

Welcome to STA 101!

Statistics is a confrontation with uncertainty.

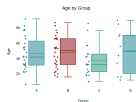
Statistics confronts uncertainty by quantifying it.

# Data analysis

Transforming messy, incomplete, imperfect data into knowledge:

subject	variable_1	variable_2
1	-1.65692830	-2.16524631
2	-0.90396488	-2.97993045
3	1.37141732	0.09720280
4	-0.43176527	0.27970110
5	0.40649190	0.69143221
6	1.47092198	4.47233461
7	-0.78625051	-1.24276055
8	0.64835135	-0.06749005
9	0.06363568	0.33517580

ggplot  
|>



	mean	median	std
Variable 1	0.795	-0.292	0.200
Variable 2	0.343	0.616	0.834
Variable 3	1.587	0.508	2.501

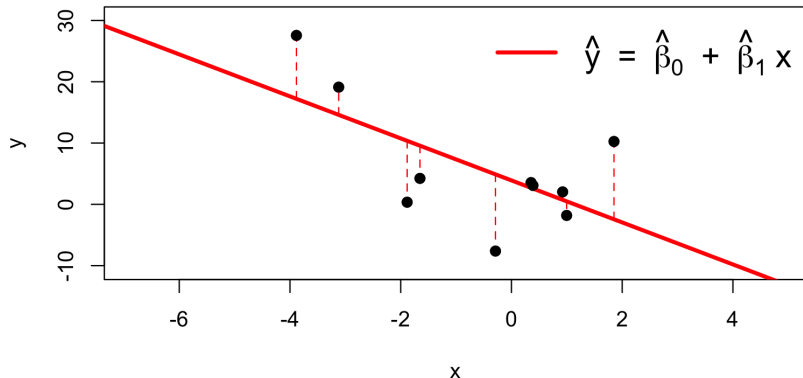
## Statistical inference

Quantifying our uncertainty about that knowledge:

- **Question:** What's the number?
- **Answer:** best-guess  $\pm$  margin-of-error

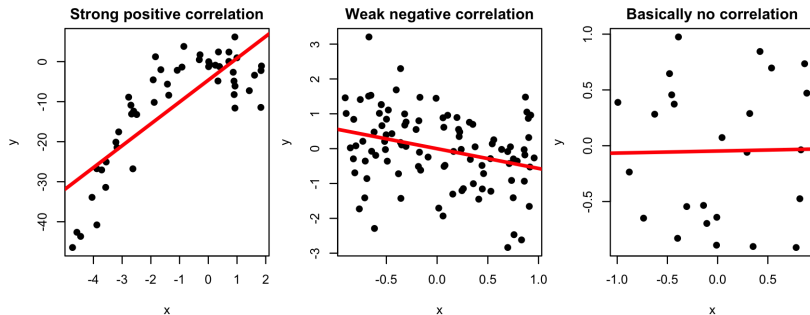
## Recap: simple linear regression

Concisely summarize the observed association between two variables using a line of best fit:



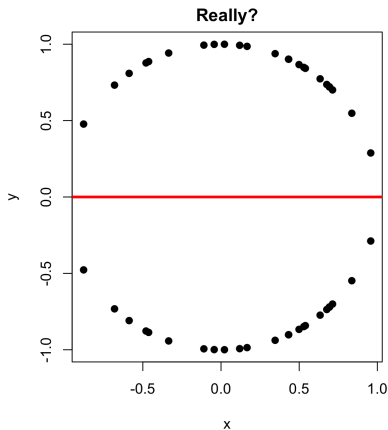
The slope and intercept estimates  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  are chosen to minimize the sum of squared deviations from the line (residuals).

# Interpreting the coefficient estimates



- The sign of  $\hat{\beta}_1$  tells you if the variables are positively or negatively correlated;
- The magnitude of  $\hat{\beta}_1$  tells you how strong the correlation is.

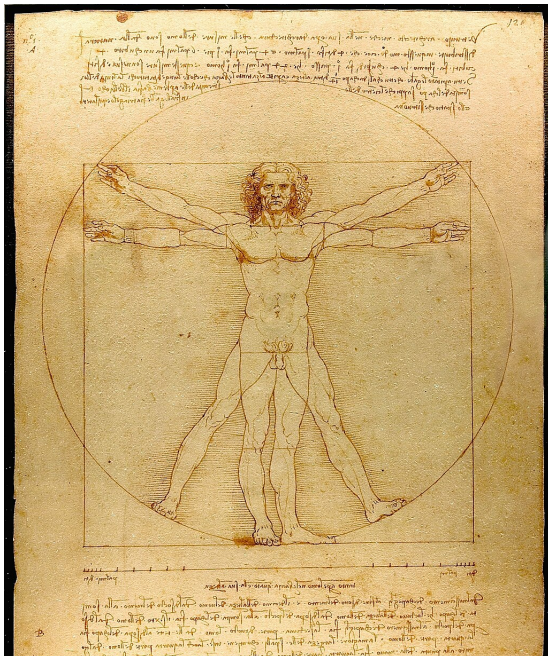
Beware: correlation is not the only kind of association...



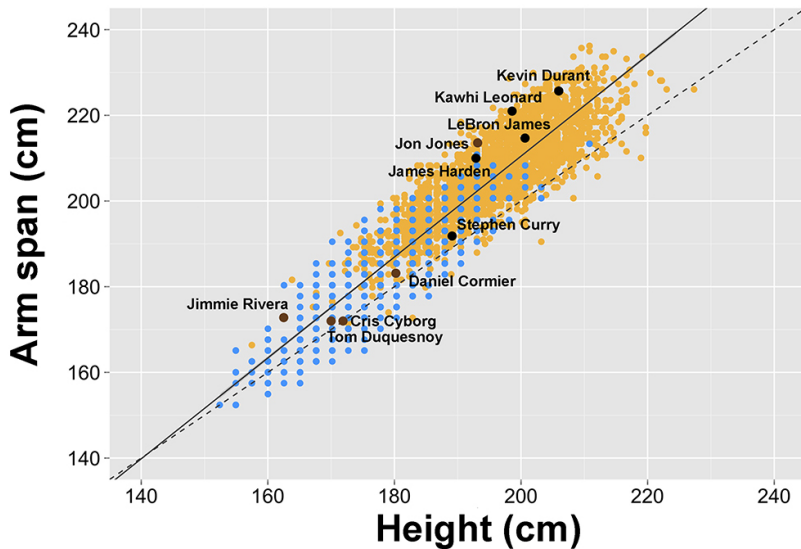
The best fit line says  $x$  and  $y$  are not correlated. Maybe so, but there is clearly *some* association.

**ABV: Always Be Visualizing**

## Example: height vs. wingspan



## Example: height vs. wingspan





## Clarification: notation

**Population:** The “idealized” linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

$\beta_0$ ,  $\beta_1$ , and  $\varepsilon_i$  are unknown. Revealed only with *infinite* data.

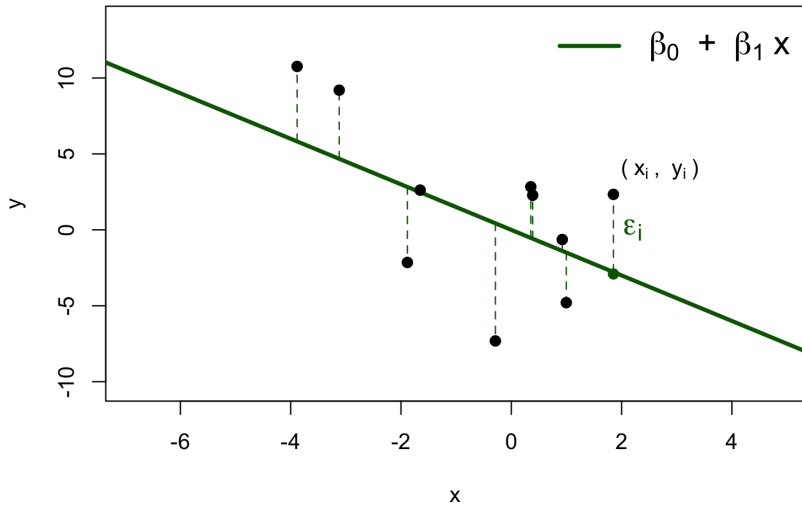
**Sample:** You collect a *finite* sample  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  and calculate the fitted regression model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

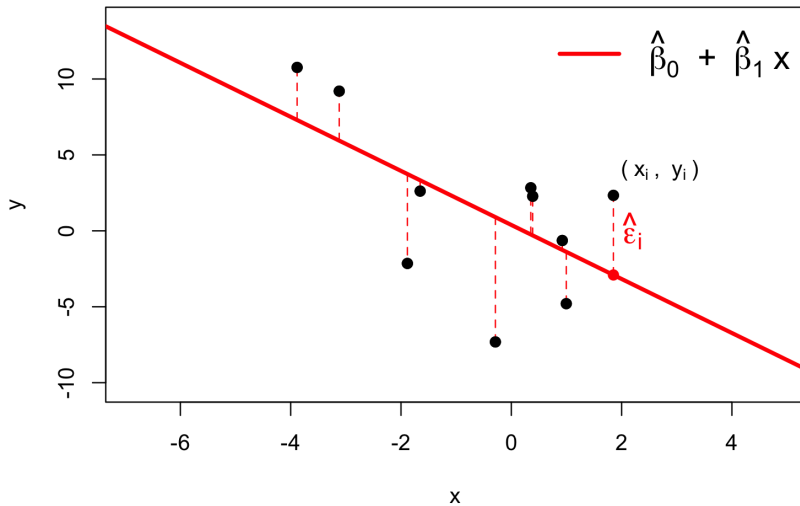
$$\hat{\varepsilon}_i = y_i - \hat{y}_i.$$

**Hopefully:** As your sample size  $n$  gets bigger, your estimates  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  get closer and closer to the ideal, population values  $\beta_0$ ,  $\beta_1$ .

# The population version



## The sample version



## Clarification: why squared error?

Why do we minimize

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

and not

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

or

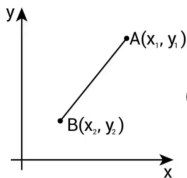
$$\sum_{i=1}^n |y_i - \hat{y}_i| \quad ?$$

Why not  $\sum_{i=1}^n (y_i - \hat{y}_i)$ ?

Because it's trash.

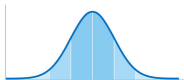
# Why choose squared error?

- Squared error is computationally convenient (take my word for it);
- Squared error is intimately related to the *mean*, while absolute error is intimately related to the *median* (take my word for it);
- Squared error plays nice with the geometry of Euclidean space:



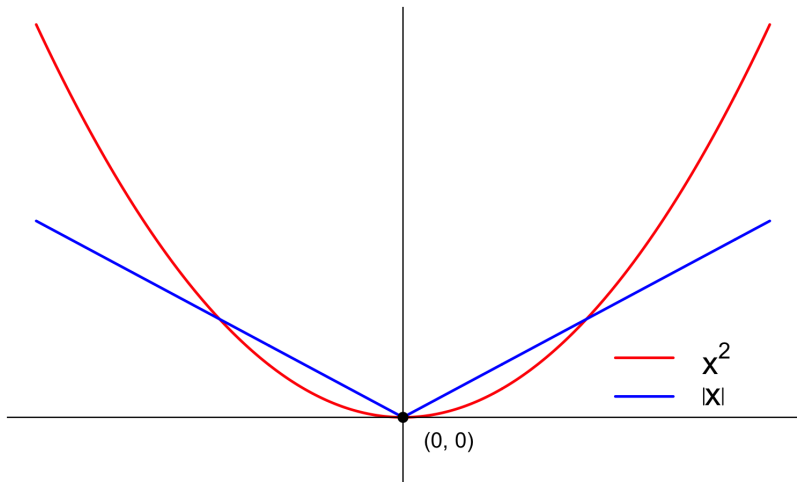
$$d(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- Squared error plays nice with the *bell curve*:



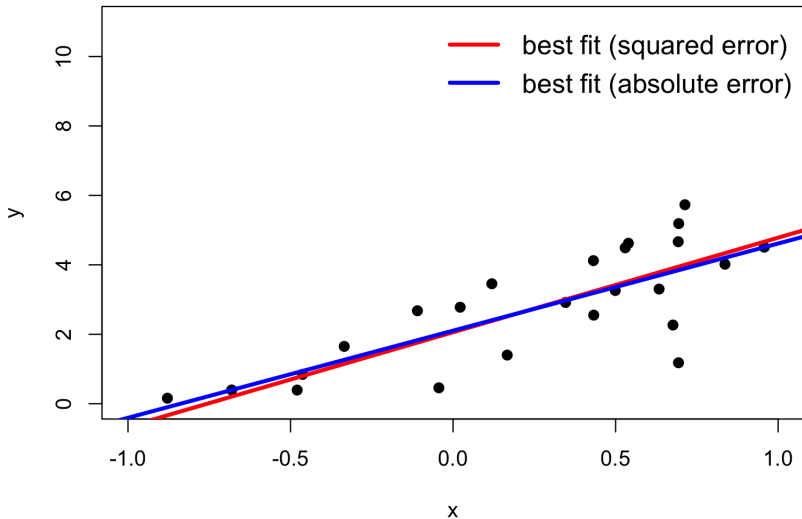
$$p(x) \propto \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

But absolute error is still no joke!



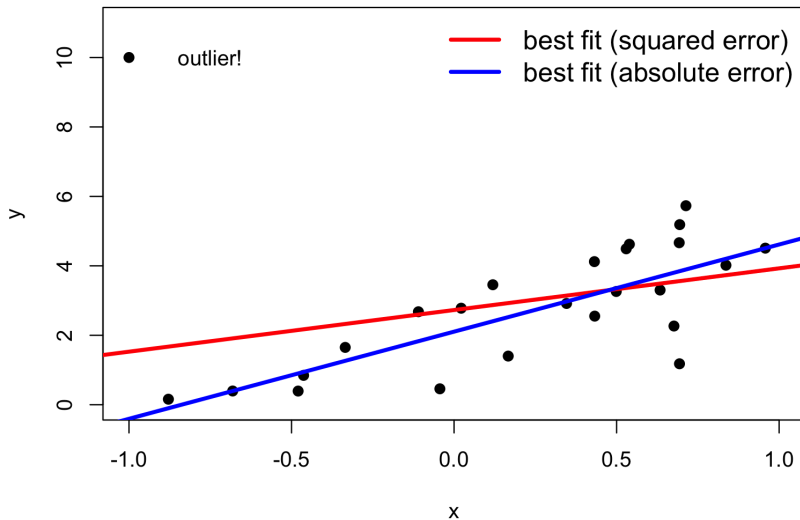
Squared error gives increasingly more weight to data points that are far away from the others (outliers); absolute error is more chill.

“Well-behaved” data: these are basically the same



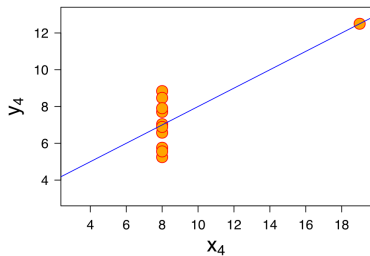
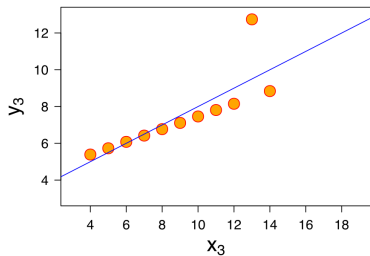
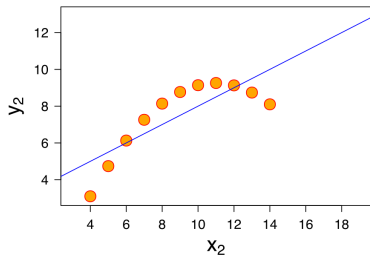
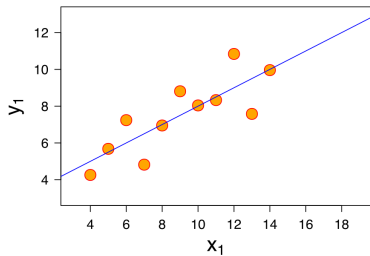


## Add a single outlier...



Regression based on absolute error may be “more robust.”

For the upteenth time...



## Square vs. absolute error: bottom line

*As I said last time, squared error and absolute error are “first class citizens” from a conceptual point of view. They both have pros and cons, and you may prefer one or the other depending on what you are trying to do.*

*But throughout this course (and most courses you might take), we focus on the squared error version.*

## Today's topic

*Multiple* linear regression.

### Goal

Study the association between multiple variables (not just two).

### Subtext

Assess the causal impact of one variable on another *while accounting for other factors*.

### Warning

Association alone does not imply causation.

## Linear regression with two predictors

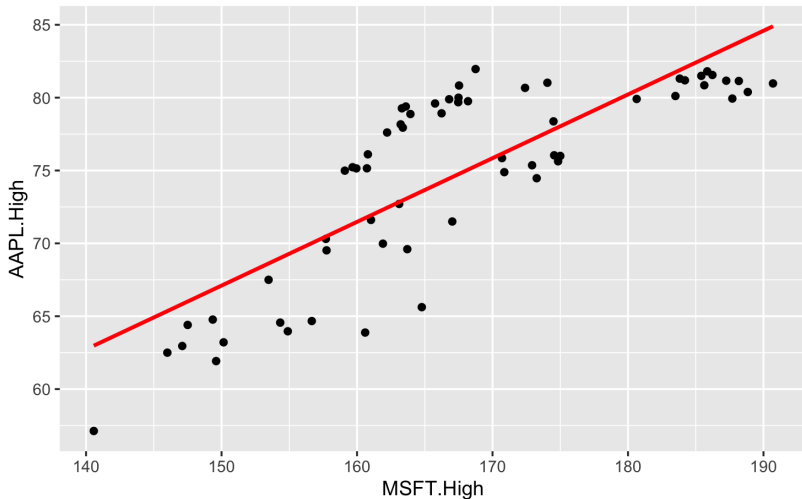
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

- $y$ : *outcome or response variable*;
- $x_1, x_2$ : *predictors, covariates, regressors, features, ...*;
- $\beta_0, \beta_1, \beta_2$ : *coefficients or parameters*;
- $\varepsilon$ : *error or residual*;

This model predicts  $y$  given  $x_1$  and  $x_2$ .

## Recall: stock prices (first quarter of 2020)

$$\widehat{\text{AAPL}} = \underbrace{\hat{\beta}_0}_{1.52} + \underbrace{\hat{\beta}_1}_{0.437} \text{MSFT}$$

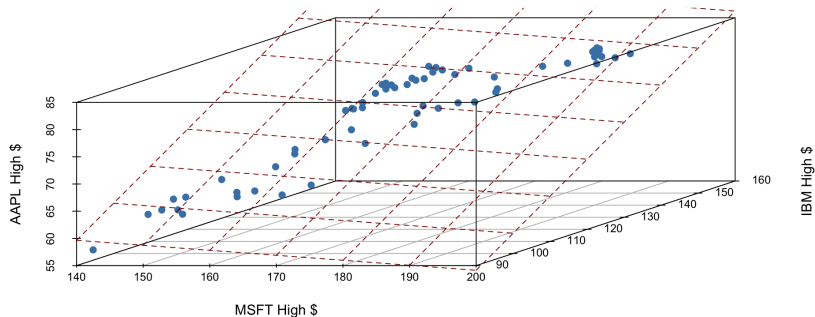


## Include a third stock (IBM)

Model:

$$\widehat{\text{AAPL}} = \underbrace{\hat{\beta}_0}_{31.24} + \underbrace{\hat{\beta}_1}_{-0.091} \text{MSFT} + \underbrace{\hat{\beta}_2}_{0.458} \text{IBM}$$

2 predictors + 1 outcome = 3 dimensions:



The line of best fit becomes a **plane** of best fit. Already hard to visualize. Becomes impossible in higher dimensions.

# Multiple linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_p + \varepsilon.$$

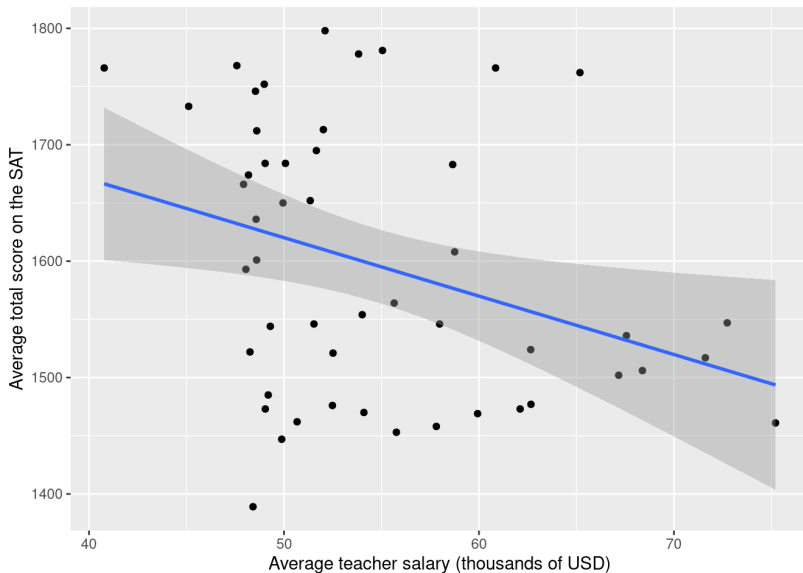
- $y$ : *outcome or response* variable;
- $x_1, x_2, \dots, x_p$ : *predictors, covariates, regressors, features, ...*;
- $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ : *coefficients or parameters*;
- $\varepsilon$ : *error or residual*;

This model predicts  $y$  given  $x_1, x_2, x_3, \dots, x_p$ .

The “concise” numerical summary is a **hyperplane** of best fit, which human beings cannot visualize.



# Why do we need more predictors?



## Why do we need more predictors?

