

## Two or more variables (one group)

### Correlation and regression

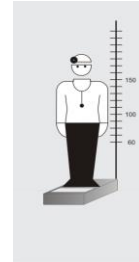
The relationship between two variables.

Method to estimate the relationship between variables.

## Correlation of two variables

Example:  
Is there any relationship between the height and weight?

Experiment:



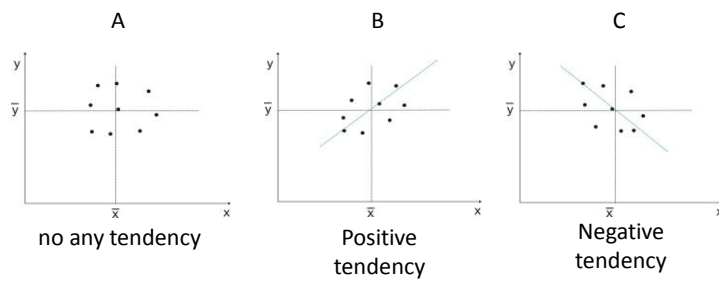
Data pairs:

No.	Height (cm)	Weight (kg)
1	150	61
2	170	70
3	166	75
4	174	70
5	180	72
6	155	50
7	172	65
8	161	59
9	177	81

## Graphic representation

E.g.: height is the x and weight is the y.

Possible situations:



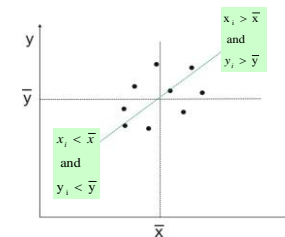
## Covariance

$$Q_{xy} = \sum_i [(x_i - \bar{x}) \cdot (y_i - \bar{y})]$$

$$\text{cov}(x, y) = \frac{Q_{xy}}{n - 1}$$

(n: no. of elements)

Positive tendency:

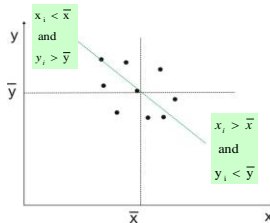


**Frequently:**

if  $x_i < \bar{x}$  then  $y_i < \bar{y}$   
or  $x_i > \bar{x}$  then  $y_i > \bar{y}$

**Consequence:**  $Q_{xy} > 0$ .

Negative tendency:



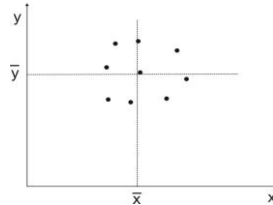
Frequently:

if  $x_i < \bar{x}$  then  $y_i > \bar{y}$

or  $x_i > \bar{x}$  then  $y_i < \bar{y}$

**Consequence:**  $Q_{xy} < 0$ .

No tendency:



The y values are independent from the x-values!

**Consequence:**  $Q_{xy} = 0$ .  
(if  $n \rightarrow \infty$ )

## Pearson's correlation coefficient

$$r = \frac{\text{cov}(x, y)}{s_x \cdot s_y} = \frac{Q_{xy}}{\sqrt{Q_x \cdot Q_y}}$$

Possible range for r:

$$-1 \leq r \leq 1$$

Covariance divided by the squareroot of the product of the standard deviation of the two variables (= standardized covariance).  
The measure of the relationship.

**In the population:**

$r = 0$  no correlation,

$r \neq 0$  correlation (strength is proportional to the actual value of  $r$ .)

## Coefficient of determination

$$r^2$$

The coefficient of the determination tells us how strong is the relationship.  
Expresses how much percent of the variability of the y values may be accounted by the variability of the independent variable or variables.

## Correlation t-test

Calculated  $r$  is the estimation of the  $r$  in the population. This fluctuates around the theoretical value.  
(e.g.  $r_{\text{calc}} = 0.1$  ?)

$$H_0: r = 0! \longrightarrow t = r \sqrt{\frac{n-2}{1-r^2}} \longrightarrow \text{d.f.: } n-2$$

**Decision:** based on t-value. Look previous cases!

**Condition:** at least one of the variable has normal distribution.

## Non-normal distribution or ordinal data

Example: blood pressure measurements.  
Relationship between the two methods.



The distribution is skewed, non-normal.  
Pearson's r is false.

## Spearman's rank-correlation

Example: diastolic pressure.  
(only a few cases!)

case	Cuff	rank	finger	rank
1	80	4.5	58	1
2	65	2	79	5
3	70	3	66	2
4	80	4.5	93	6
5	60	1	75	4
6	82	6	71	3
...				

$$r_s = \frac{\sum (R_x - \bar{R}_x)(R_y - \bar{R}_y)}{\sqrt{\sum (R_x - \bar{R}_x)^2 \cdot \sum (R_y - \bar{R}_y)^2}}$$

$R_x$ : rank of the x variable  
 $R_y$ : rank of the y variable.

The test and the decision are the same as in the case of Pearson's r.

## Linear regression

If the variables have normal distribution, the relationship is linear, and we can describe with straightline.

The regression model:  
(to predict the y values)

$$y_i = ax_i + b + h_i$$

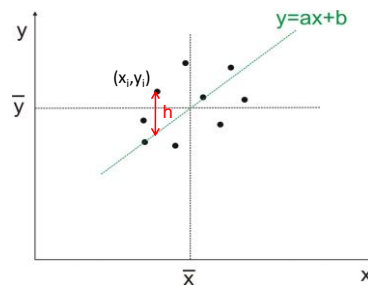
In regression analysis :

$$y_i = b_0 + bx_i + h_i$$

$y$ : dependent variable

$x$ : independent (explanatory) variable

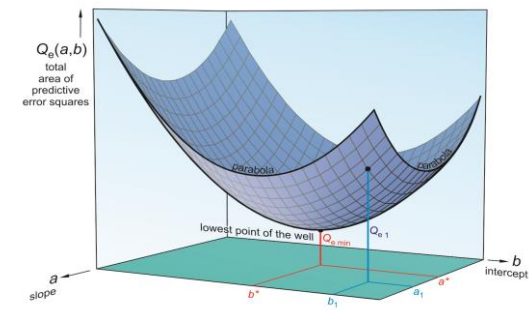
$h_i$ : error term =  $y_i - (ax_i + b)$ .  
(the difference between the actual value and the predicted value.)



## Least-squares method

$$Q_e = \sum_i h_i^2 = \sum_i (y_i - (ax_i + b))^2$$

The  $x_i$  and  $y_i$  are measured values.  
Unknown values are the  $a$  and  $b$ !



## Which is the best straightline?

$Q_e$  has minimum!



$$a^* = \frac{Q_{xy}}{Q_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b^* = \bar{y} - a^* \bar{x}$$

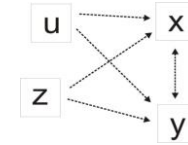
Prediction of insulin sensitivity by BMI.

$r^2$ : coefficient of determination.  
How much part of the variance of  $y$  is explained by  $x$ .

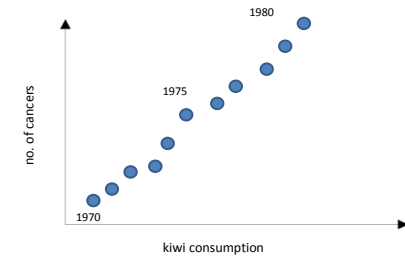
indep.	regr. coeff.	st. error	t	p	decision
BMI	-0.077	0.018	-4.25	0.0011	significant
$r^2$	0.6				

## Interpretation of the correlation

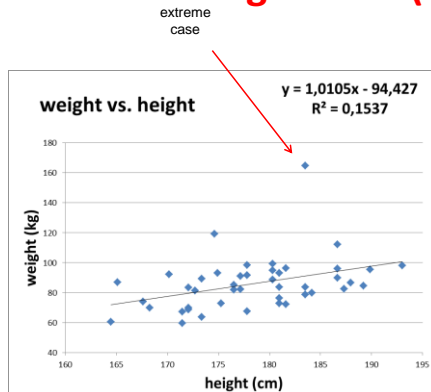
Not necessary being direct causality. (But may be!)  
In the background there may be quantities, effects that influence both measured variables.



Example:  
There is positive correlation,  
but we can't suppose causality.



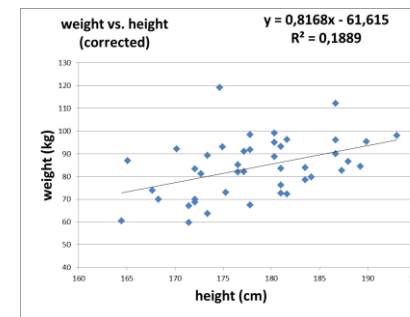
## Linear regression (example 1)



correlation  
t-test

n	44
r	0,3920
t	2,761
p	0,85%

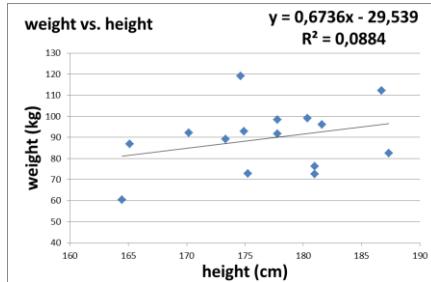
## Linear regression (without extreme case)



correlation  
t-test

n	43
r	0,4347
t	3,090
p	0,36%

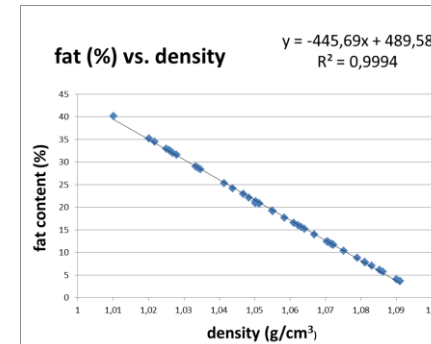
## Linear regression (effect of n)



correlation  
t-test

n	15
r	0,2973
t	1,123
p	28,20%

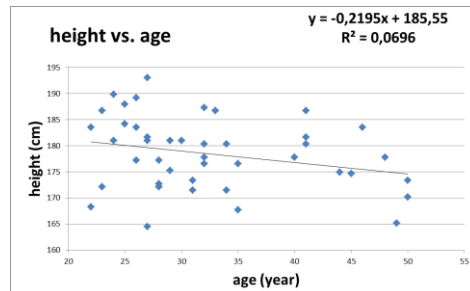
## Linear regression (example 2)



correlation  
t-test

n	44
r	-0,9997
t	-271,242
p	9,33E-70

## Linear regression (example 3)



correlation  
t-test

n	44
r	-0,2638
t	-1,772
p	8,36%

## Multilinear regression

Two or more independent variables.

The regression model:  
(to predict the y values)

$$y_i = b_0 + \sum_j b_j x_{j,i} + h_i$$

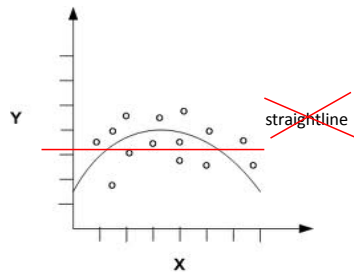
Predict the insulin sensitivity (y)!

Conclusion: only the BMI influences the sensitivity. About the 64% of the variation of the sensitivity explained by the BMI.

indep.	regr. coeff.	st. error	t	p	decision
Age	-0.0045	0.0041	-1.09	0.3	not sign.
BMI	-0.068	0.02	-.344	0.0055	significant
r²	0.639				

## Non-linear regression

On the base of the model (e.g.):



Polinomial:  $y = b_0 + \sum_i b_i x^i + h$

Exponential:  $y = hab^x$

Power:  $y = hax^b$