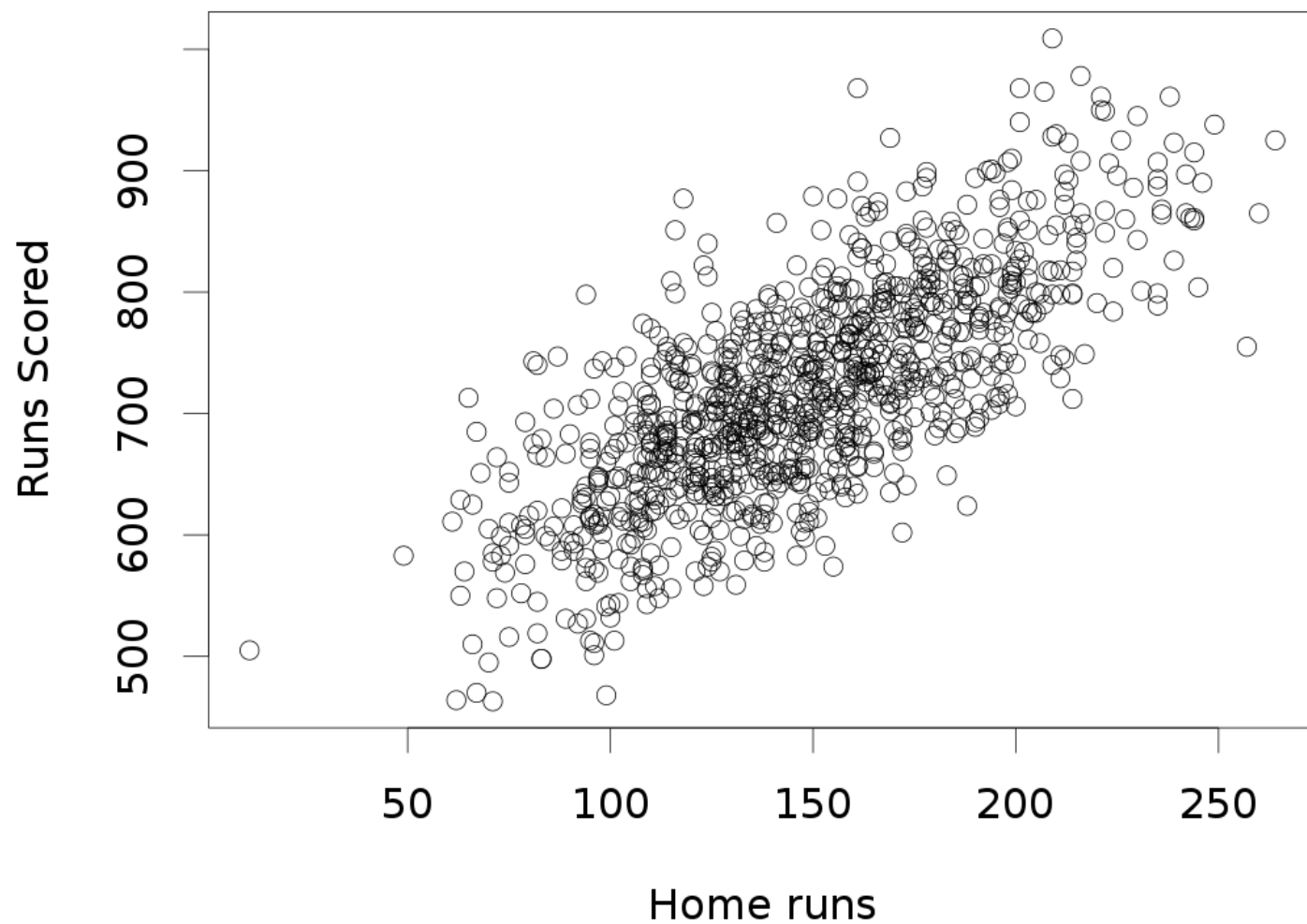
BA: Batting Average: H/AB

SLG: Slugging percentage: $(1 \cdot 1B + 2 \cdot 2B + 3 \cdot 3B + 4 \cdot HR)/AB$

On-base plus Slugging (OPS): $OBP + SLG$

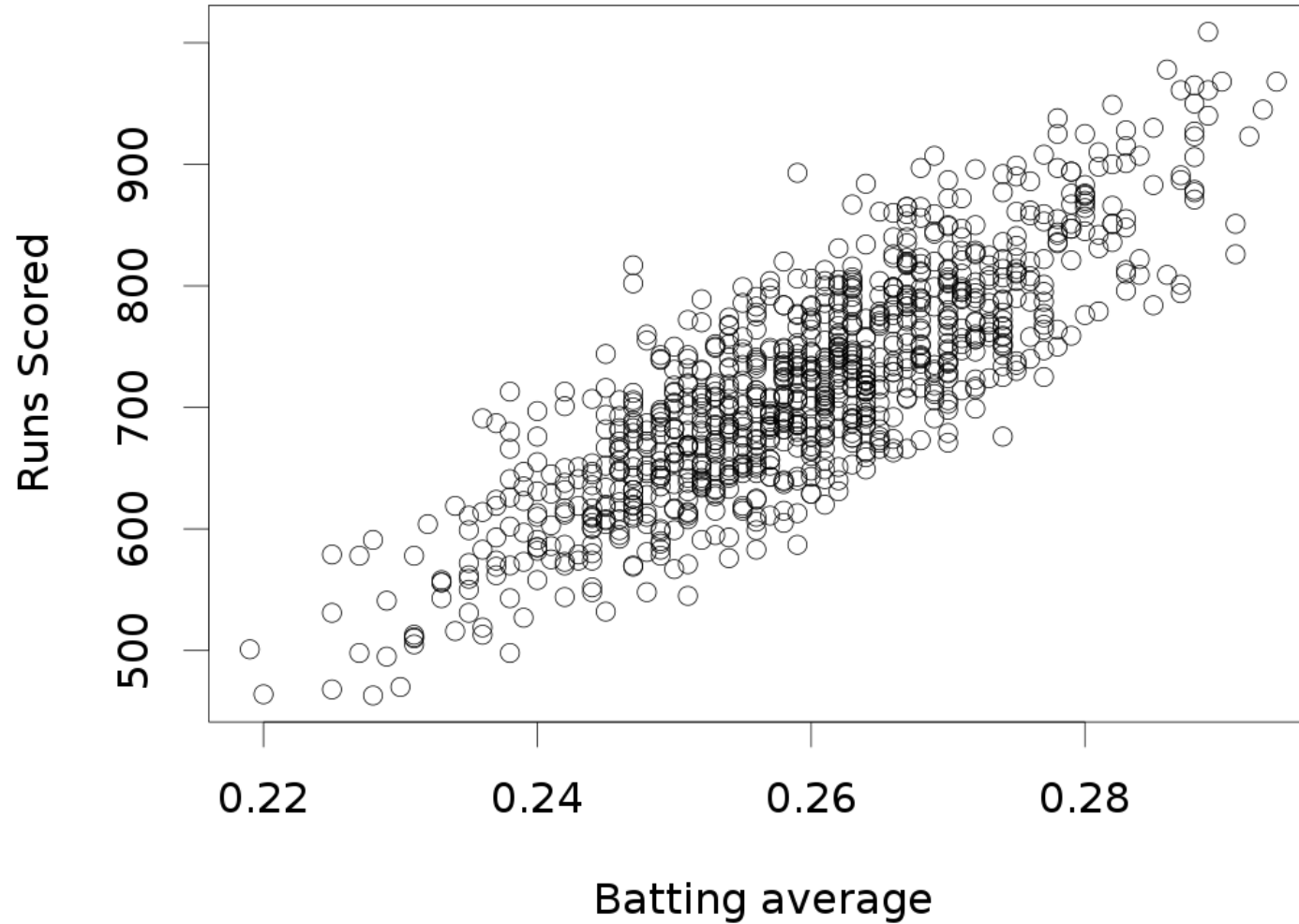
Correlation between HR and runs

$r = 0.74$



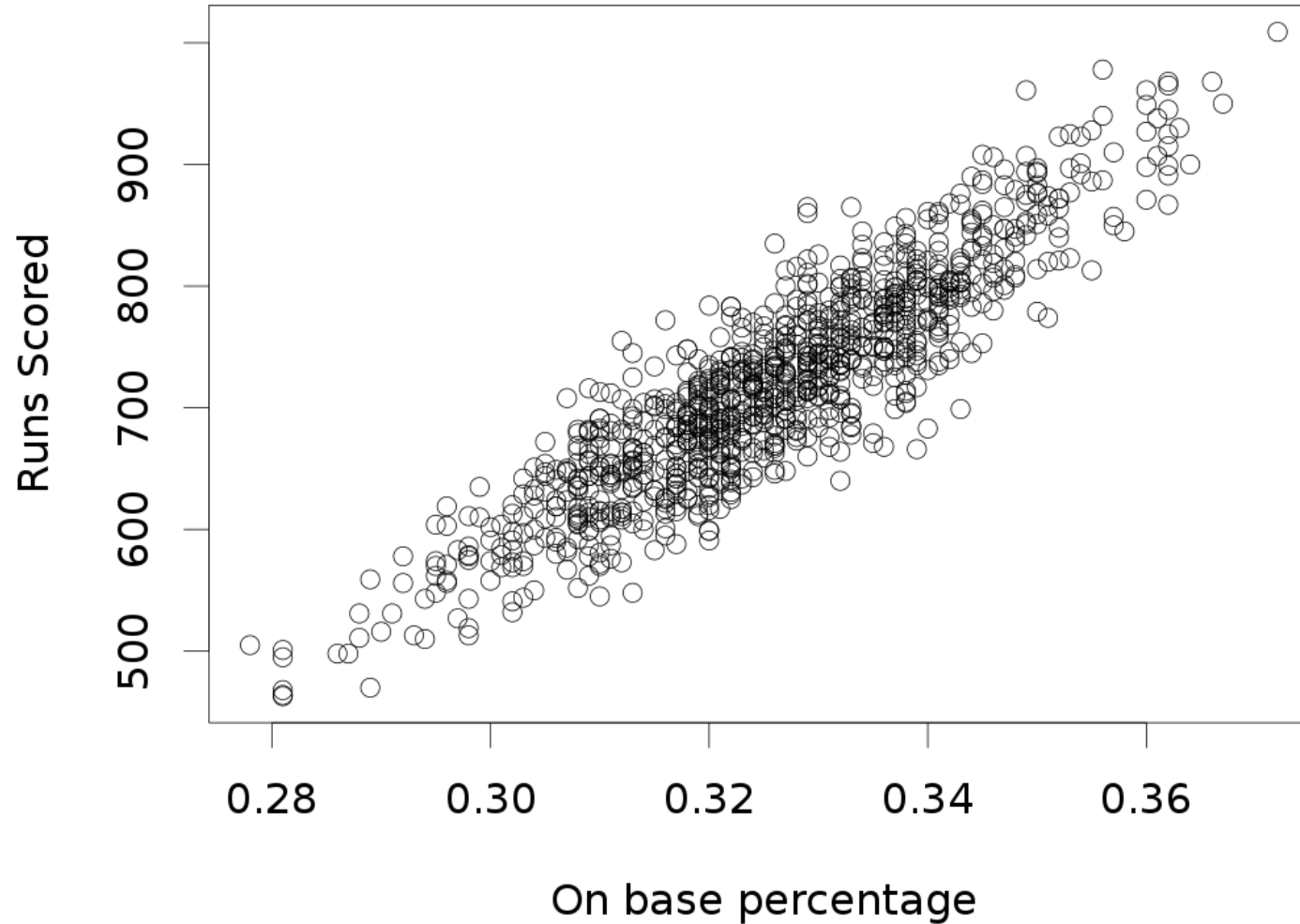
Correlation between BA and runs

$r = 0.83$



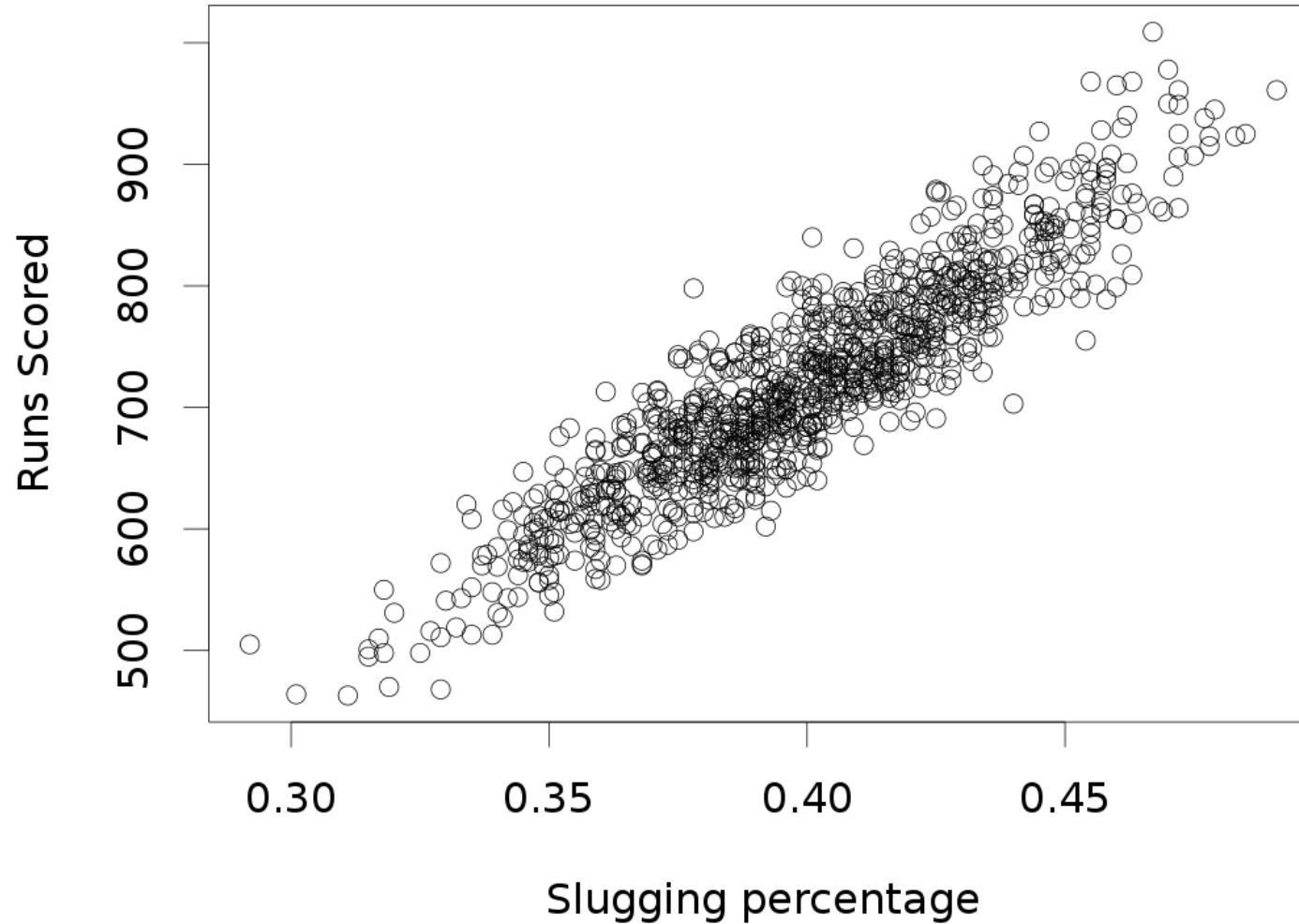
Correlation between OBP and runs

$r = 0.9$



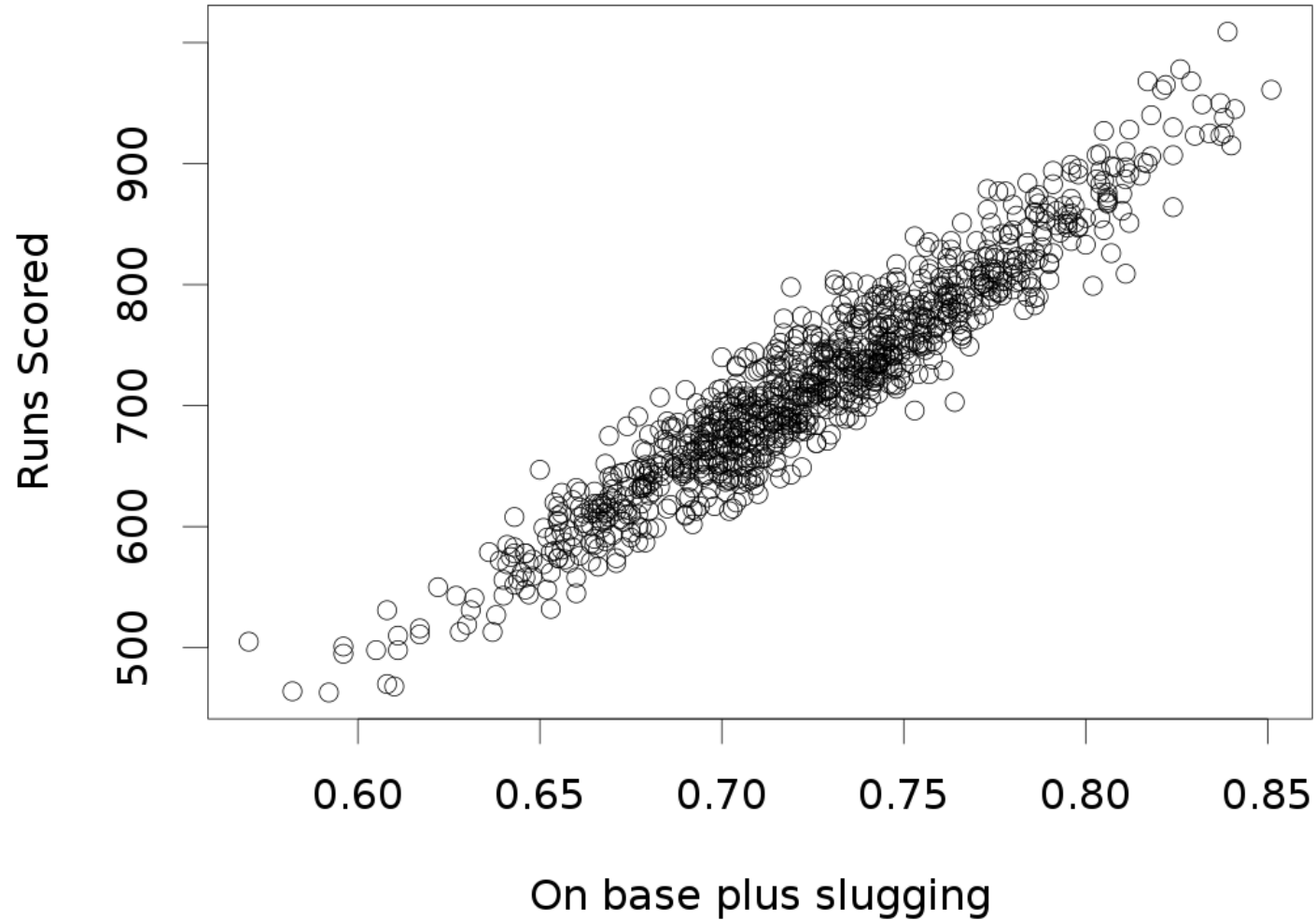
Correlation between Slug and runs

$r = 0.91$



Correlation between OPS and runs

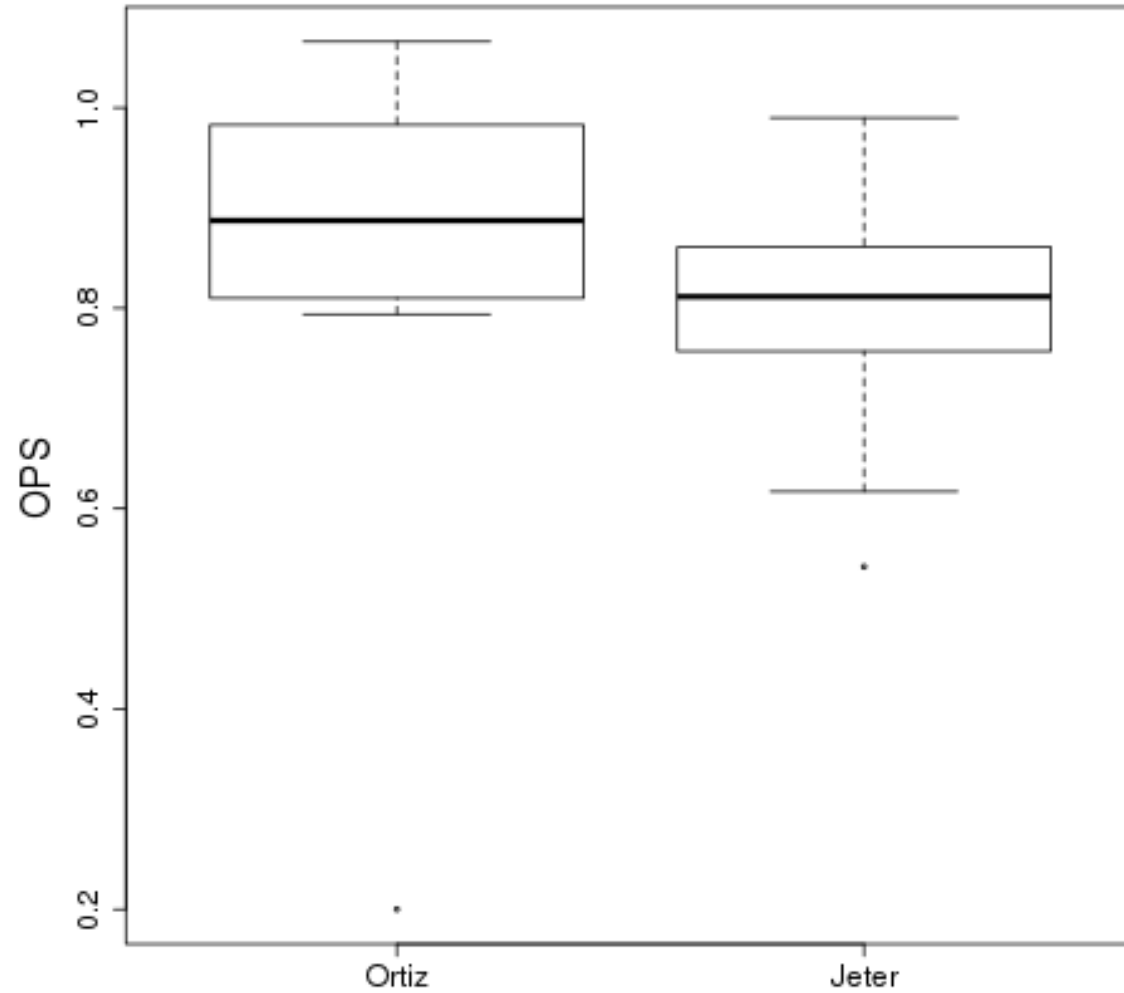
$r = 0.95$



The winner...

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

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



Are you convinced that David Ortiz is better?

Career home runs

- Jeter: 260
- Ortiz: 541

Career grand slams

- Jeter: 1
- Ortiz: 11

Ortiz has a better on-base plus slugging!

[Onion infographic](#)

[Other Onion articles](#)



Even better statistics?

On-base plus slugging (OPS) was the “best” statistic we came up with

- i.e., it had the highest correlation with runs

Do you think we could come up with a better statistic than OPS?

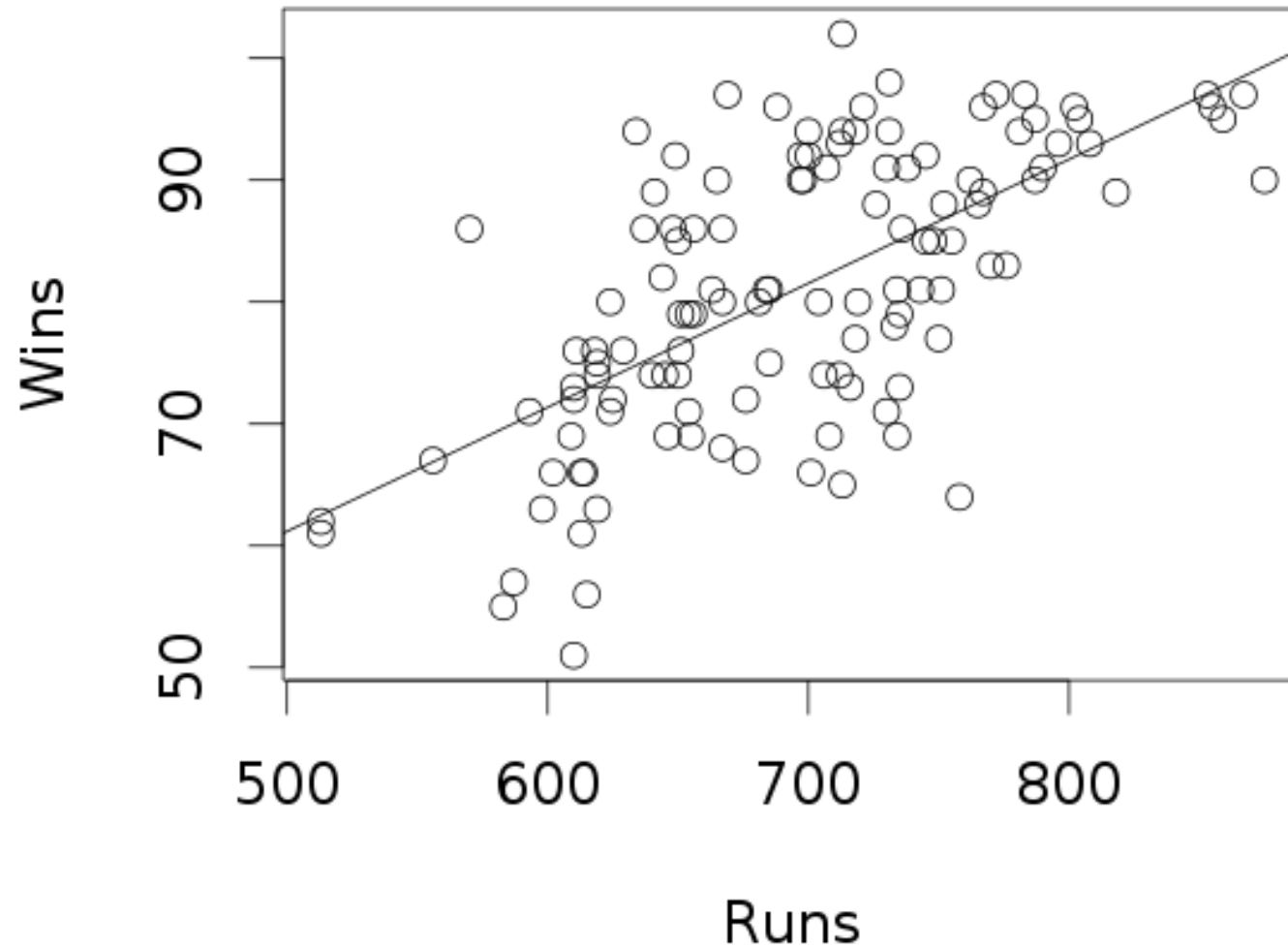
We will get back to this later, but first, regression!

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: wins as a function of runs scored



Equation for a line

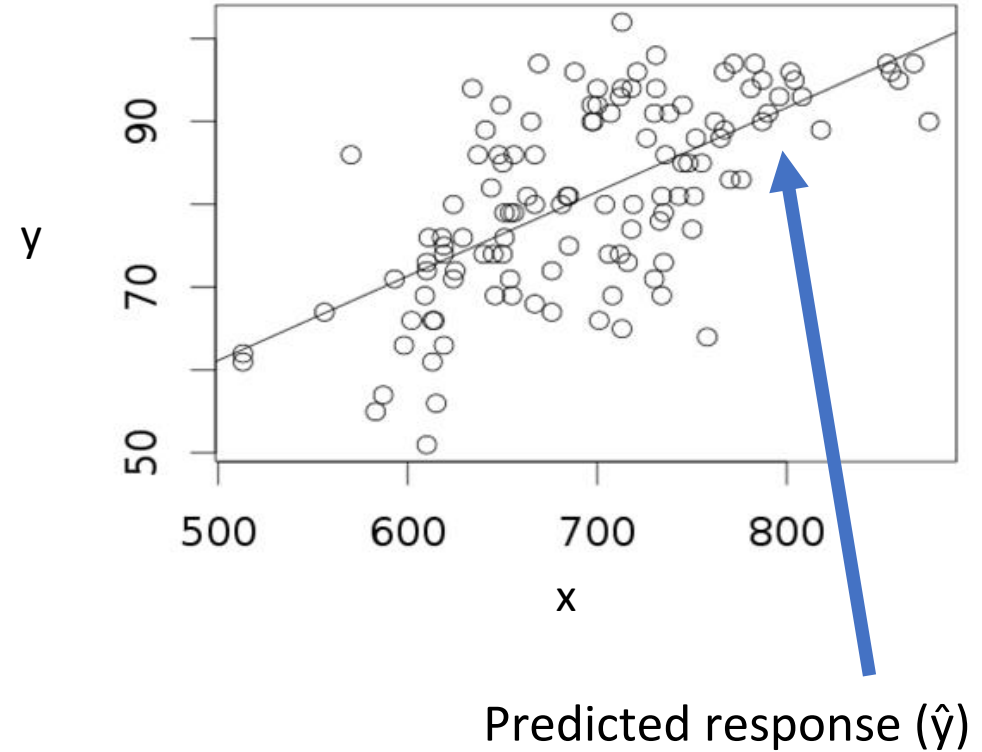
$$\hat{y} = a + b \cdot x$$

$$\hat{y} = f(x)$$

Constants
(coefficients)

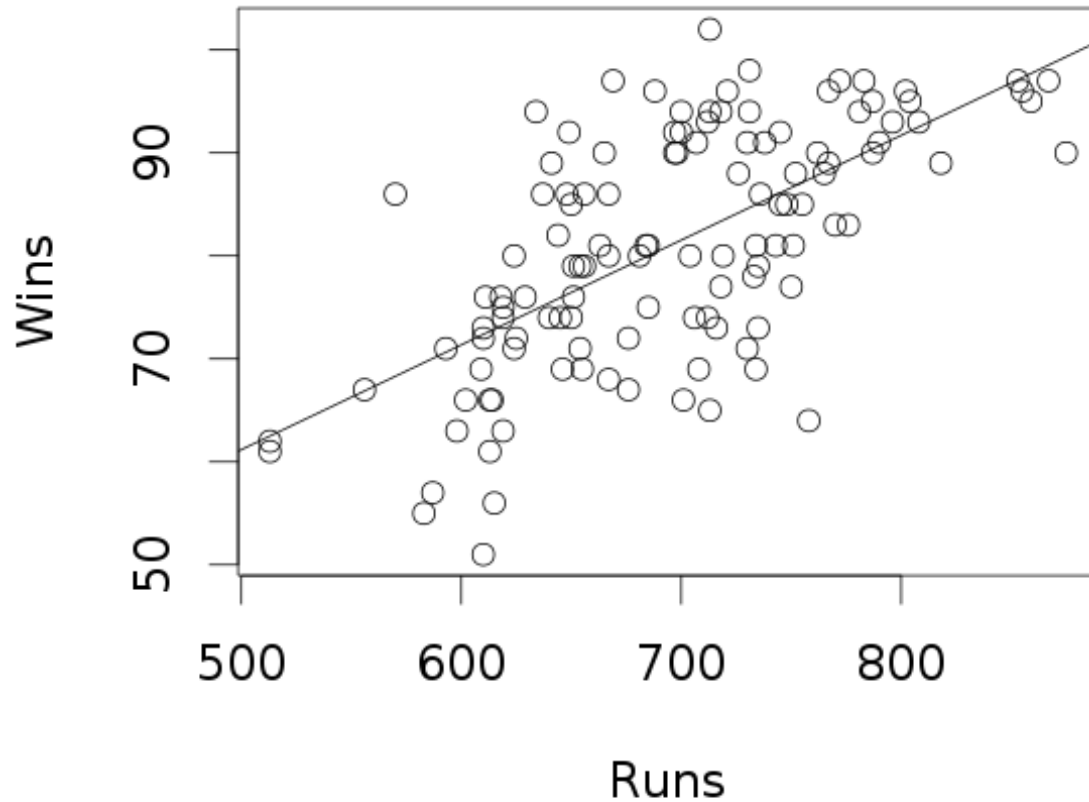
Predicted response (\hat{y})

Explanatory variable (x)



\hat{y} is our prediction (best guess) of what we think y should be based on the value x

Wins runs regression



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

$$a = 14.47$$

$$b = .088$$

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

If a team scores 700 runs, how many wins (\hat{w}) would we predict?

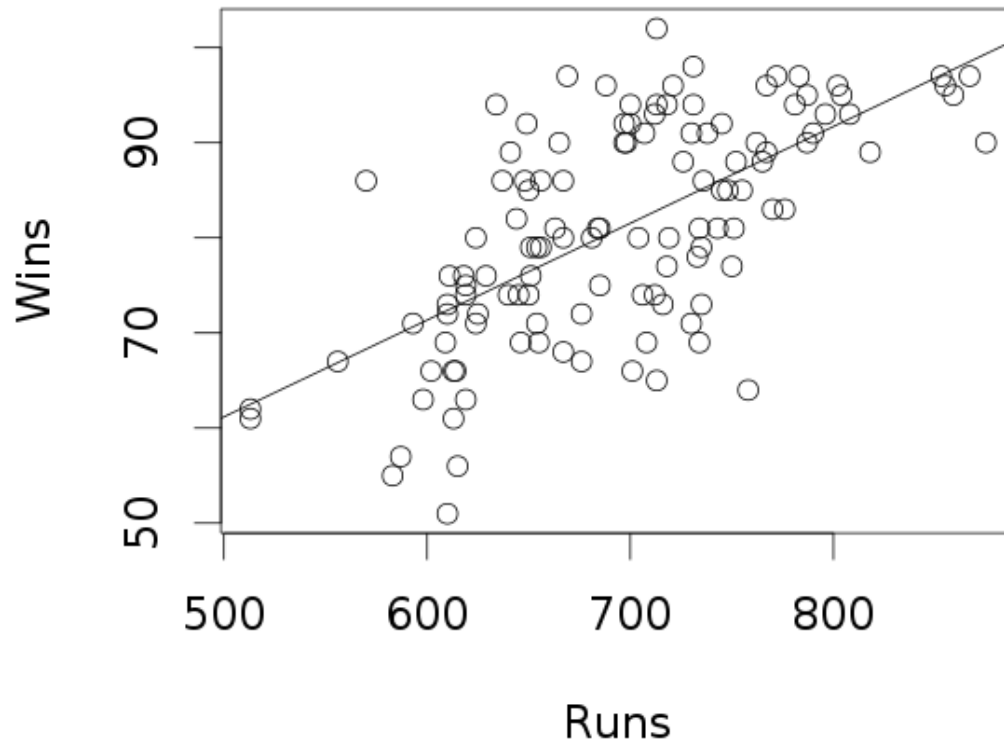
Interpreting the slope and intercept

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

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



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

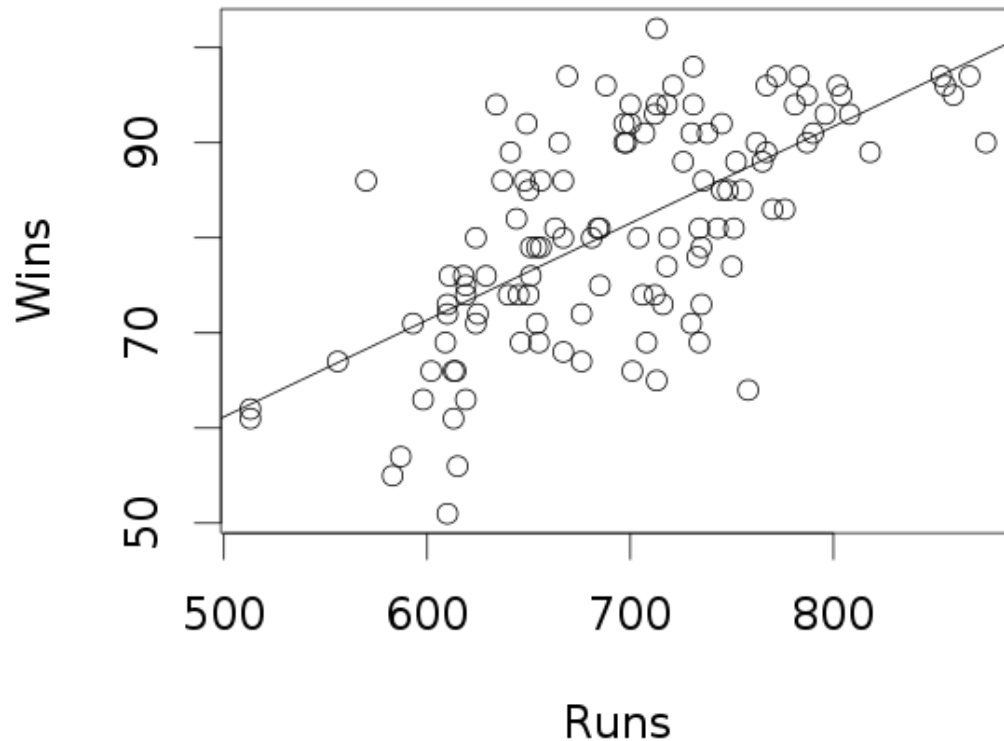
$$a = 14.47$$

$$b = .088$$

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

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?

Using the regression line to make predictions



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

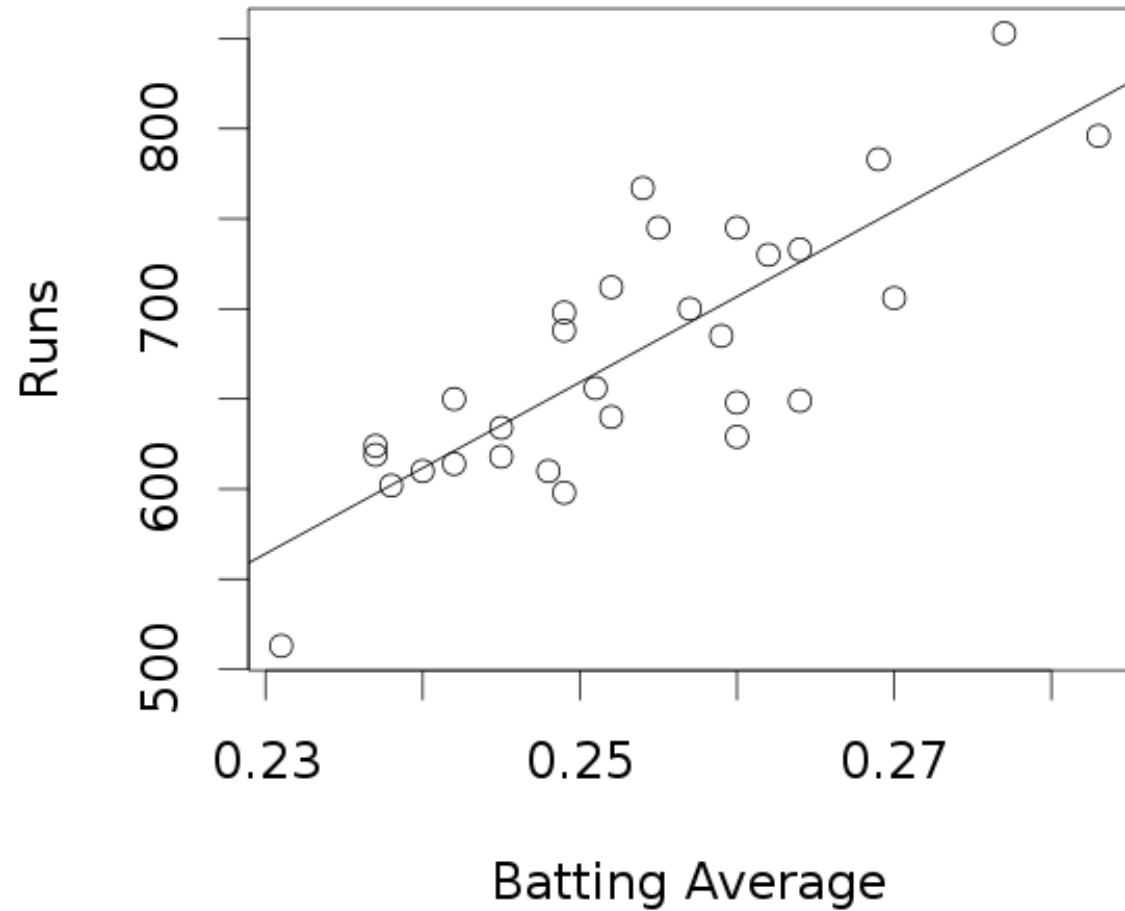
$$a = 14.47$$

$$b = .088$$

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

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: Runs and BA regression



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

$$a = -770.2$$

$$b = 5739.5$$

Write the equation for predicting runs as a function of BA

$$\hat{r} = -770.2 + 5739.5 \cdot BA$$

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

$$\hat{r} = -770.2 + 5739.5 \cdot BA$$

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

$$\hat{r} = -770.2 + 5739.5 \cdot .270$$

$$a = -770.2$$

$$\hat{r} = 772.465$$

$$b = 5739.5$$

How about if a team batting .250?

Finding the coefficients a and b

To use the linear regression equation, we need to know what the coefficients a and b are

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

The Y123 class textbook discuss one way to calculate these regression coefficients yourself based on first calculating the correlation coefficient r

$$a = \bar{y} - b \cdot \bar{x}$$

Where:

$$b = r \cdot \frac{s_y}{s_x}$$

\bar{y} and \bar{x} are the means of the y 's and x 's `np.mean()`

s_y and s_x are standard deviations of the y 's and x 's `np.std()`

r is the correlation coefficient

We will discuss other ways for finding a and b (i.e., the line of "best fit") later in the class

Finding the coefficients a and b

There are there a few popular Python libraries that are used for creating linear regression models which are:

- Numpy
 - `b1, b0 = np.polyfit(ndarray_x, ndarray_y, degree)`
- Scikit-learn
- Statsmodels

For now we will use the “formula notation” from Statsmodels package:

```
import statsmodels.formula.api as smf
```

I will go over the main ideas now, and you will try this out on lab 9

- If there is time at the end of class you can work on the lab as well

Linear regression in Python

Let's build a regression model for predicted the number of runs (\hat{r}) a team will score based on the team's batting average (BA): $\hat{r} = a + b \cdot \text{BA}$

Exercise 1.1: You will start by plotting the data with the regression line using:

```
tb.scatter('x', 'y', fit_line = True)
```

(If you are using Pandas just plot it without the regression line)

Linear regression in Python

We can build a linear model using the Statsmodels syntax:

```
lm = smf.ols('y ~ x', data = my_df).fit()
```

We can extract **a** (the intercept) and **b** (the slope) from the model using:

```
params = lm.params
```

Exercise 1.2: You will get the linear regression coefficients for predicting runs from batting average and write down the linear regression equation

Linear regression in Python

Building the linear model:

```
lm = smf.ols('R ~ BA', data = teams_2013).fit()
```

Get the coefficients:

```
the_params = lm.params
```

```
Intercept    -526.921684
```

```
BA           4744.561329
```

```
dtype: float64
```

Can you write an equation using these coefficients?

Making predictions

Once we have fit the regression model (and have the regression equation) we can use it to make predictions

To do this we can use the use:

```
sm_predictions = lm.predict(the_data)
```

Where `the_data` is a table that has as column with the same name as the x variable that was used when the model was fit

Making predictions

We can also write our own function to make predictions based on the regression coefficients that were learned

```
def make_predictions(the_coefs, x_vals):
```

```
    ...
```

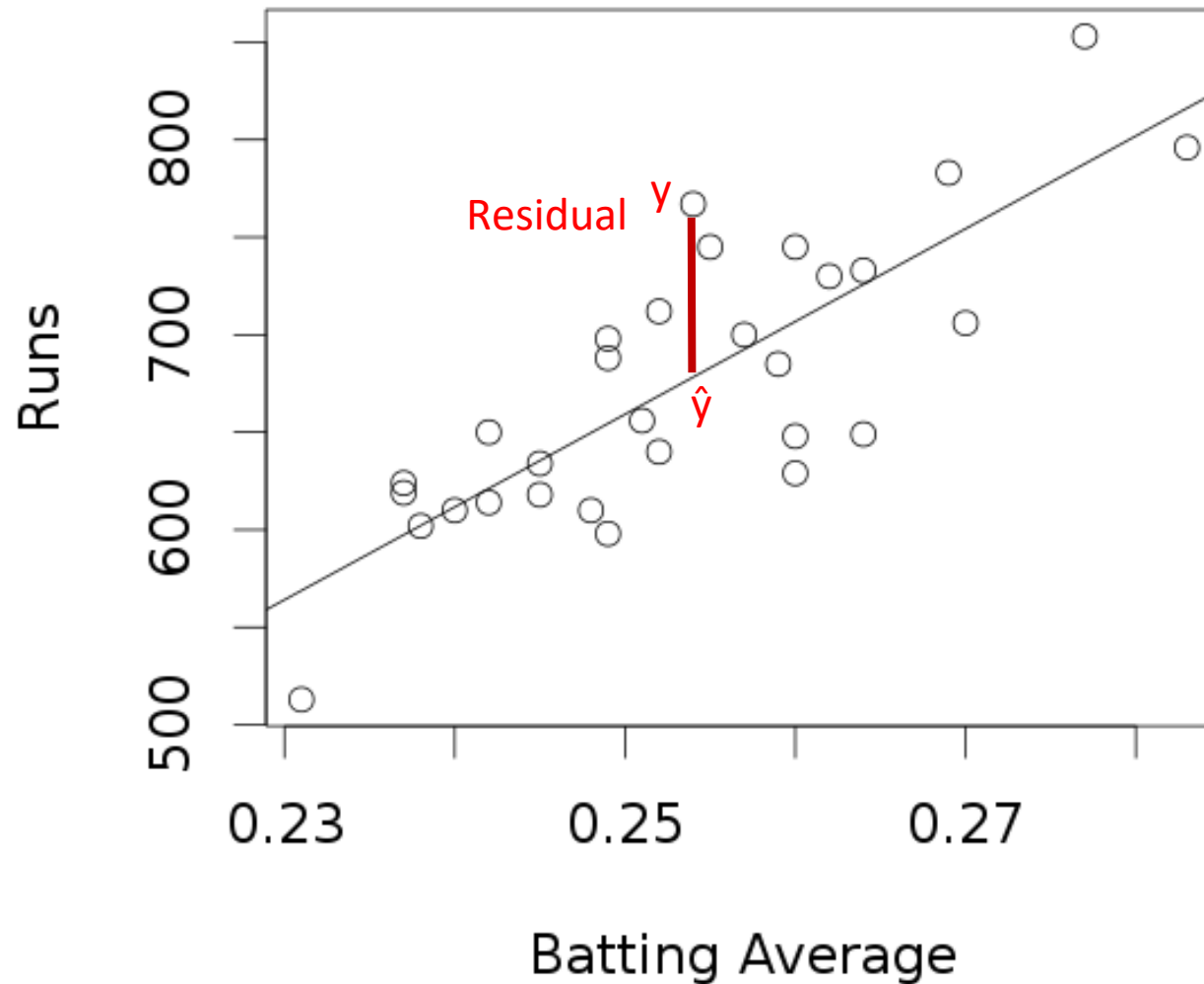
You will do this on part 2 of the homework!

Residuals

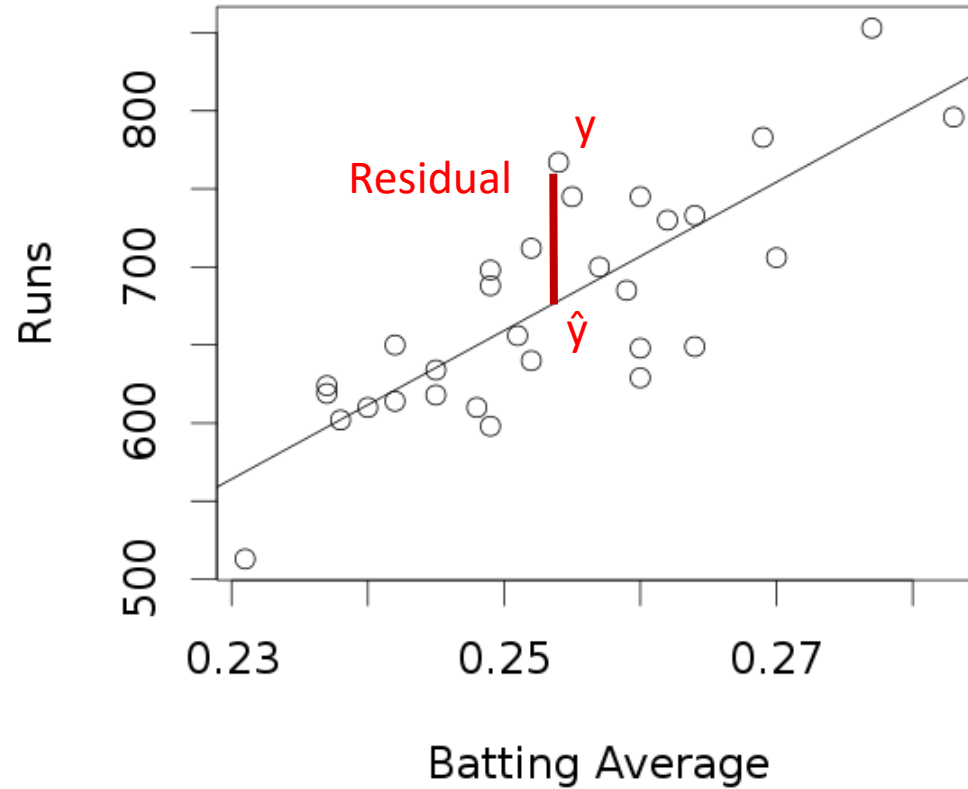
The **residual** at a data value is the difference between the observed (y) and predicted value (\hat{y}) of the response variable

$$\text{Residual} = \text{Observed} - \text{Predicted} = y - \hat{y}$$

Run vs. batting average (2013)

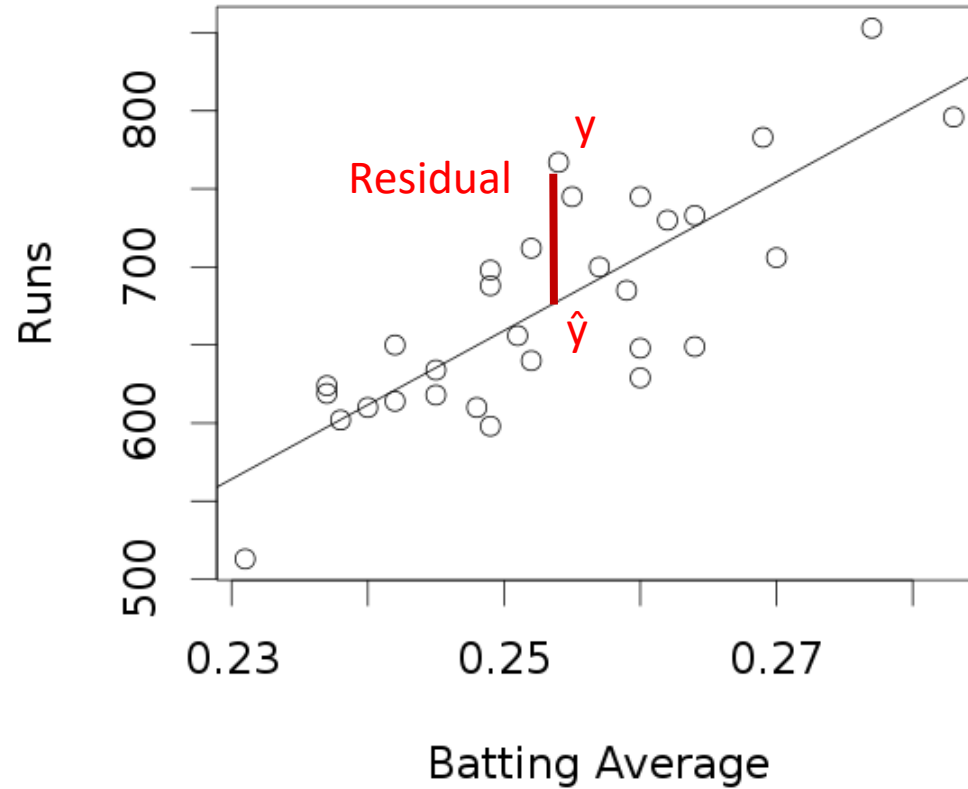


Measuring goodness of fit



If the residuals are small, then the line does a good job describing the data

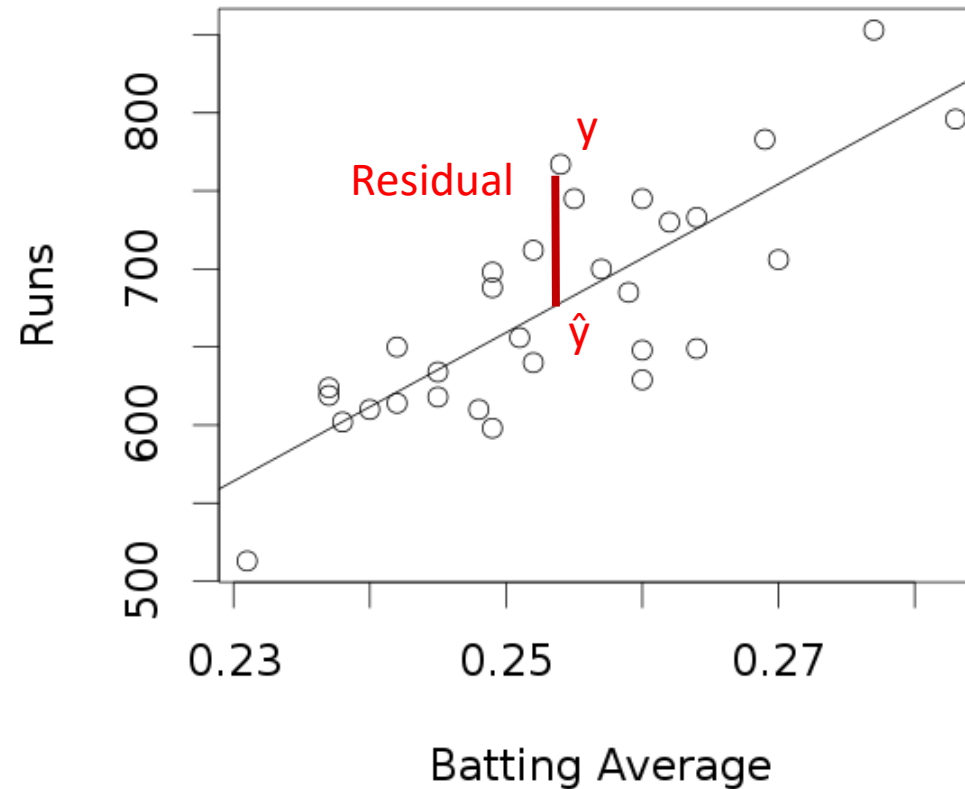
Measuring goodness of fit



We can measure how well the line fits the data using the equation:

$$MSE = \frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2$$

Measuring goodness of fit



See if you can change the slope (a) and intercept (b) to find a line with a small MSE: https://asterius.hampshire.edu:3939/baseball_regression/

Calculating residuals for the runs as a function of batting average

	Obs x (BA)	Runs obs (y) <i>(r in this case)</i>	Runs pred (\hat{y}) <i>(\hat{r} in this case)</i>	Residuals ($y - \hat{y}$) <i>($r - \hat{r}$ in this case)</i>
ARI	.259	685		
ATL	.249	688		
BAL	.260	745		
BOS	.277	853		
CHA	.249	598		

$$\hat{r} = -529.8 + 4755.7 \cdot BA$$

Residuals for the runs as a function of batting average

	Obs x (BA)	Runs obs (y) <i>(r in this case)</i>	Runs pred (\hat{y}) <i>(\hat{r} in this case)</i>	Residuals ($y - \hat{y}$) <i>($r - \hat{r}$ in this case)</i>
ARI	.259	685	702.0	
ATL	.249	688		
BAL	.260	745		
BOS	.277	853		
CHA	.249	598		

$$\hat{r} = -529.8 + 4755.7 \cdot BA$$

Residuals for the runs as a function of batting average

	Obs x (BA)	Runs obs (y) <i>(r in this case)</i>	Runs pred (\hat{y}) <i>(\hat{r} in this case)</i>	Residuals ($y - \hat{y}$) <i>($r - \hat{r}$ in this case)</i>
ARI	.259	685	702.0	-17.0
ATL	.249	688		
BAL	.260	745		
BOS	.277	853		
CHA	.249	598		

$$\hat{r} = -529.8 + 4755.7 \cdot BA$$

Residuals for the runs as a function of batting average

	Obs x (BA)	Runs obs (y) <i>(r in this case)</i>	Runs pred (\hat{y}) <i>(\hat{r} in this case)</i>	Residuals ($y - \hat{y}$) <i>($r - \hat{r}$ in this case)</i>
ARI	.259	685	702.0	-17.0
ATL	.249	688		
BAL	.260	745		
BOS	.277	853		
CHA	.249	598		

Fill in the predicted values (\hat{y}) and the residuals ($y - \hat{y}$) for ATL, BAL, BOS and CHA

$$\hat{r} = -529.8 + 4755.7 \cdot BA$$

Residuals for the runs as a function of batting average

	Obs x (BA)	Runs obs (y) <i>(r in this case)</i>	Runs pred (\hat{y}) <i>(\hat{r} in this case)</i>	Residuals ($y - \hat{y}$) <i>($r - \hat{r}$ in this case)</i>
ARI	.259	685	702.0	-17.0
ATL	.249	688	654.4	33.6
BAL	.260	745	706.7	38.3
BOS	.277	853	787.6	65.4
CHA	.249	598	654.4	-56.4

$$\hat{r} = -529.8 + 4755.7 \cdot BA$$

Sum of squared residuals

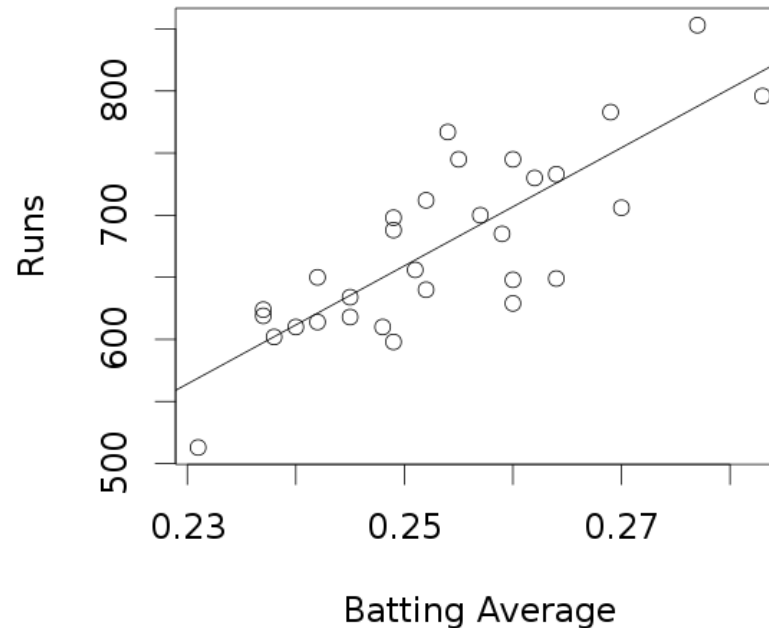
	Runs obs (y)	Runs pred (\hat{y})	Residuals (y - \hat{y})	Residuals² (y - \hat{y})²
ARI	685.0	702.0	-17.0	287.5
ATL	688.0	654.4	33.6	1129.0
BAL	745.0	706.7	38.3	1465.9
BOS	853.0	787.6	65.4	4282.4
CHA	598.0	654.4	-56.4	3181.0

Taking the average of these deviations yields the Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2$$

Root Mean Squared Error

The square root of the MSE (RMSE) gives a sense of how much a points typically differ from the regression line

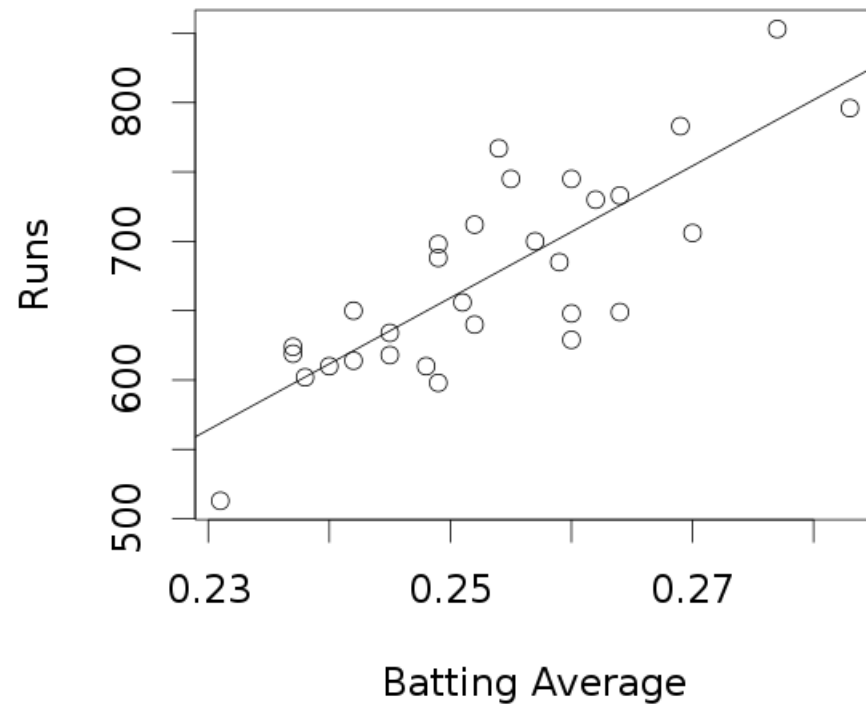


RMSE for predicting runs from BA is 51.35

- i.e., predictions typically off by 51 runs

Question

Where did the regression line come from again?



$$a = \bar{y} - b \cdot \bar{x}$$

$$b = r \cdot \frac{s_y}{s_x}$$

Where:

\bar{y} and \bar{x} are the means of the y's and x's

s_y and s_x are standard deviations of the y's and x's

r is the correlation coefficient

Least squares line

The **least squares line**, also called “**the line of best fit**”, is the line which minimizes the sum of squared residuals

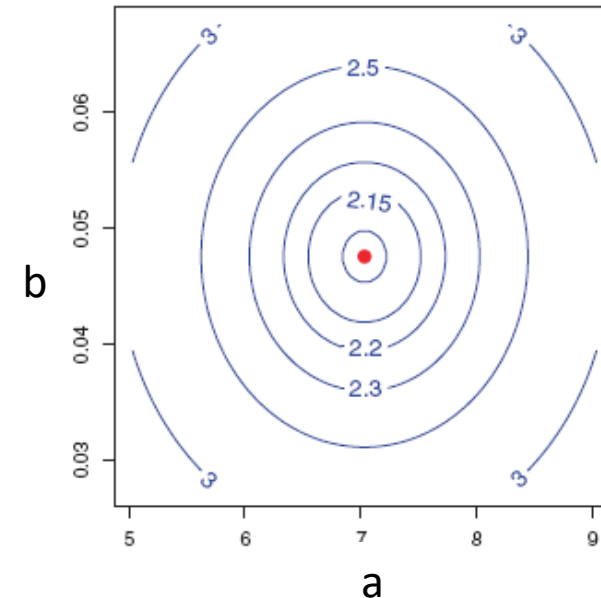
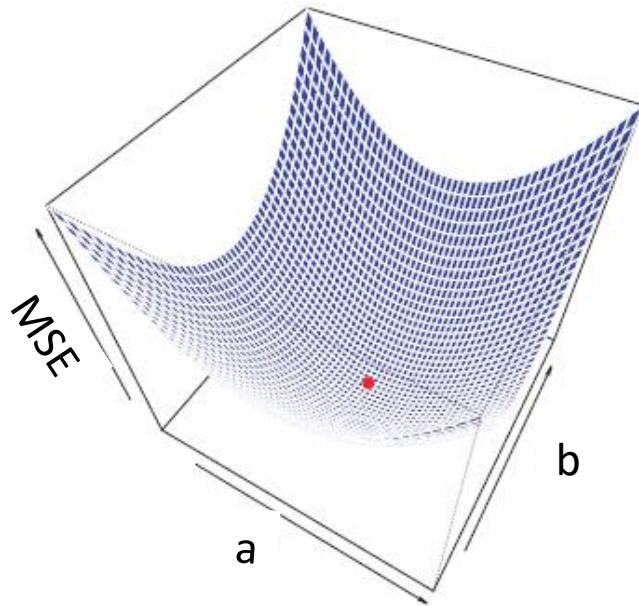
i.e., the least squares line are the coefficients a , and b that minimize the Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2$$

Finding the coefficients a and b

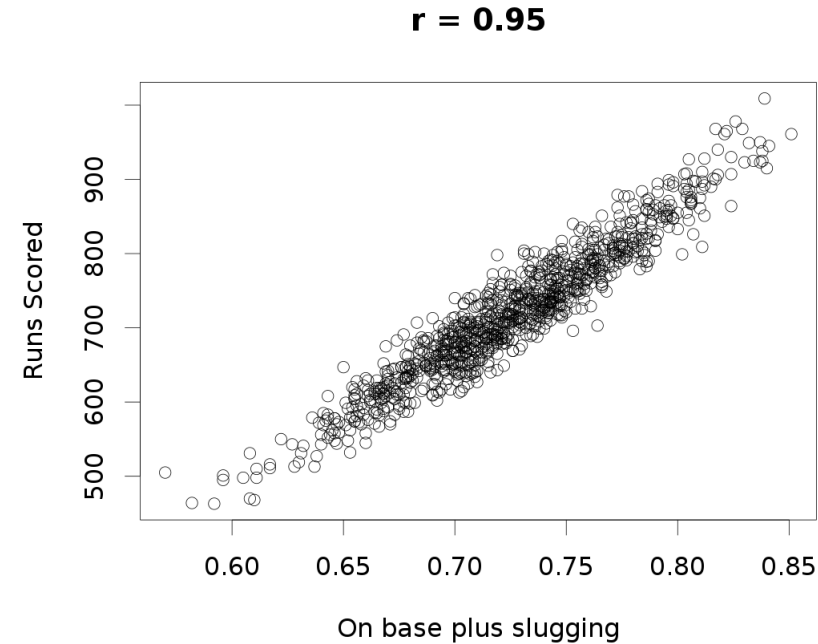
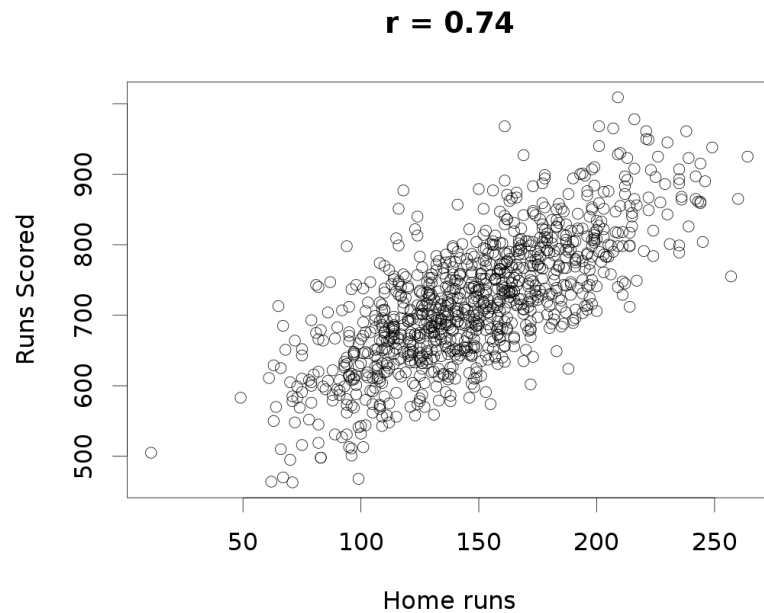
The regression coefficients can be found using calculus:

- Find a and b to minimize the mean squared error (MSE)
- This can be done by setting the partial derivative of the MSE with respect for a and b to 0 and solving for a and b



Compare batting measures based on RMSE

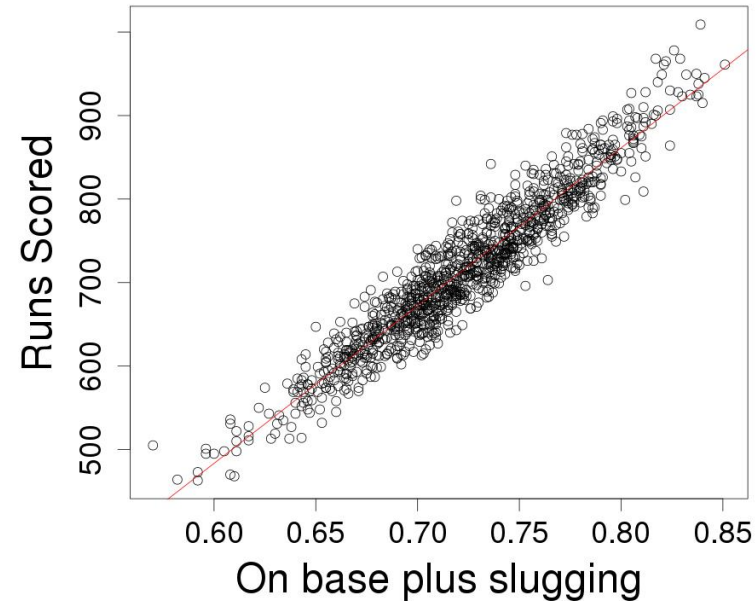
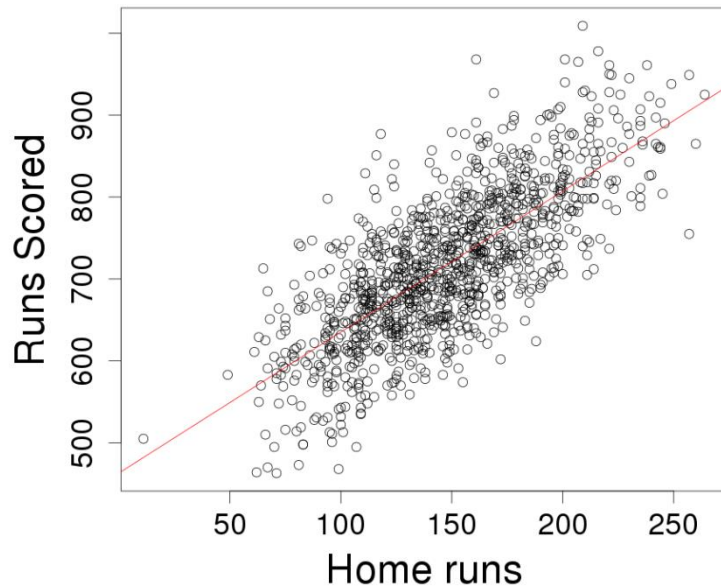
On homework 8 you found the 'best' statistic based on the correlation between each statistic and runs scored



Compare batting measures based on RMSE

We can also find the ‘best’ statistic based on the root mean squared error (RMSE)

- i.e., which statistic leads to a model with the minimal squared residuals



Compare batting measures based on RMSE

	RMSE
HR	60.76
BA	
OBP	
Slug	
OPS	

Compare batting measures based on RMSE

	RMSE
HR	60.76
BA	51.35
OBP	
Slug	
OPS	

Compare batting measures based on RMSE

	RMSE
HR	60.76
BA	51.35
OBP	39.74
Slug	36.95
OPS	27.46

Compare batting measures based on RMSE

	RMSE	r
HR	60.76	0.74
BA	51.35	0.82
OBP	39.74	0.90
Slug	36.95	0.91
OPS	27.46	0.95

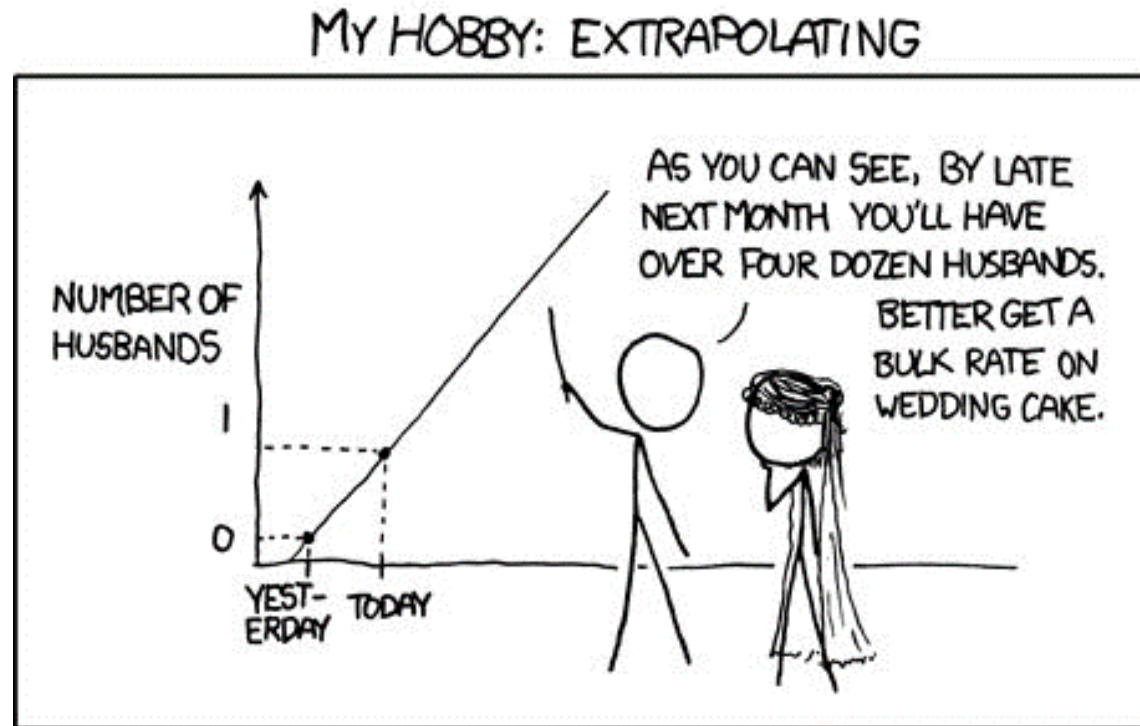
$$r^2 = 1 - \text{MSE}/\text{var}(y) \cdot [(n-1)/n]$$

Note: the variance function in numpy divides by n not n -1 so the relationship using numpy is:

$$r^2 = 1 - \text{MSE}/\text{var}(y)$$

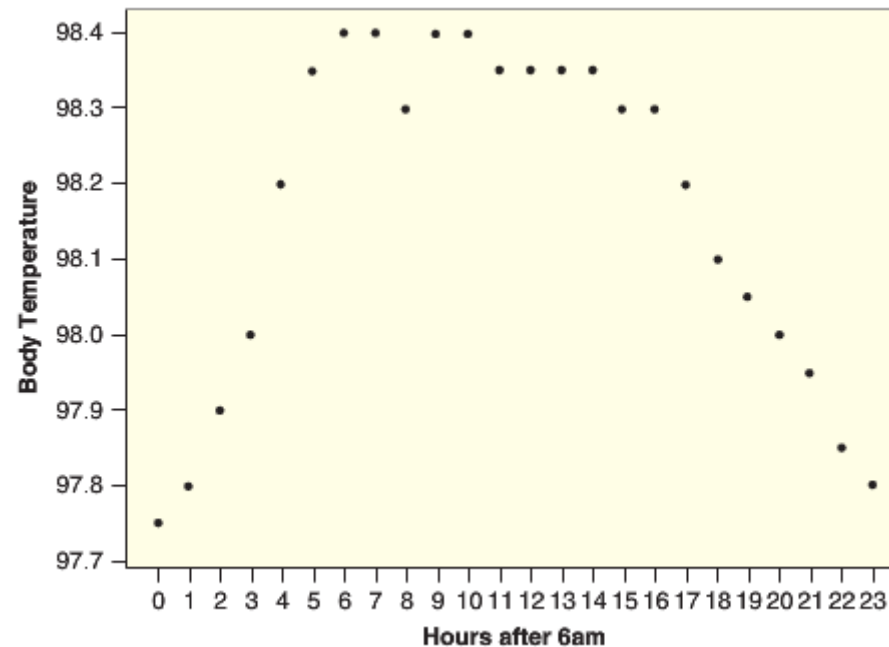
Regression caution # 1

Avoid trying to apply the regression line to predict values far from those that were used to create the line. i.e., do not extrapolate too far



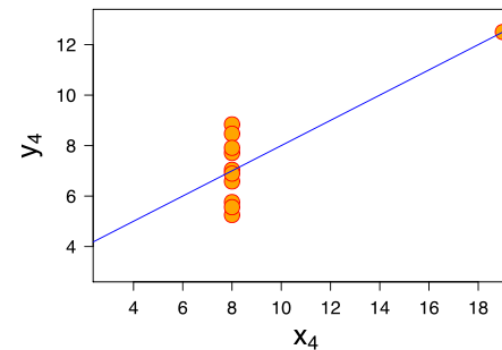
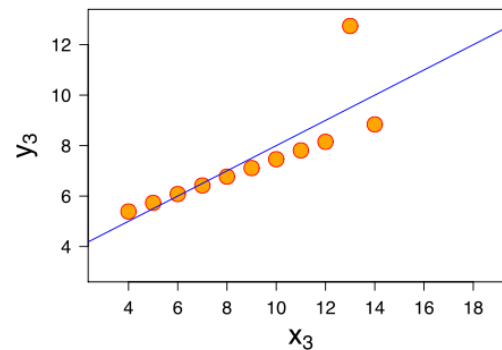
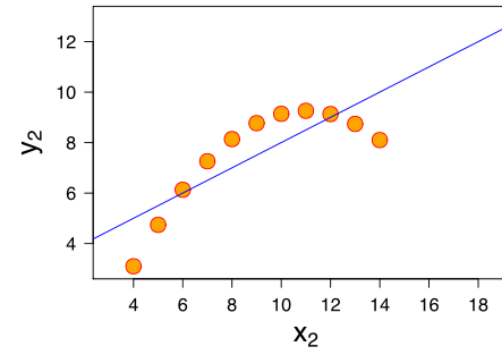
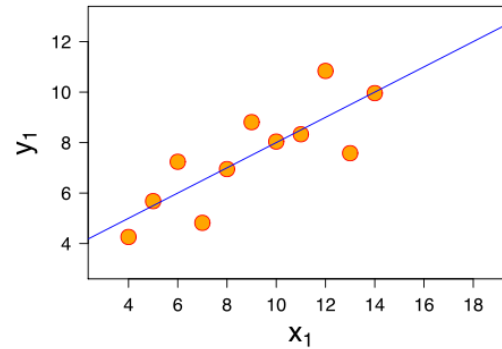
Regression caution # 2

Plot the data! Regression lines are only appropriate when there is a linear trend in the data



Regression caution #3

Be aware of outliers – they can have an huge effect on the regression line.



Back to: who is better Ortiz vs. Jeter?



Derek Jeter



David Ortiz

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’

Batting average:

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

Slugging percentage:

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

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 ?

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

We can use multiple regression, to find these optimal b_i 's !

Multiple regression

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

For multiple linear regression our equation has the form of:

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

We estimate coefficients using a data set to make predictions \hat{y}

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

Multiple regression

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

There are many uses for multiple regression models including:

- To make predictions as accurately as possible
- To understand which predictor variables are related to the response variable
- To create new statistics (“metrics”) that give a useful numerical description of a phenomenon



What are the optimal weights?

Any ideas for the best b_i 's ?

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

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

```
lm = smf.ols('R ~ BB + HBP + H + X2B + X3B + HR', data = teams_2013).fit()
```

```
the_params = lm.params
```

What are the optimal weights?

	b_i
(Intercept)	-497.44
HBP	0.42
BB	0.34
X1B	0.56
X2B	0.75
X3B	1.40
HR	1.44

`lm.params`

Do these coefficients
make sense?

Can you write this in the form of an equation?

What are the optimal weights?

	b_i
(Intercept)	-497.44
HBP	0.42
BB	0.34
X1B	0.56
X2B	0.75
X3B	1.40
HR	1.44

$$\hat{r} = .34 \cdot BB + .42 \cdot HBP + .56 \cdot 1B + .75 \cdot 2B + 1.40 \cdot 3B + 1.44 \cdot HR - 497.44$$

How low can you go?

Can you come create additional variables in the team_batting Table that will lead to a statistic with even higher R^2 ?