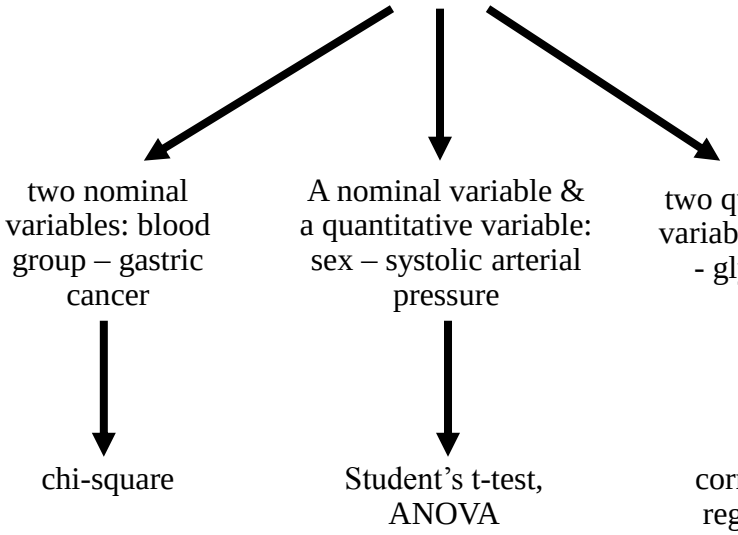
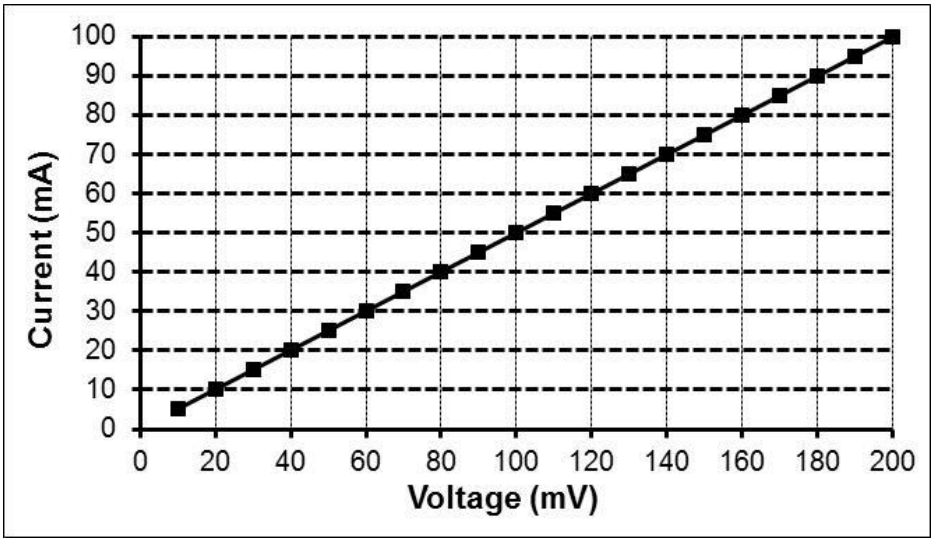
Simple linear regression

- Prof. Giuseppe Verlato
- Unit of Epidemiology & Medical Statistics, Dept. of Diagnostics & Public Health, University of Verona

Statistics with two variables

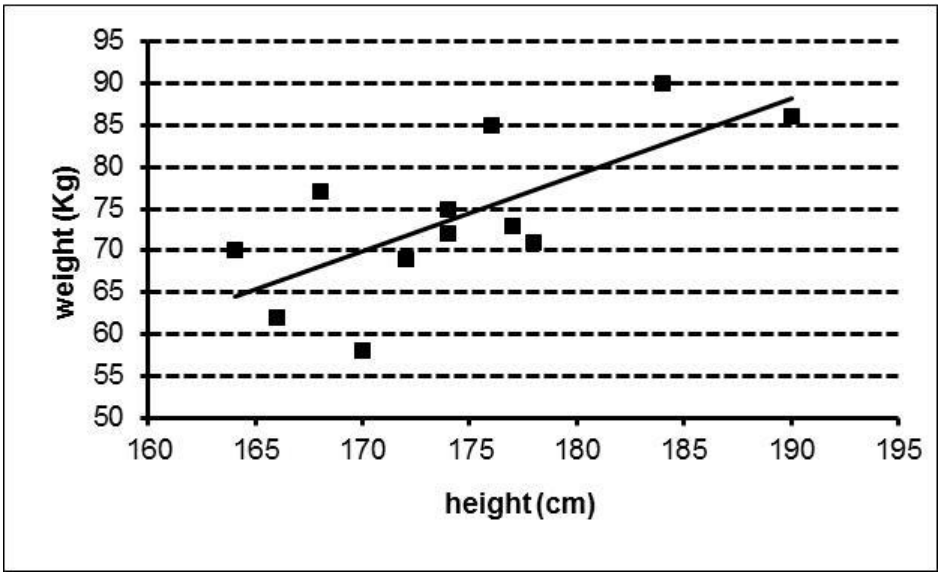


PERFECT correlation between variable X (electric potential difference V) and variable Y (current): 1st Ohm's law



Conductance = $\Delta I / \Delta V$

There are no PERFECT relations in Medicine, as a variable Y is affected not only by a variable X, but also by several other variables, mostly unknown (the so-called biologic variability)

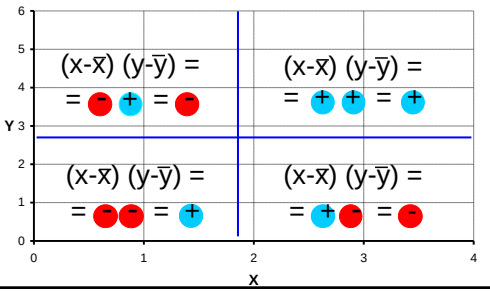


Measures of variability

Univariate statistics → Sum of squares

Bivariate statistics → Sum of products

	Heuristic equation	Empirical equation	
Sum of squares	$\Sigma(x-\bar{x})^2$	$\Sigma x^2-(\Sigma x)^2/n$	always>=0
Sum of products	$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma xy-(\Sigma x*\Sigma y)/n$	<0, =0, >0



Measures of variability

Univariate statistics → variance = SSq / (n-**1**)

Bivariate statistics → covariance=sum of products/(n-**2**)

$$\hat{COV}[X, Y] = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n - 2}$$

2 random samples of size *n*

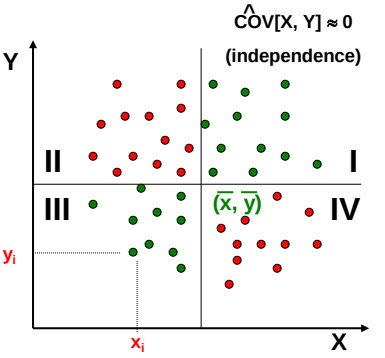
⇒ *n* observations: $(x_i - \bar{x})(y_i - \bar{y})$

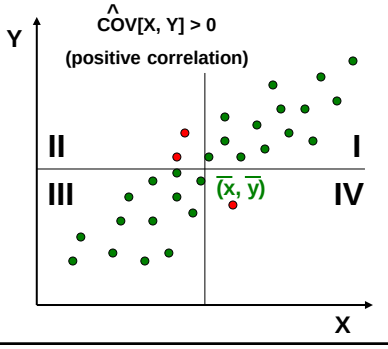
I quadrant - *positive values*

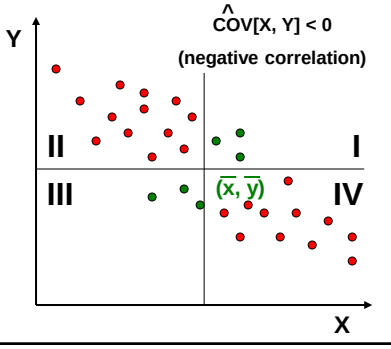
II quadrant - *negative values*

III quadrant – *positive values*

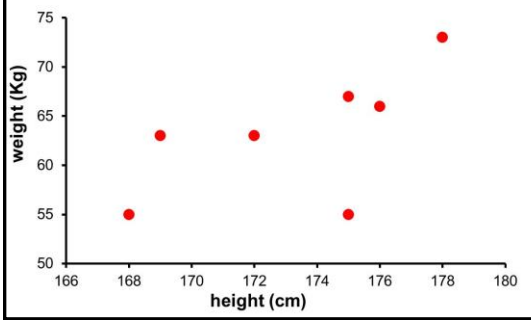
IV quadrant – *negative values*







	height (cm)	weight (Kg)	xy	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
	172	63	10836	-1,3	-0,1	0,2
	178	73	12994	4,7	9,9	46,5
	175	67	11725	1,7	3,9	6,6
	175	55	9625	1,7	-8,1	-14,0
	176	66	11616	2,7	2,9	7,8
	169	63	10647	-4,3	-0,1	0,6
	168	55	9240	-5,3	-8,1	43,0
Total	1213	442	76683			90,7
Mean	173,3	63,1				



Sum of products =

$\Sigma(x-\bar{x})(y-\bar{y}) = 90.7$

$\Sigma xy - (\Sigma x\ \Sigma y)/\ n =$

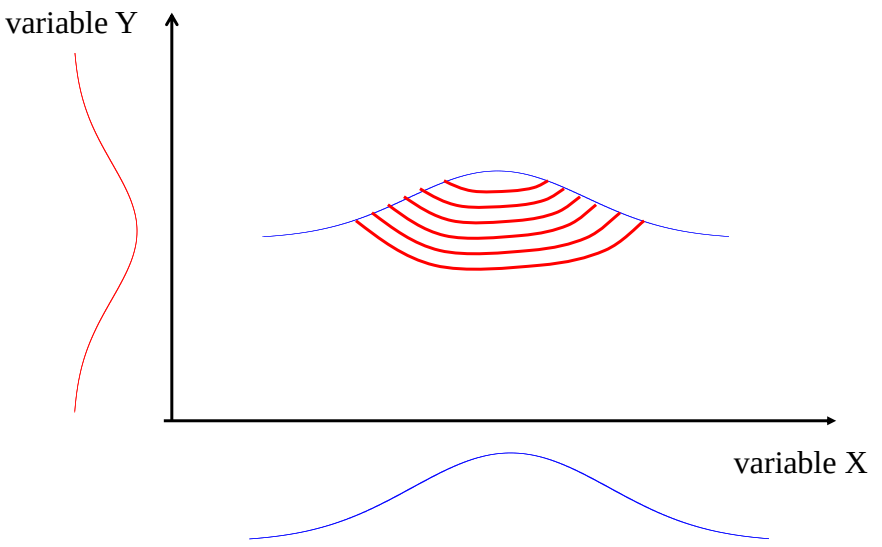
$76683 - 1213*442/7 =$

$76683 - 76592.3 = 90.7$

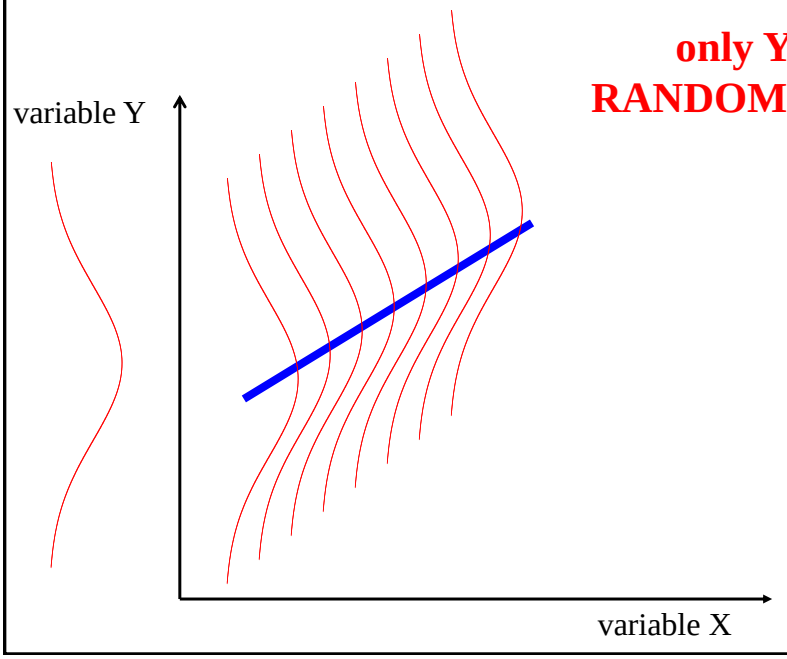
Regression = asymmetric relation:
a random variable (Y) depends on a fixed variable (X)

Correlation = symmetric relation: the two variables are at the same level

Correlation:
Both X and Y are RANDOM variables



Regression model



only Y is a RANDOM variable

Model of simple linear regression - 1

Systematic part of the model

Response variable (dependent)

$$y = \beta_0 + \beta x + \varepsilon$$

error term

intercept

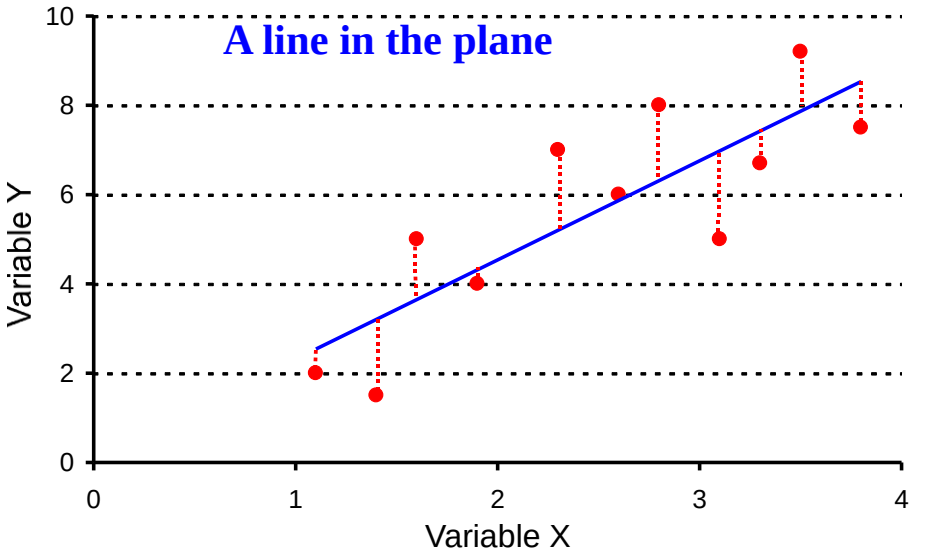
Linear regression coefficient

Unknown parameters of the model, estimated from available data

Explanatory variable (predictive, independent)

Model of simple linear regression - 2

$y = \beta_0 + \beta x + \varepsilon$

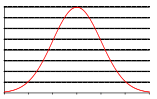


Model of simple linear regression - 3

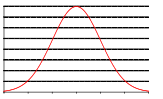
$$y = \beta_0 + \beta x + \varepsilon$$

Linear predictor, deterministic part of the model, without random variability

Response variable



error term, probabilistic part



Error terms, and hence response variable, are NORMALLY distributed

Model of simple linear regression - 4

Weight (Y) depends on height (X_1)

$$E(y) = \beta_0 + \beta_1 x_1$$

E(y) = expected value (mean) of weight in those individuals with that particular height

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

y = weight of a given subject, which depends on height (systematic part of the model), but also on other individual known and unknown characteristics (ε, probabilistic part)

Model of simple linear regression - 5

- **Unknown "real" model**

$$y = \beta_0 + \beta_1 x + \varepsilon$$

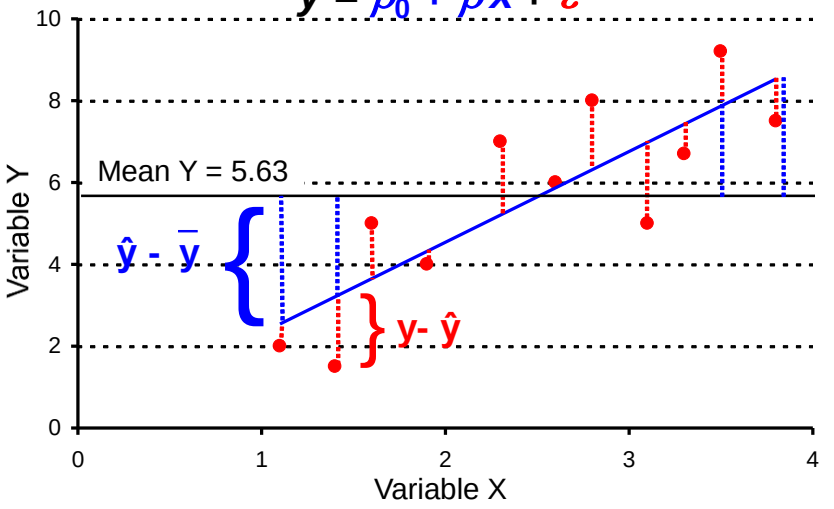
- **Estimated linear regression**

$$\hat{y} = b_0 + b_1 x$$

$$y = b_0 + b_1 x + e$$

DECOMPOSING total sum of squares in simple linear regression - 1

$$y = \beta_0 + \beta x + \varepsilon$$



$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

DECOMPOSING total Sum of Squares (SSq) in simple linear regression - 2

Total variability: $(y - \bar{y})$ = Variability explained by the regression: $(\hat{y} - \bar{y})$ + Residual variability: $(y - \hat{y})$

It can be demonstrated that:

Total SSq, SST: $\Sigma(y - \bar{y})^2$ = Regression-explained SSq, SSR: $\Sigma (\hat{y} - \bar{y})^2$ + Residual (Error, unexplained) SSq, SSE: $\Sigma(y - \hat{y})^2$

Correlation

Correlation coefficient (r) is a dimensionless number, ranging from -1 to +1

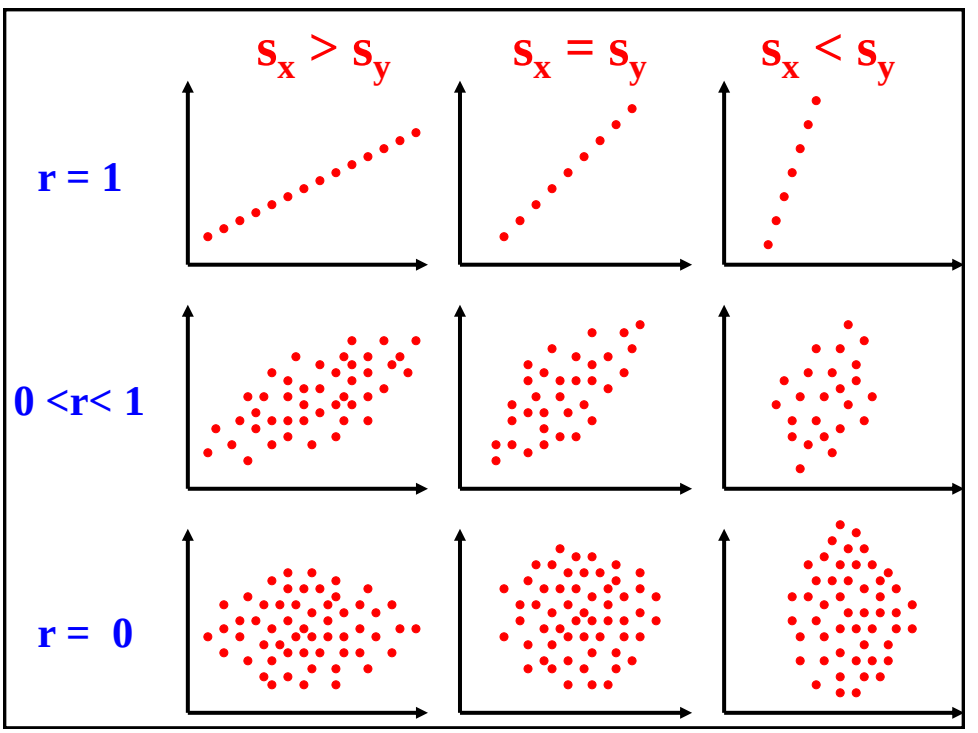
r = -1 points are perfectly aligned along a declining line

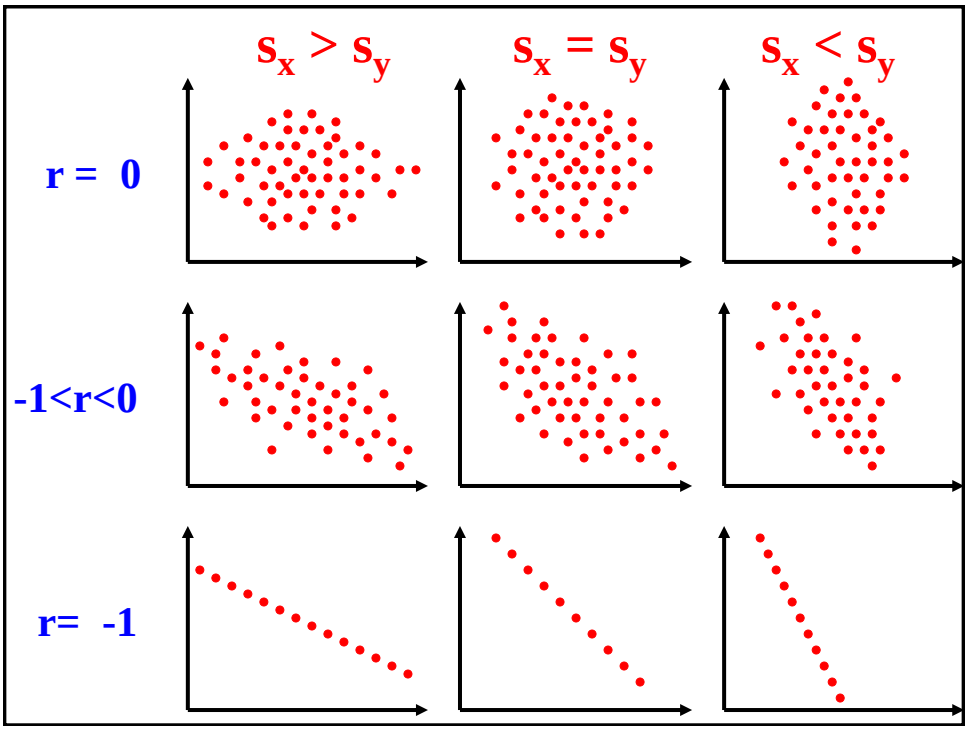
r = 0 points are randomly scattered, without an increasing or decreasing trend

r = +1 points are perfectly aligned along an increasing line

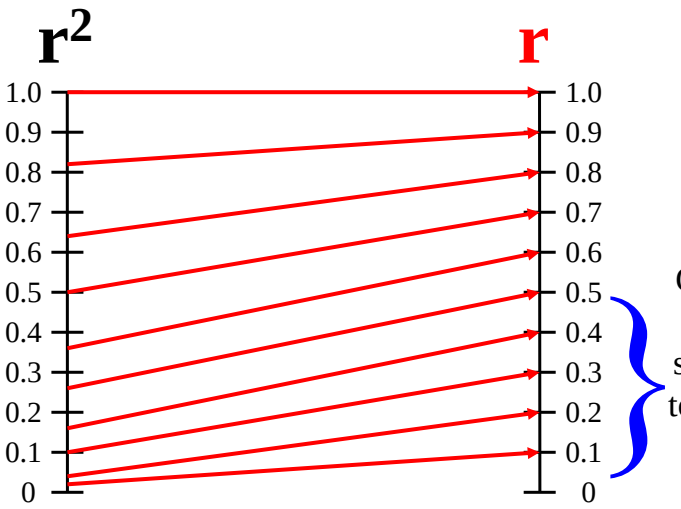
$$r = \frac{\text{Sum of products}_{xy}}{\sqrt{SSq_x * SSq_y}}$$

$$r^2 = \frac{\text{SSq explained by regression, SSR}}{\text{total SSq, SST}} = \frac{\Sigma (\hat{y} - \bar{y})^2}{\Sigma (y - \bar{y})^2}$$





Most correlations between biological variables are rather weak: the coefficient of determination (r^2) ranges between 0 and 0.5.



Going from r^2 to r, by extracting the square root, allows to amplify the scale in the range of weak correlation.

Simple linear regression

One should identify the line, that best fits the scatter points, i.e. that roughly goes through the middle of all the scatter points.

LEAST SQUARES METHOD

The line that minimizes residual (Error) Sum of Squares, SSE, $\Sigma(y - \hat{y})^2$, is selected.

Simple linear regression

b_1 = linear regression coefficient, slope

$$b_1 = \frac{\text{Sum of products}_{xy}}{\text{Sum of squares}_x}$$

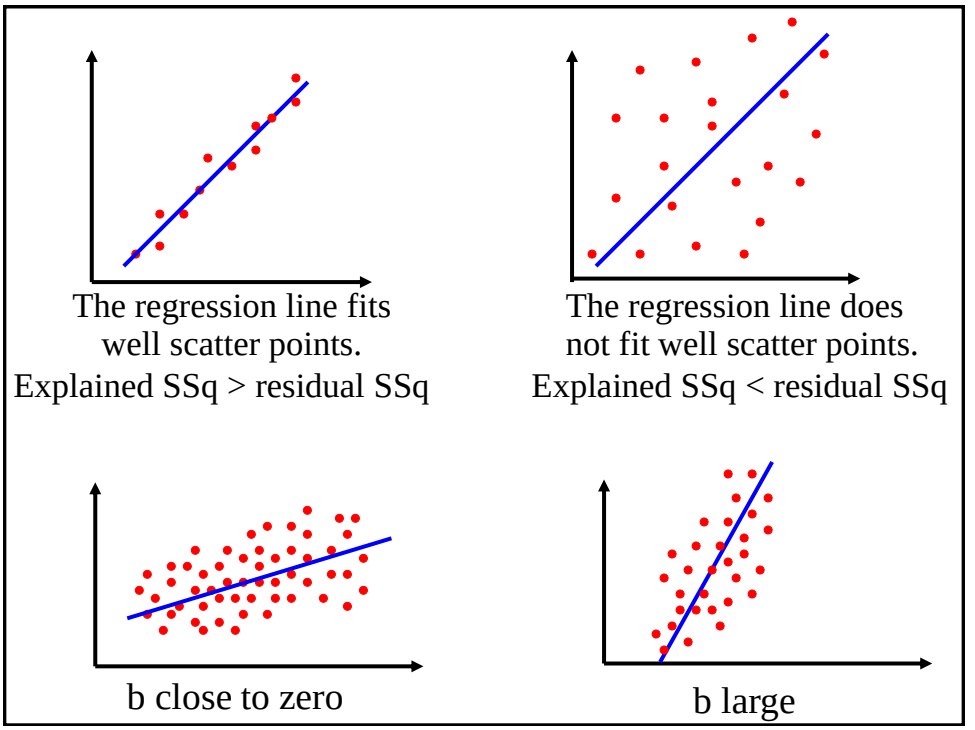
b_1 ranges from $-\infty$ to $+\infty$

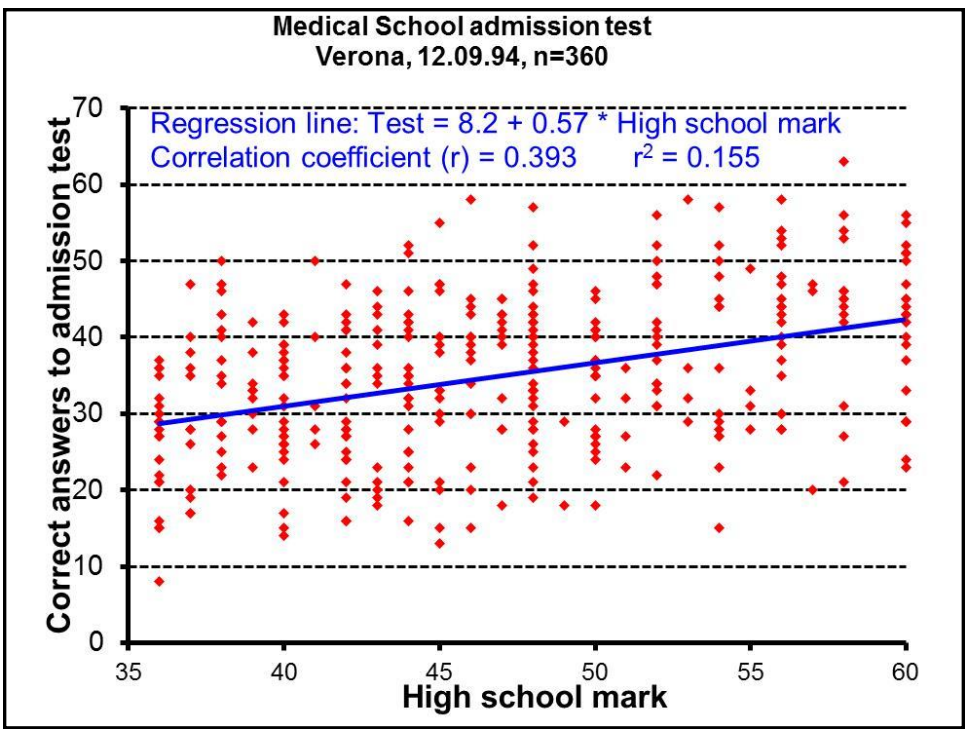
Its measurement unit is the ratio of Y measurement unit to X measurement unit.

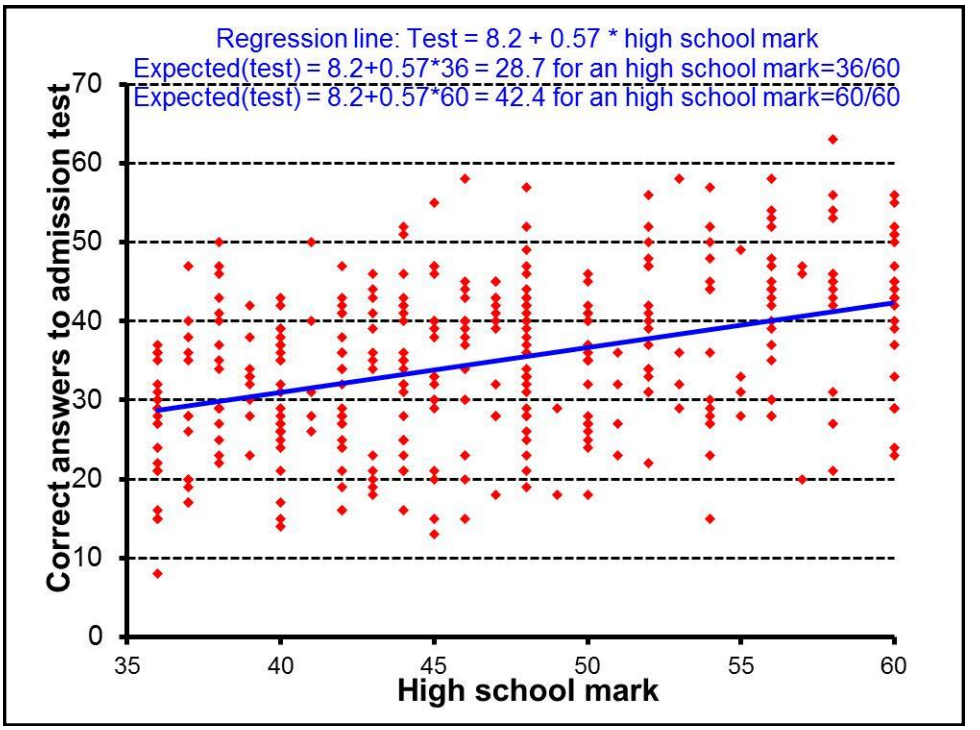
Hence the absolute value of b_1 depends on the measurement units adopted.

b_0 = intercept

$$b_0 = \bar{y} - b_1 \bar{x}$$


The regression line fits well scatter points.
Explained SSq > residual SSq
The regression line does not fit well scatter points.
Explained SSq < residual SSq
b close to zero
b large


Medical School admission test
Verona, 12.09.94, n=360
Regression line: Test = 8.2 + 0.57 * High school mark
Correlation coefficient (r) = 0.393 r² = 0.155
Correct answers to admission test
High school mark
0
10
20
30
40
50
60
70
35
40
45
50
55
60



Correlation

Correlation coefficient (r) = 0.393

$r^2 = 0.155$

$$r^2 = \frac{\text{SSq explained by the regression, SSR}}{\text{Total SSq, SST}} = \frac{\Sigma (\hat{y} - \bar{y})^2}{\Sigma (y - \bar{y})^2}$$

15.5% of total variability (SSq) in admission test score is explained by variability in high school final mark.

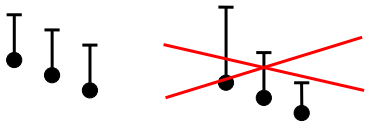
Simple linear regression

$y = \beta_0 + \beta x + \varepsilon$

ASSUMPTIONS

1) Expected value of errors, E(ε), must be ZERO

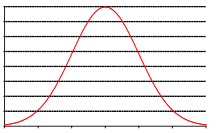
2) HOMOSCEDASTICITY (The variance of errors remains constant)



3) INDEPENDENCE of errors

if test tubes (*provette* in Italian) are not adequately washed between subsequent tests, a test is affected by the previous one

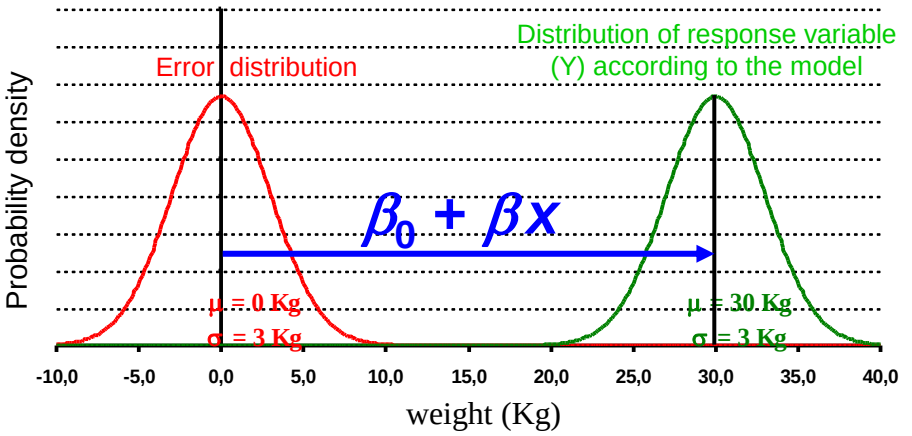
4) Errors are NORMALLY distributed

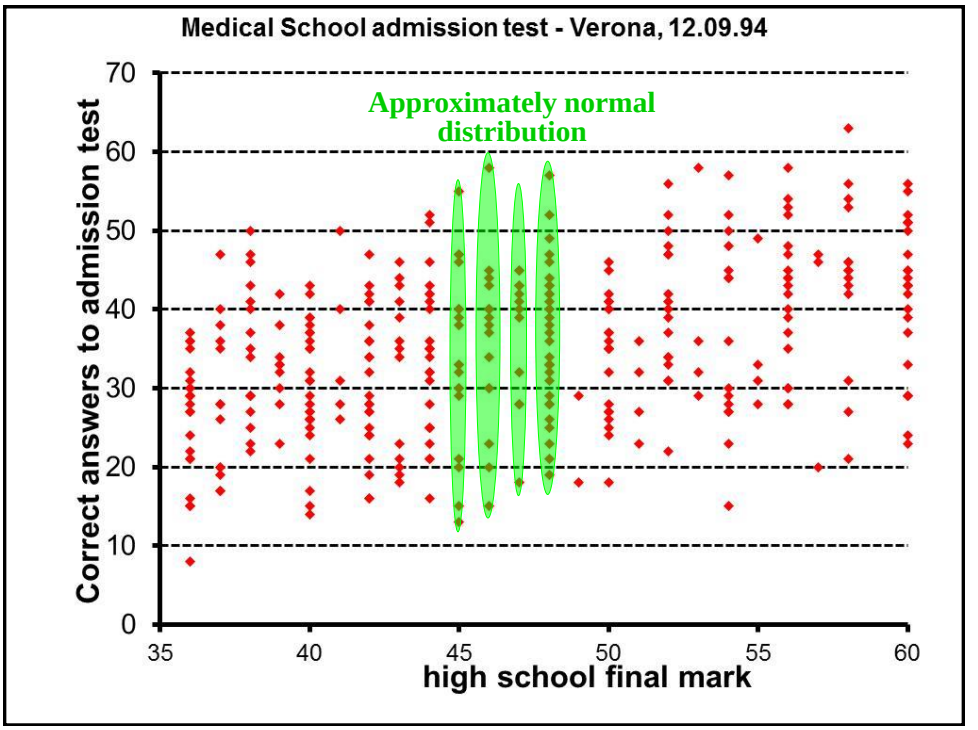


ASSUMPTIONS of simple linear regression

1) The expected value of errors, E(ε), must be ZERO

4) Errors are NORMALLY distributed





EXAMPLE:

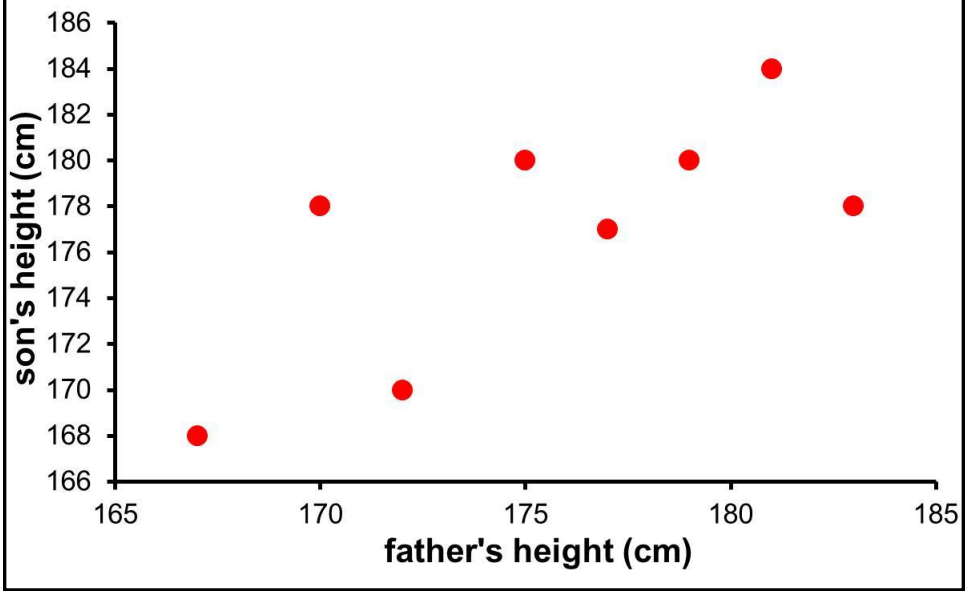
Is sons' height associated with fathers' height ?

In other words, do shorter fathers have shorter sons ?

Do taller fathers have taller sons ?

Father	Son
167 cm	168 cm
175 cm	180 cm
183 cm	178 cm
170 cm	178 cm
181 cm	184 cm
172 cm	170 cm
177 cm	177 cm
179 cm	180 cm

1° STEP: graphic representation (scatterplot) of observed values



2° STEP: a statistical model is assumed, potentially useful to interpret data

The following linear model is assumed: $y = \beta_0 + \beta x + \varepsilon$

(*son's height*) = β_0 + β (*father's height*) + ε

(sons sharing the same father have similar, but not exactly equal height, even if ομοιομετρός, i.e having also the same mother)

3° STEP: uni-variate and bi-variate descriptive statistics

	Σx	Σx^2	N	Σxy	mean
Father	1404	246,618			175.5
			8	248,483	
Son	1415	250,477			176.875

Sum of products= $\Sigma xy - \Sigma x \, \Sigma y/n = 248.483 – 1404*1415/8 = 150.5$

	SSq	Variance	Standard dev.
Father	216	30.86	5.55
Son	198.875	28.41	5.33

4° STEP: Estimation of model parameters by the method of least squares

$$b_1 = \frac{\text{Sum of products}_{xy}}{\text{Sum of squares}_x} = 150.5/216 = 0.697 \text{ cm/cm}$$

$$b_0 = \bar{y} - b_1 \bar{x} = 176.875 – 0.697 *175.5 = 54.59 \text{ cm}$$

Regression line:

Son's height (cm)= 54.6 cm + 0.697 cm/cm * father's height (cm)

When father's height increases by 1 cm, son's height increases on average by 7 mm.

5° STEP: Computing correlation coefficient

$$r = \frac{\text{Sum of products}_{xy}}{\sqrt{SSq_x * SSq_y}} = \frac{150.5}{\sqrt{216 * 198.9}} = 0.726$$

$$r^2 = 0.7261^2 = 0.527$$

52.7% of total variability in sons' height is explained by variability in fathers' height.

6° STEP: Inference on model parameters: Is the theoretical model "supported" by observed data?

Student's *t*-test, based on b_1 (estimate of β_1)

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$
Two-tailed test

Significance level = 5%
Degrees of freedom = n-2 = 8-2 = 6
Critical threshold = $t_{6,\ 0.025} = 2.447$

$$t = \frac{b-0}{SE_b} = \frac{b}{\sqrt{var_{res}/SSq_x}} = \frac{0.697}{\sqrt{15.67 / 216}} = 2.588$$

$SSq_{res} = SSq_y - \sum products_{xy}^2 / SSq_x = 198.9 - 150.5^2/216 = 94.01$
$variance_{res} = SSq_{res} / (n-2) = 94.01 / 6 = 15.67$

6° STEP: Inference on model parameters: Is the theoretical model "supported" by observed data?

Student's *t*-test, based on r (estimate of ρ)

$\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$ Two-tailed test

Significance level = 5%

Degrees of freedom = n-2 = 8-2 = 6

Critical threshold = $t_{6,\ 0.025} = 2.447$

$$t = \frac{r-0}{SE_r} = \frac{r}{\sqrt{(1-r^2)\,/(n-2)}} = \frac{0.726}{\sqrt{(1-0.527)\,/6}} = 2.587$$

| observed t | > critical *t* value

2.588 > 2.447

H_0 is rejected

The observed relation between sons' height and fathers' height is not due to chance, but rather represents a real phenomenon (P=0.041).

7° STEP: Prediction of Y values

Which is the expected value of Y, for x = 185 cm ?

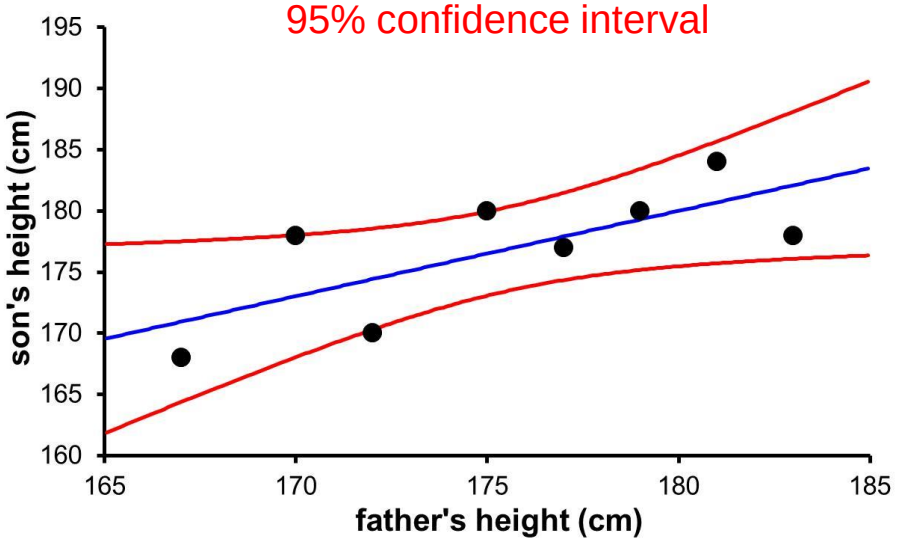
Regression line:

$\hat{y}$ = 54.6 cm + 0.697 cm/cm *185 cm = 183.49 cm

$$SE_{\hat{y}} = \sqrt{var_{res}\,[1/n + (x-\bar{x})^2 / SSq_x\,]} =$$

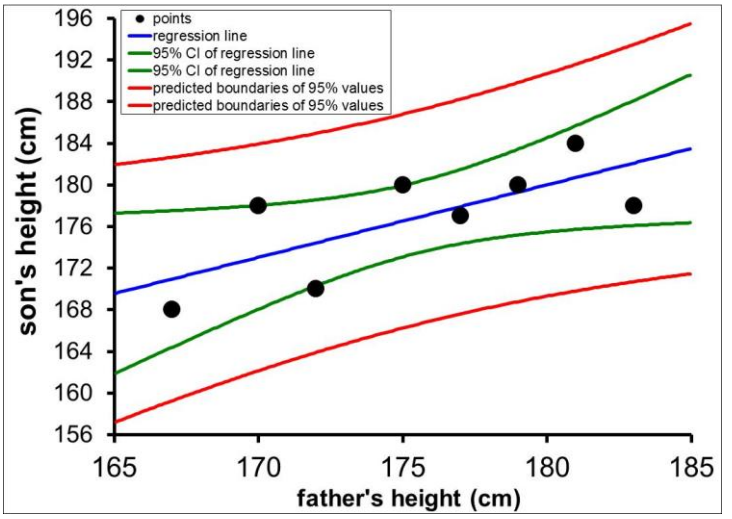
$$= \sqrt{15.67\,[1/8 + (185-175.5)^2 / 216\,]} = 2.916$$

$$IC_{95\%} = \hat{y} \pm t_{\nu,\alpha/2} * SE_{\hat{y}} = 183.49 \pm 2.447 * 2.916 = \begin{bmatrix} 190.63 \\ 176.36 \end{bmatrix}$$



Uncertainty in the estimates of the regression line increases as X moves away from X mean (175.5 cm)

Regression line,
95% confidence interval of the regression line,
prediction interval containing 95% of single observations



Uncertainty is greater in predicting single observations than in estimating expected values (regression line)