Linear regression continued

Outline for today

Better know a player: Ken Griffey Jr.

Review and continuation of linear regression

Better know a player

Ken Griffey Jr.

Thoughts on Big Data Baseball?

Please keep up with the reading and the quote and reactions

We will have a brief discussion of the book next class

Meetings on Friday

If anyone would like to meet with me on Friday to discuss anything related to the class, please sign up for a time here:

https://emeyers.youcanbook.me

Also, midterm self-evaluations were due yesterday so please complete them if you have not done so already

Review

How to create z-score in R

Z-score formula is? $$z_i = \frac{x_i - \bar{x}}{s}$$

Suppose we have two data frames

- Jeter.data # Derek Jeter's yearly data
- PA500.Jeter7th # Data from all players in Jeter's 7th season

How do we create a z-score for Jeter's 7th season BA?

1. BA.Jeter.7th <- Jeter.data[7,]$BA
2. mean.all.players.7th <- mean(PA500.Jeter7th$BA)
3. sd.all.players.7th <- sd(PA500.Jeter7th$BA)
4. zscore.Jeter <- (BA.Jeter.7th - mean.all.players.7th)/
sd.all.players.7th

Regression

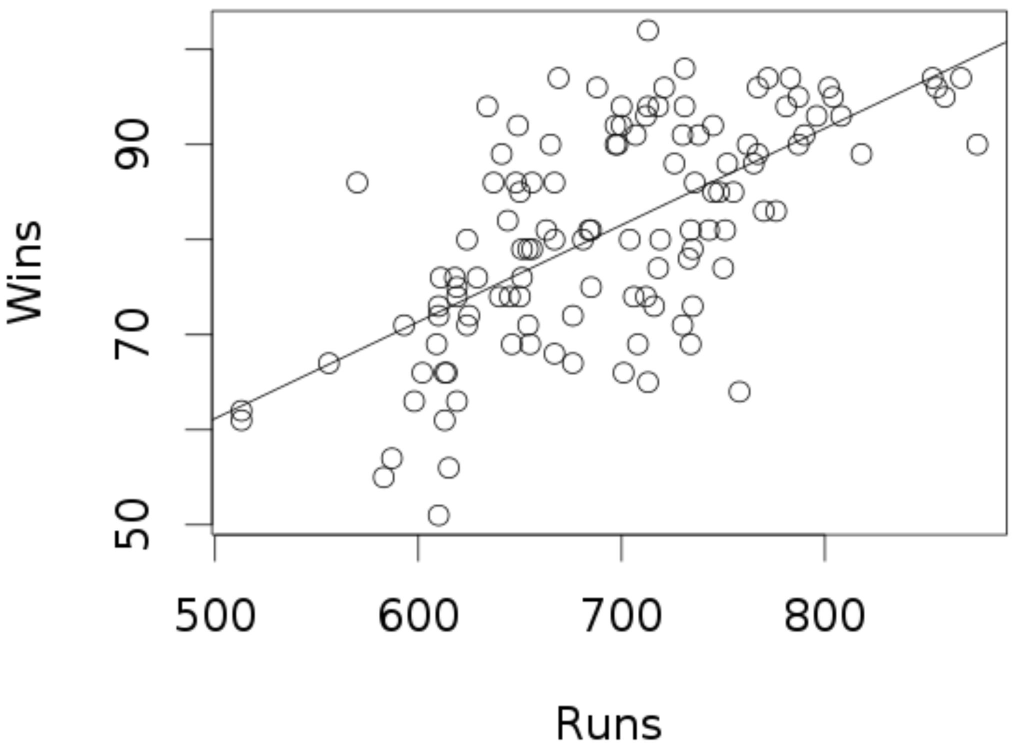
Regression is method of using one variable to predict the value of a second variable

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

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

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

Wins runs regression



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

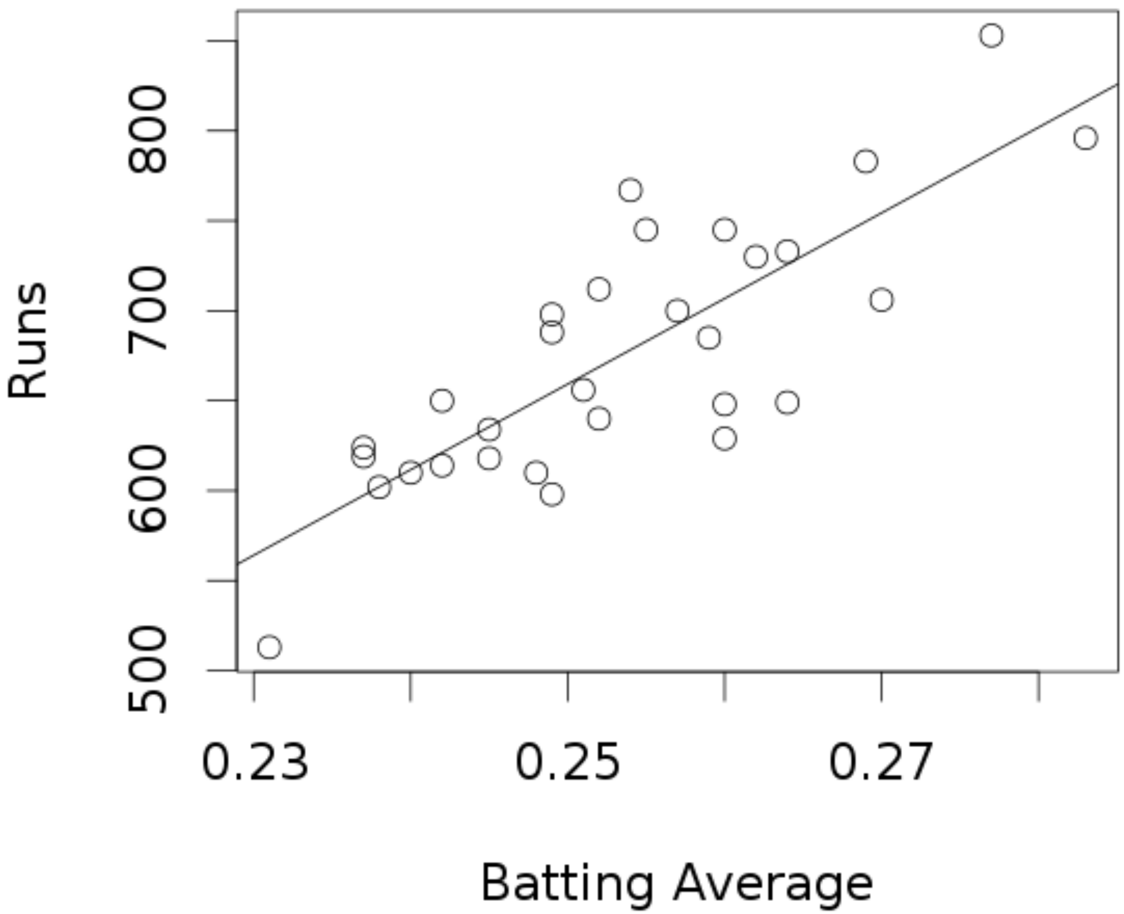
$a = \mathit{14.47}$

$b = \mathit{.088}$

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

1. An additional win for ~11 additional runs scored
   - Now we know the value of a run! It's worth 1/11th of a win!
2. There will be 14.47 wins if you score 0 runs all season

Runs and BA regression



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

$a = -529.8$

$b = 4755.7$

Write the equation for predicting runs as a function of BA

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

Linear regression in R

Link to download data is on Moodle (see Feb 28):

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

We can build a linear model using:

```
fit  <-  lm(y ~ x, data = my_df)
```

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

```
coef(fit)
```

What are y and x for predicting runs from batting average?

y = R and x = BA

See if you can use R to get the linear regression coefficients for predicting runs from batting average and write down the linear regression equation

Linear regression in R

Building the linear model:

```
> fit <- lm(R ~ BA, data = team.batting.162)
```

Get the coefficients:

```
> coef(fit)
(Intercept)          BA
  -786.211    5797.657
```

Writing an equation using these coefficients

$$\hat{r} = -786.2 + 5797.7 \cdot BA$$

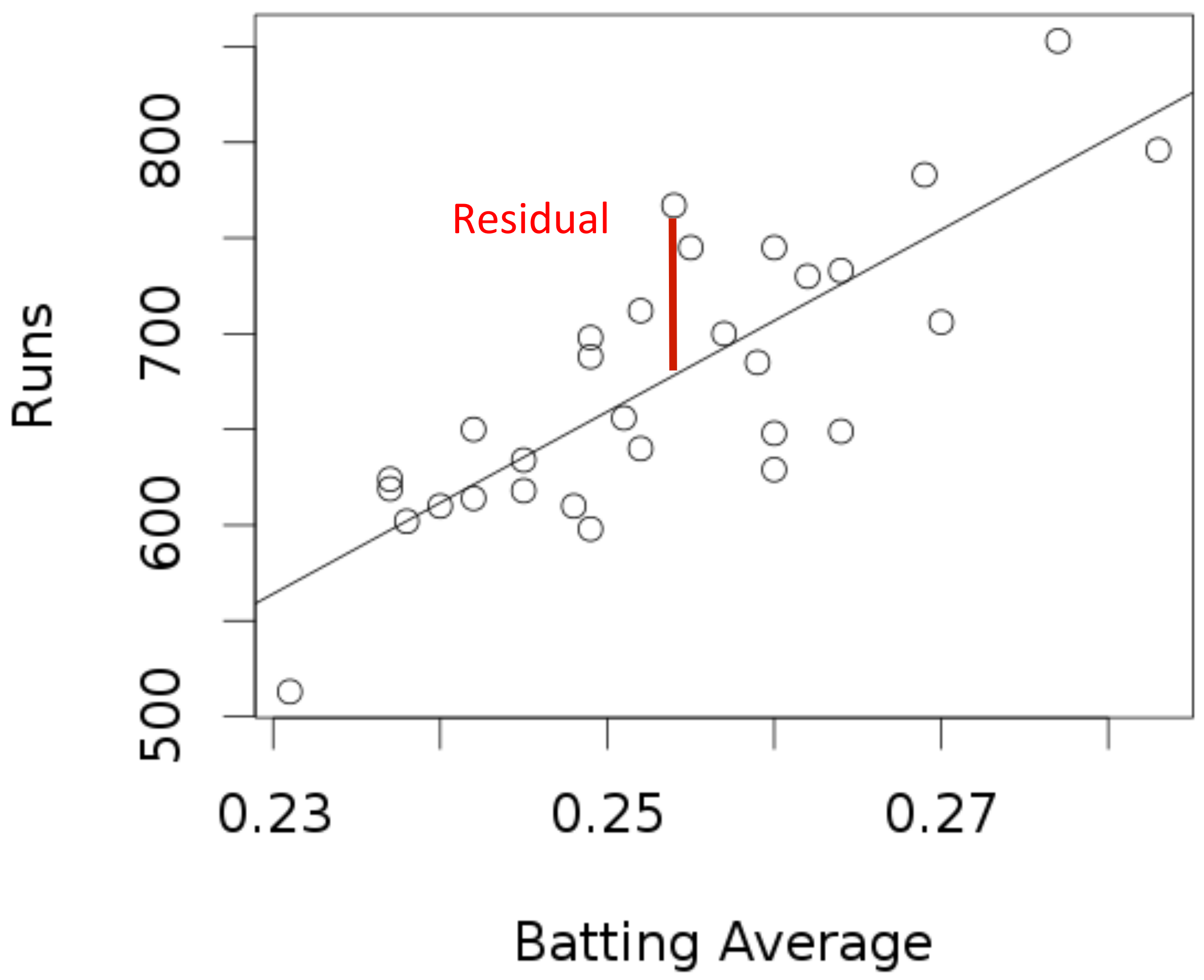
Can you use this equation to predict the amount of run ($\hat{r}$) a team score if they have a particular BA of .300?

Residuals

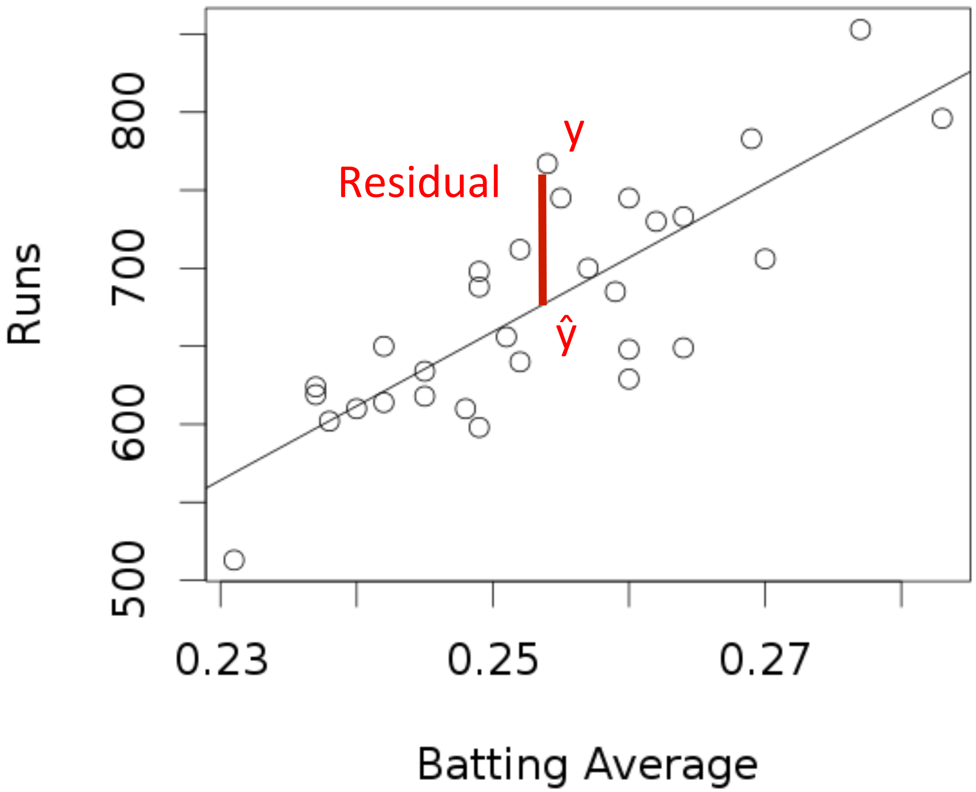
The **residual** at a data value is the difference between the observed (y) and predicted value ($\hat{y}$) of the response variable

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

Run vs. batting average (2013)

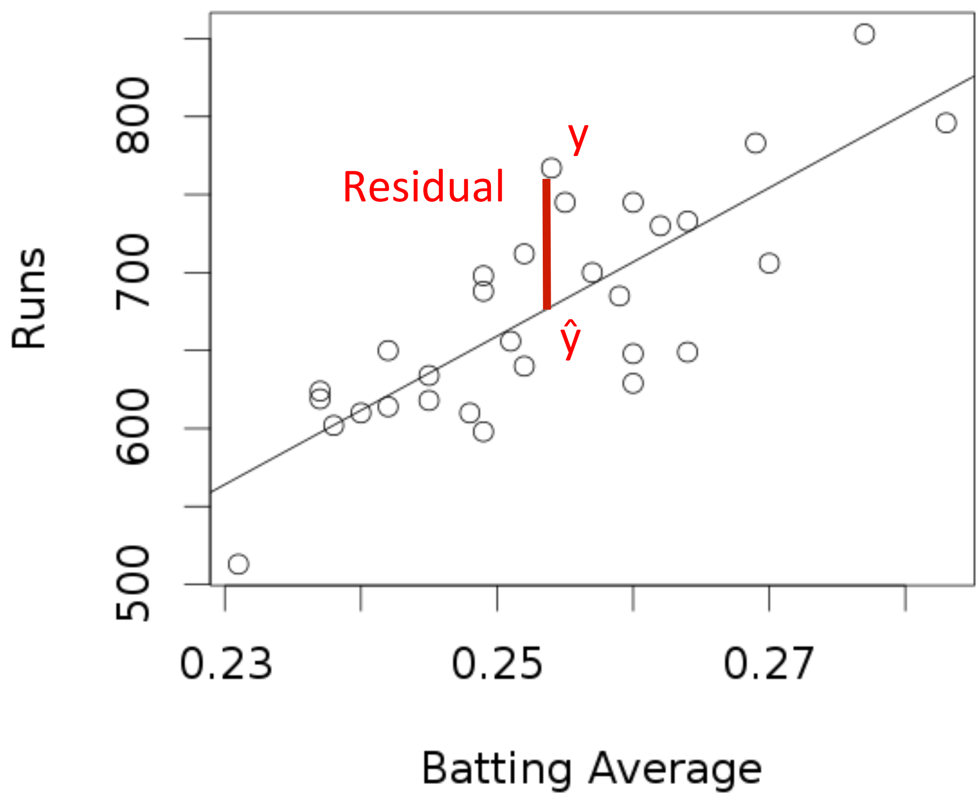


Measuring goodness of fit



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

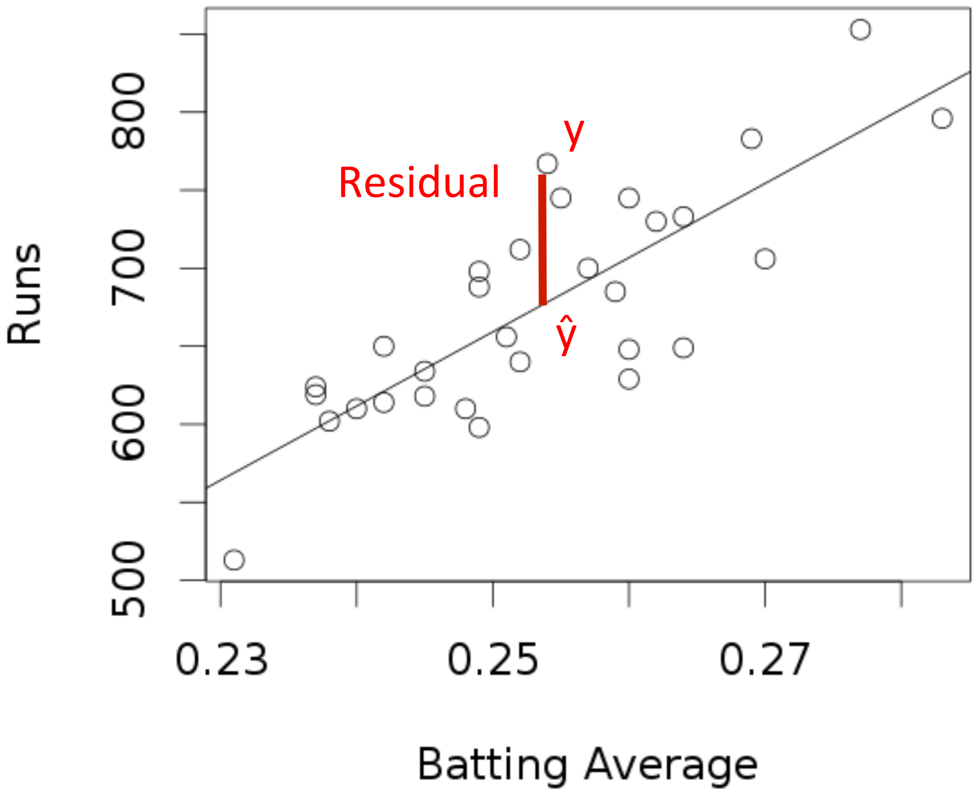
Measuring goodness of fit



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

$$MSE = \frac{1}{n}\sum_i^n(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_i^n(a + b \cdot x_i - y_i)^2$$

Measuring goodness of fit



See if you can change the slope (a) and intercept (b) to find a line with a small MSE (link to the app is on Moodle)

https://emeyers.shinyapps.io/baseball_regression_app/

Calculating residuals for the runs as a function of batting average

	BA	**Runs obs (y)** *(r in this case)*	**Runs pred ($\hat{y}$)** *($\hat{r}$ in this case)*	**Residuals ($y - \hat{y}$)** *($r - \hat{r}$ in this case)*
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

	BA	**Runs obs (y)** *(r in this case)*	**Runs pred ($\hat{y}$)** *($\hat{r}$ in this case)*	**Residuals ($y - \hat{y}$)** *($r - \hat{r}$ in this case)*
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

	BA	**Runs obs (y)** *(r in this case)*	**Runs pred ($\hat{y}$)** *($\hat{r}$ in this case)*	**Residuals ($y - \hat{y}$)** *($r - \hat{r}$ in this case)*
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

	BA	**Runs obs (y)** *(r in this case)*	**Runs pred ($\hat{y}$)** *($\hat{r}$ in this case)*	**Residuals (y - $\hat{y}$)** *(r - $\hat{r}$ in this case)*
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

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

Residuals for the runs as a function of batting average

	BA	**Runs obs (y)** *(r in this case)*	**Runs pred (ŷ)** *(r̂ in this case)*	**Residuals (y - ŷ)** *(r - r̂ in this case)*
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

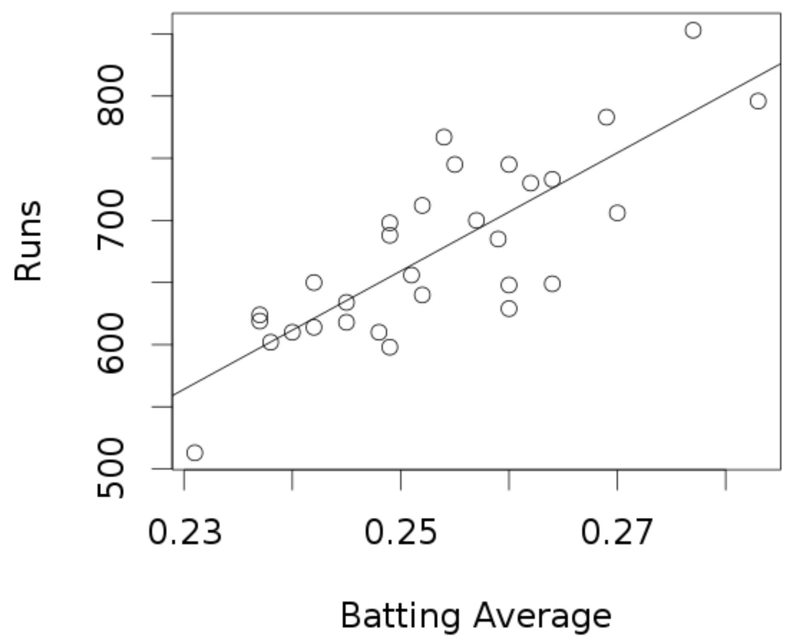
	Runs obs (y)	Runs pred ($\hat{y}$)	Residuals ($y - \hat{y}$)	Residuals^2 $(y - \hat{y})^2$
ARI	685.0	702.0	-17.0	287.5
ATL	688.0	654.4	33.6	1129.0
BAL	745.0	706.7	38.3	1465.9
BOS	853.0	787.6	65.4	4282.4
CHA	598.0	654.4	-56.4	3181.0

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

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

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

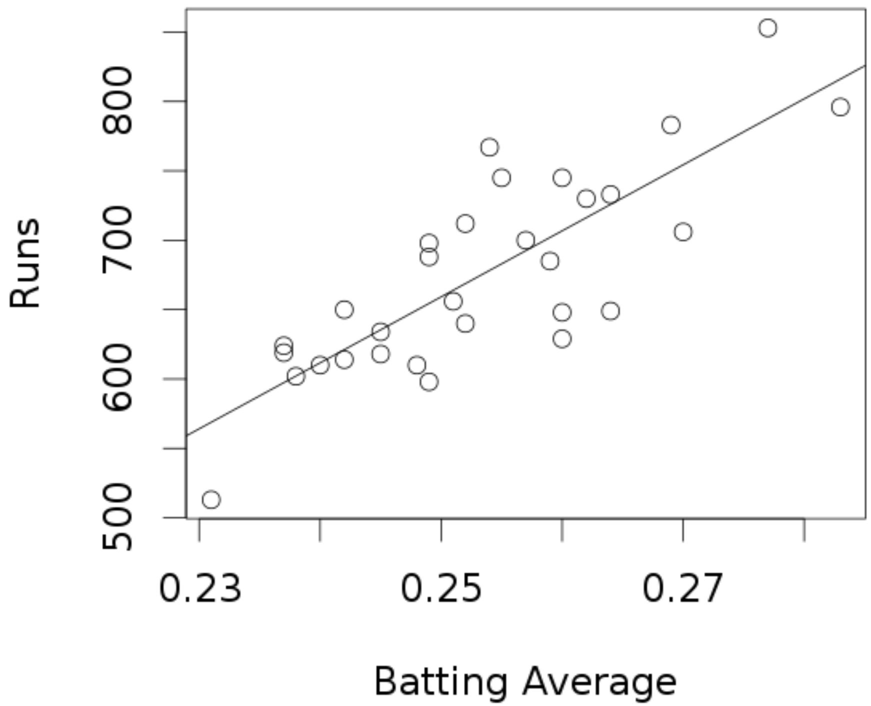


RMSE for predicting runs from BA is 51.35

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

Question

Where did the regression line come from?



Least squares line

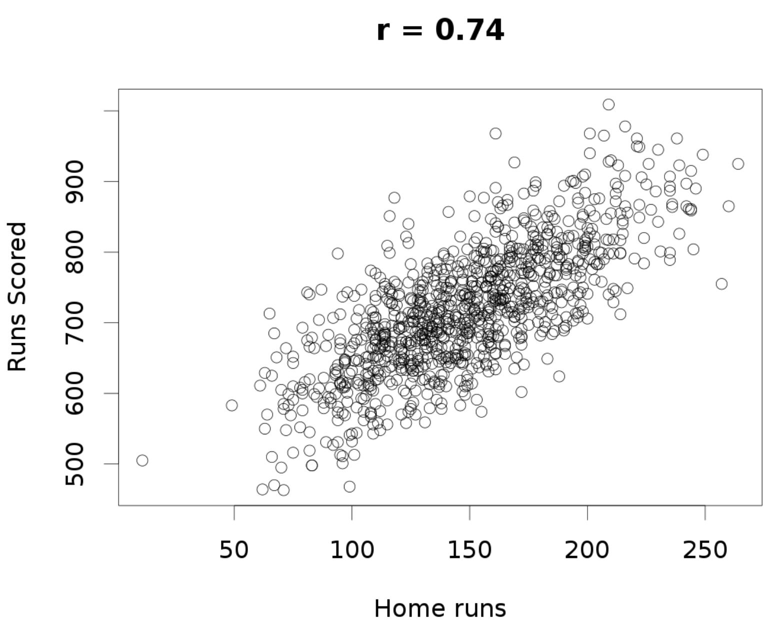
The **least squares line**, also called “**the line of best fit”**, is the line which minimizes the sum of squared residuals

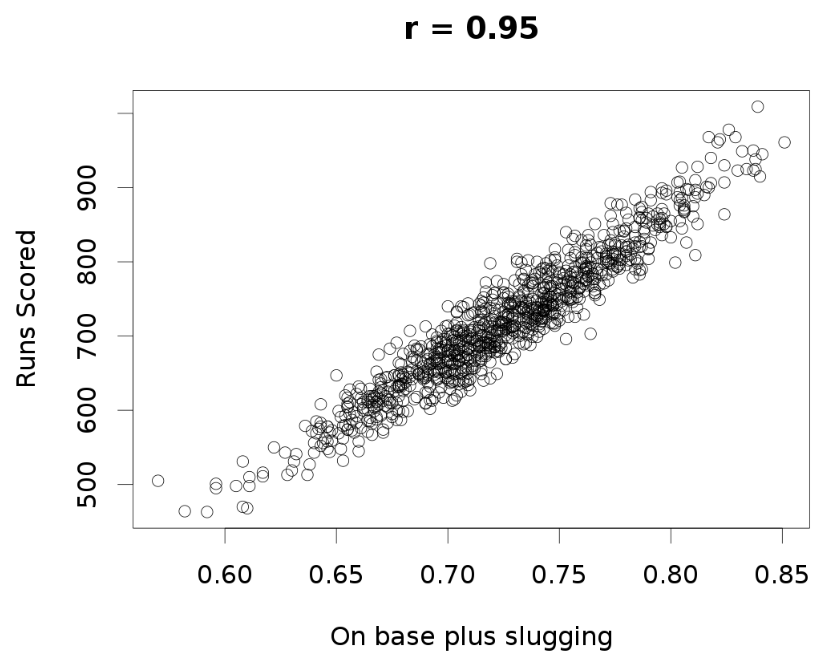
i.e., the least squares line is the line that minimizes the Mean Squared Error (MSE)

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

Compare batting measures based on RMSE

Last class we found the 'best' statistic based on the correlation between each statistic and runs scored

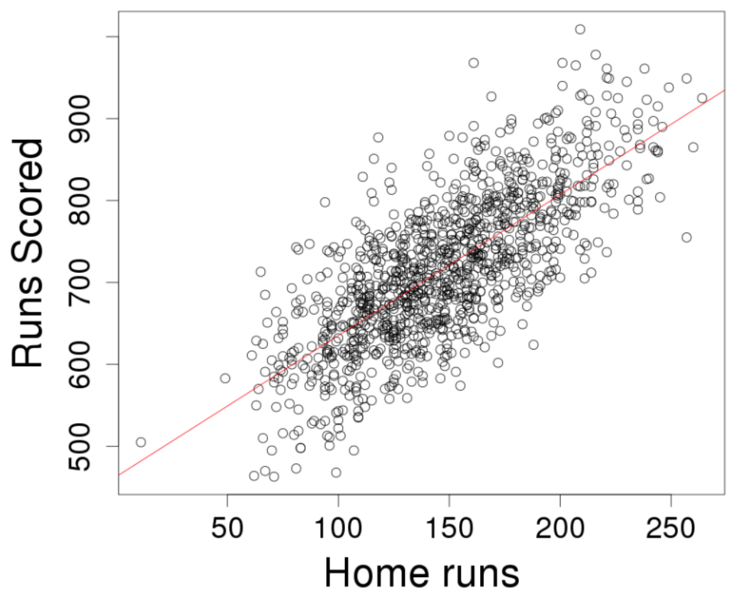


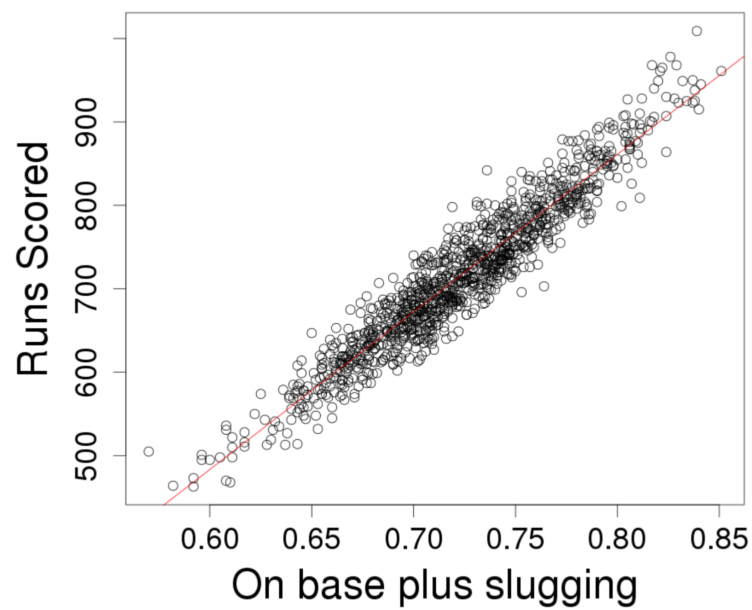


Compare batting measures based on RMSE

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





How to calculate the RMSE in R

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

Let’s calculate the MSE when predicting a team’s Run totals from HR totals

In this case, what are the y_i’s?

How would we get a vector of them using R?

```
> real.runs.y <- team.batting.162$R
```

What is the equation for predicting runs from home runs?

$$\hat{y} = \hat{r} = a + b \cdot HR$$

How can we get the coefficients *a* and *b* to predict runs from HR in R?

```
> fit <- lm(team.batting.162$R ~ team.batting.162$HR)
> coef(fit)
```

How can we make predictions for each point y?

```
> predicted.runs.yhat <- coef(fit)[1] + coef(fit)[2] * team.batting.162$HR
> predicted.runs.yhat <- predict(fit)     # an alternative way
```

How to calculate the RMSE in R

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

Now we have:

- A vector of real runs (y): real.runs.y
- A vector of predicted runs (ŷ): predicted.runs.yhat

How do we calculate the RMSE from these vectors?

1. create a vector of residuals by subtracting one vector from the other

```
residual.vec  <-  real.runs.y - predicted.runs.yhat
```

2. Square the values in this vector of residuals
3. Sum these squared values
4. Divide by n
5. Take the square root

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

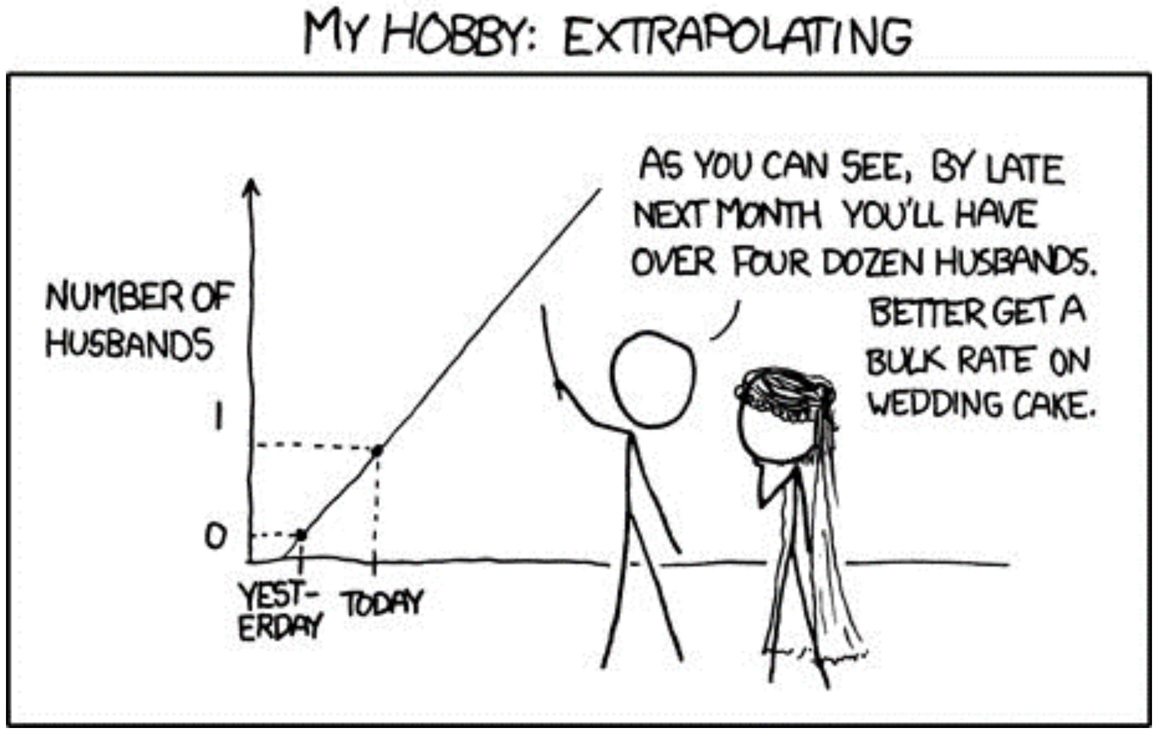
Compare batting measures based on RMSE

	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

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

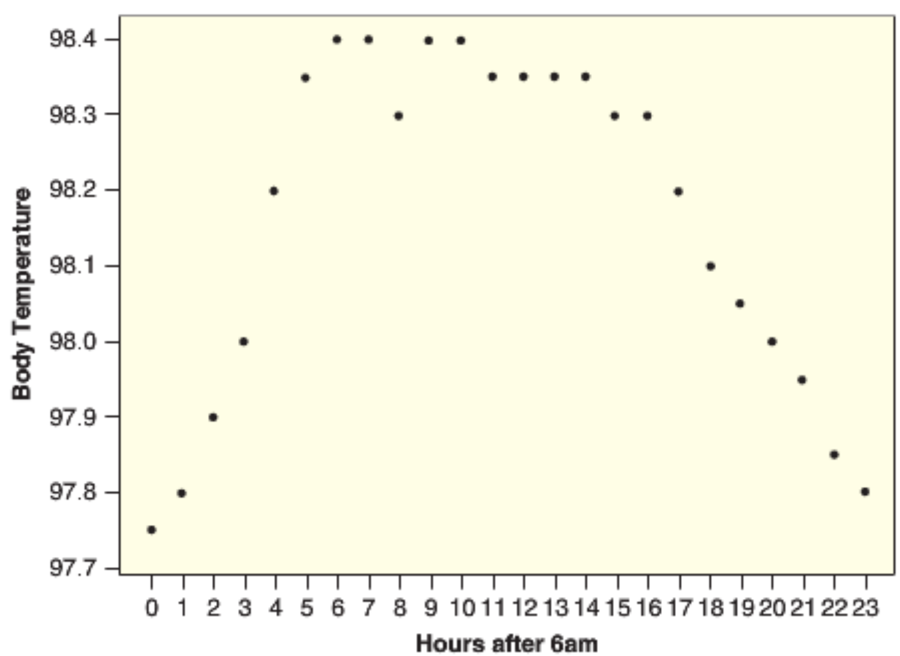
Regression caution # 1

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.

Worksheet 4

```
source('/home/shared/baseball_stats_2017/baseball_class_functions.R')
```

```
get.worksheet(4)
```

As always, worksheet is due on midnight on Sunday

- i.e., 11:59pm on March 5th

Start early and use Piazza and TA office hours!