School of Prophets

Trainer: PhD. MsC. Bondor Cosmina-Ioana

Association and prediction

A ALWAYS

S SEEK

K KNOWLEDGE

Objectives

- Correlation
- Linear regression
- Exercices

Statistical inference in the case of two variables

1. one qualitative variable
   one quantitative variable
   - compare arithmetic means of a quantitative variable on different samples (given by the categories of the qualitative variable)
2. two qualitative variable
   - compare frequencies of a qualitative variable on different samples (given by the categories of another qualitative variable)
3. two quantitative variable
   - one sample where we calculate correlation and linear regression between two quantitative variables

Inferential statistics: two quantitative variables (or ordinal)

Scatter plot

- to show the relationship between two numerical variables
- plotting dots on an X and Y axes

- objective: to show
  - trends
  - correlations (positive, negative, or none)
  - patterns
  - clusters
  - outliers

Scatter plot

- Dot in the Scatter chart
  - represents an individual data
    - its position depend on the value of variable X and Y

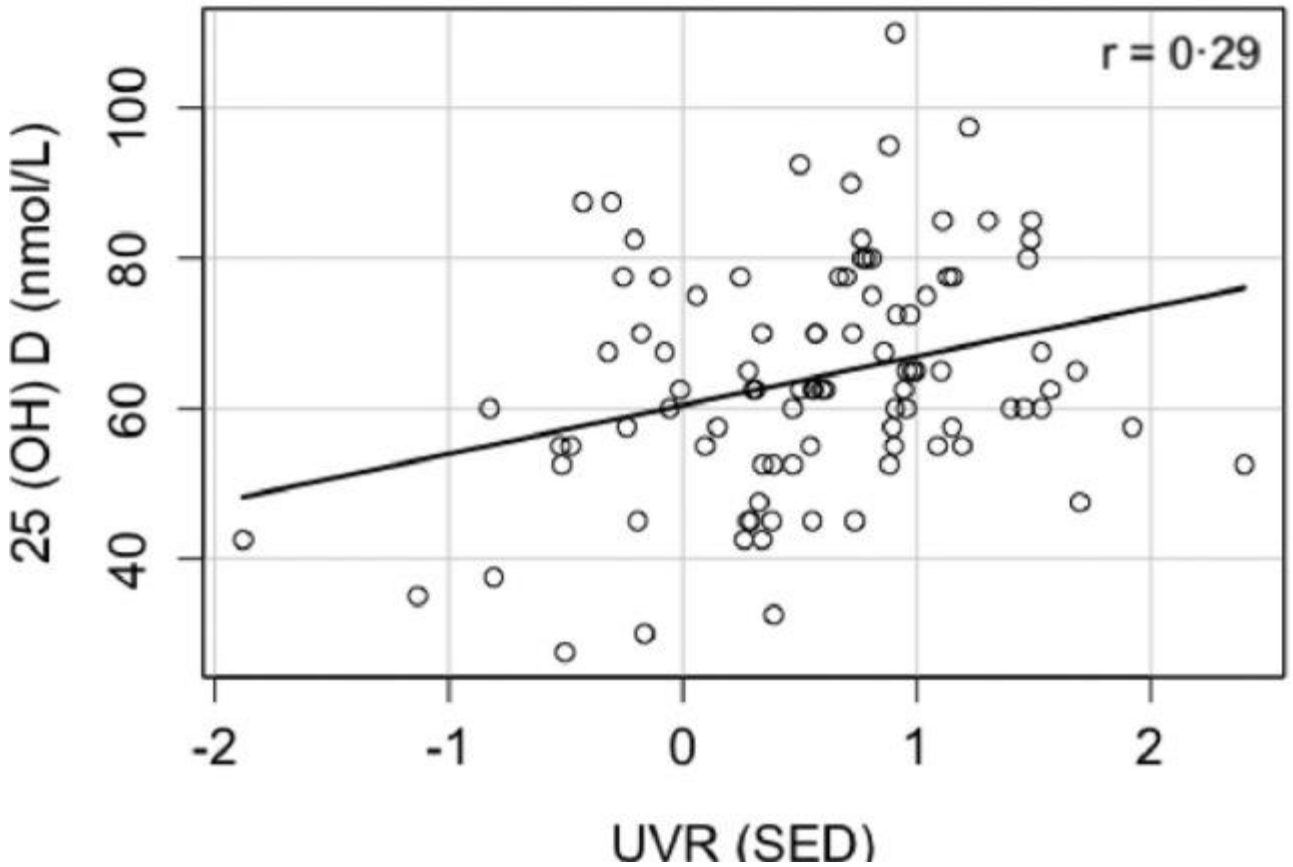
!

- 2 continuous or discrete variables
  - X and Y
- Y is correlated with X?

- ex.
  - X = Vitamin D2 (25(OH)D) concentration nmol/L
  - Y = exposure to ultraviolet radiation (UVR)
  - vitamin D2 variation is associated with UVR variation?

	A	B
1	UVR (SED)	vitamin D2 (25(OH)D) concentration
2	27	50
3	25	55
4	21	42
5	27	60
6	24	45
7	35	75
8	30	80
9	28	65
10	29	65
11	30	70
12	32	85
13	24	58
14	25	59
15	27	61
16	25	66
17	26	62
18	29	63
19	28	60
20	26	54
21	28	67
22	28	70
23	23	49
24	29	65
25	25	60
26	27	60
27	26	59

101 women aged 35 years or older in 2019



- Aim: to assess the association between vitamin D2 (25(OH)D) concentration and level of exposure to ultraviolet radiation (UVR)

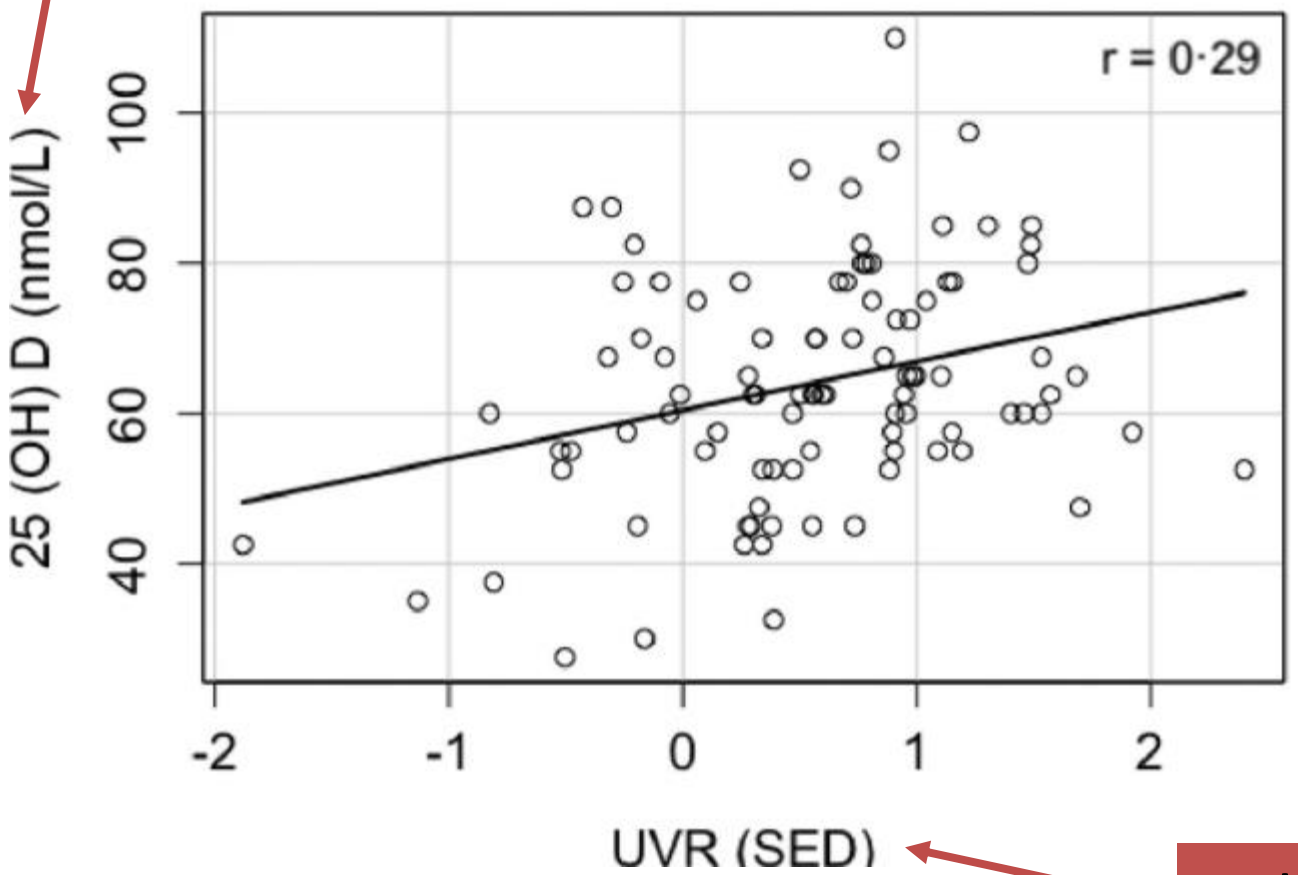
Santana, Keila Valente de Souza de et al. Association between vitamin D status and lifestyle factors in Brazilian women: Implications of Sun Exposure Levels, Diet, and Health. eClinicalMedicine, Volume 47, 101400

- 2 continuous or discrete variables
  - different unity of measurements
- Independent variable OX axes
  - BMI
- Dependent variable OY axes
  - Vitamin D2

	A	B
1	UVR (SED)	vitamin D2 (25(OH)D) concentration
2	27	50
3	25	55
4	21	42
5	27	60
6	24	45
7	35	75
8	30	80
9	28	65
10	29	65
11	30	70
12	32	85
13	24	58
14	25	59
15	27	61
16	25	66
17	26	62
18	29	63
19	28	60
20	26	54
21	28	67
22	28	70
23	23	49
24	29	65
25	25	60
26	27	60
27	26	59

Select the colums with the data

Dependent variable on OY axis

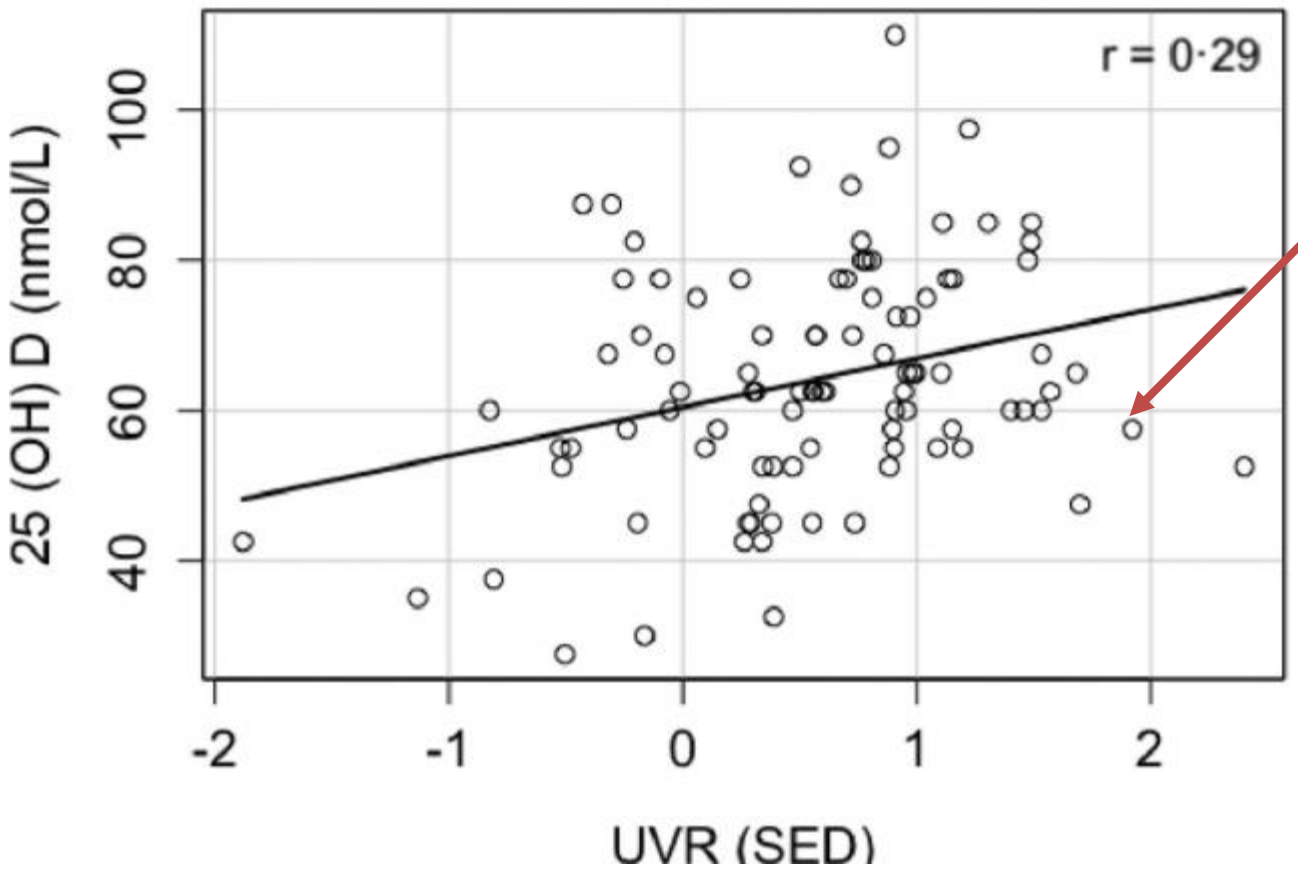


- 101 women aged 35 years or older in 2019
- Aim: to assess the association between vitamin D2 (25(OH)D) concentration and level of exposure to ultraviolet radiation (UVR)

Independent variable on OX axis

Santana, Keila Valente de Souza de et al. Association between vitamin D status and lifestyle factors in Brazilian women: Implications of Sun Exposure Levels, Diet, and Health. eClinicalMedicine, Volume 47, 101400

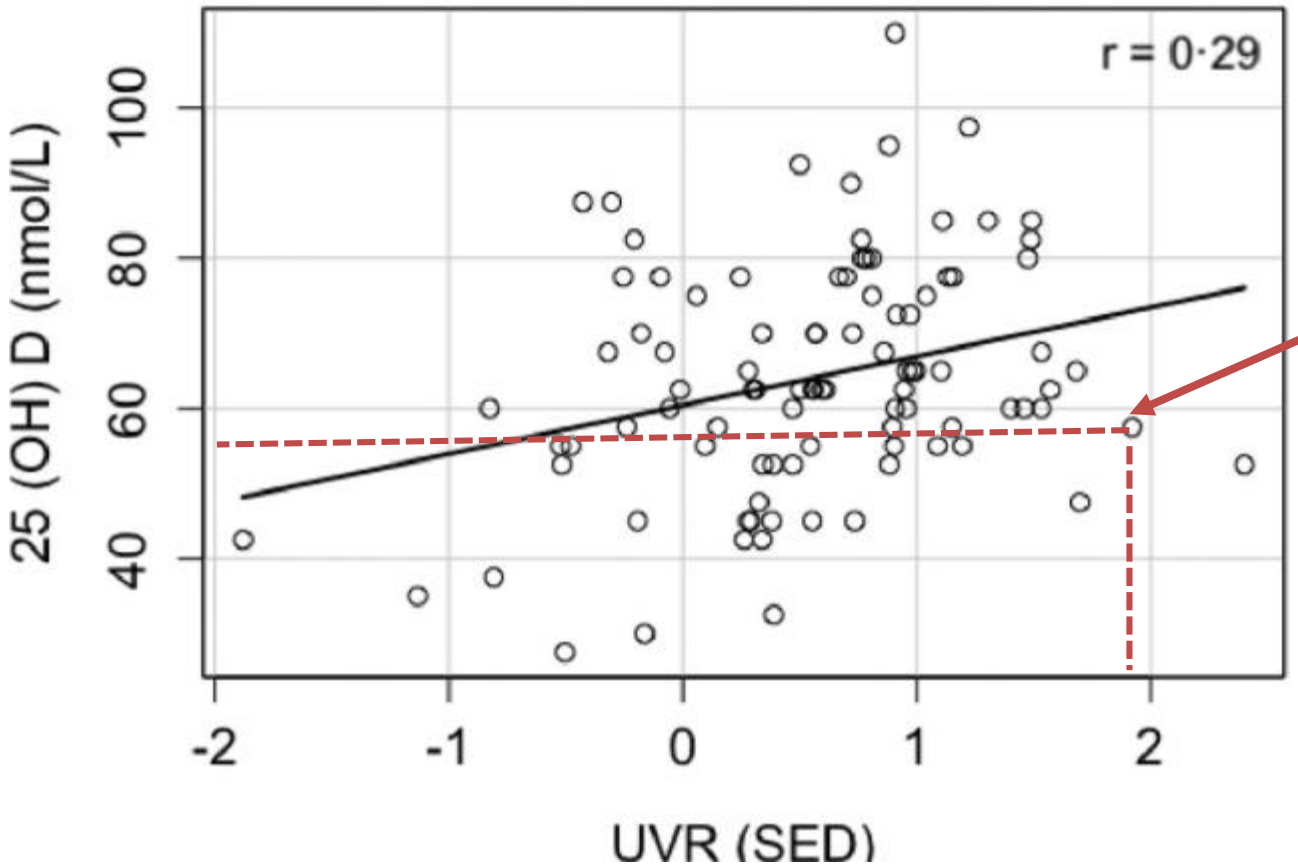
XY Scatter chart



- 101 women aged 35 years or older in 2019
- =101 dots on the chart, one dot for each patient, with his/ her's exposure to UVR on the OX axis and with his/her's quantity of vitamin D in the blood

Santana, Keila Valente de Souza de et al. Association between vitamin D status and lifestyle factors in Brazilian women: Implications of Sun Exposure Levels, Diet, and Health. eClinicalMedicine, Volume 47, 101400

XY Scatter chart



a patient with UVR= 1.9 SED and 25(OH)D=58 nmol/L

Santana, Keila Valente de Souza de et al. Association between vitamin D status and lifestyle factors in Brazilian women: Implications of Sun Exposure Levels, Diet, and Health. eClinicalMedicine, Volume 47, 101400

Correlation

- The relationship between two numerical characteristics

- How is the relationship?
- Can we predict an event?
- What error of prediction can we afford?

The Pearson correlation coefficient

If X and Y are two quantitative variables. The Pearson correlation coefficient :

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}$$

Where n – number of patients (sample size), $\overline{X},\overline{Y}$-arithmetic means of X and Y variables

The Pearson correlation coefficient r

- indicates the association between X and Y
- Always between -1 and 1

$$r \in [-1,1]$$

As $|\ r\ |$ approaches 1 the association is stronger

As $|\ r\ |$ approaches 0 the association is weaker

! We are talking about linear relationship

Colton rules

– r in [-0.25 to +0.25] → No relation

– r in (0.25 to +0.50] or in (-0.50 to -0.25] → Weak relation

– r in (0.50 to +0.75] or in (-0.75 to -0.50] → Moderate relation

– r in (0.75 to +1] or in [-1 to -0,75) → Strong relation

[Colton T. Statistics in Medicine. Little Brown and Company, New York, NY 1974]

Positive correlation – direct proportionally association

As $|\ r\ |$ approaches 1
the association is stronger

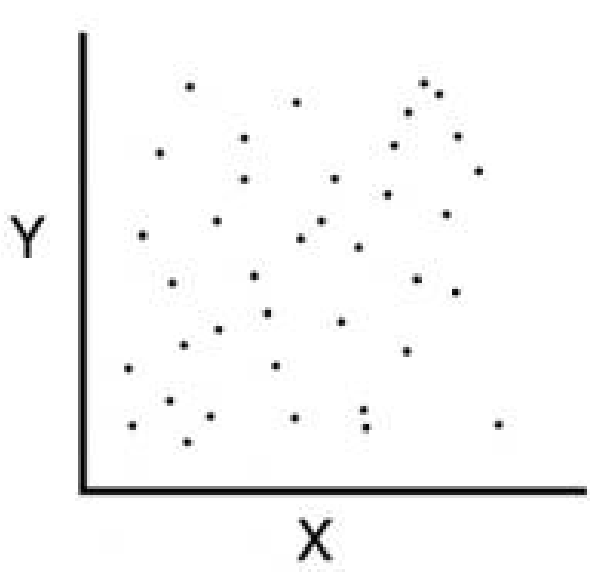
As $|\ r\ |$ approaches 0
the association is weaker

Strong correlation

r ≈ 1

Moderate correlation

r ≈ 0.5



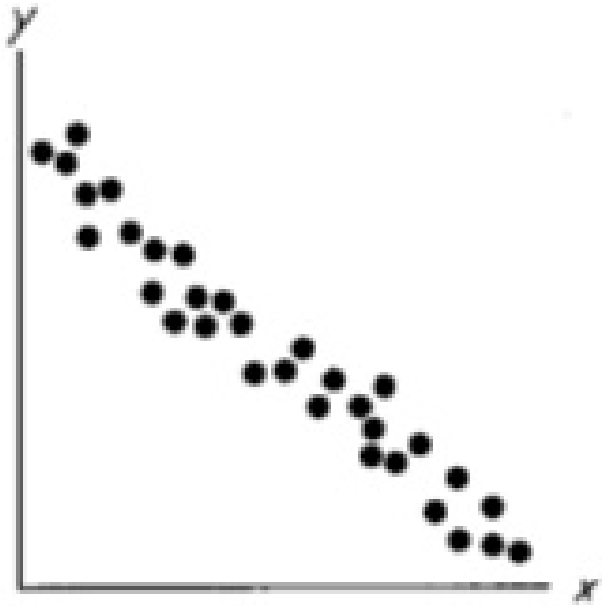
Null correlation

r ≈ 0

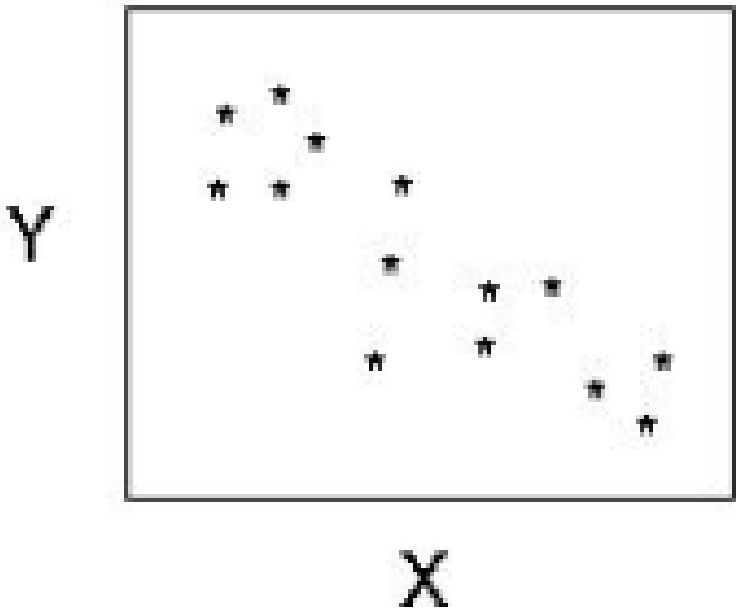
Negative correlation – indirect proportionally association

As $|\ r\ |$ approaches 1
the association is stronger

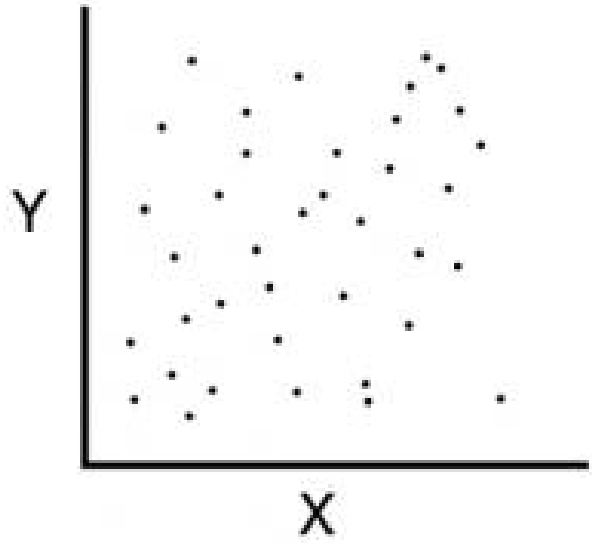
As $|\ r\ |$ approaches 0
the association is weaker



Strong correlation
r ≈ -1



Moderate correlation
r ≈ -0.5



Null correlation
r ≈ 0

The Pearson correlation coefficient r

- If r > 0 than the association between X and Y is **positive** (direct)

for low value of X – corresponds low value of Y,

for high level of X – corresponds high level of Y

X increase → Y increase

X decrease → Y decrease

Strong correlation

r ≈ 1

Moderate correlation

r ≈ 0.5

!

The Pearson correlation coefficient r

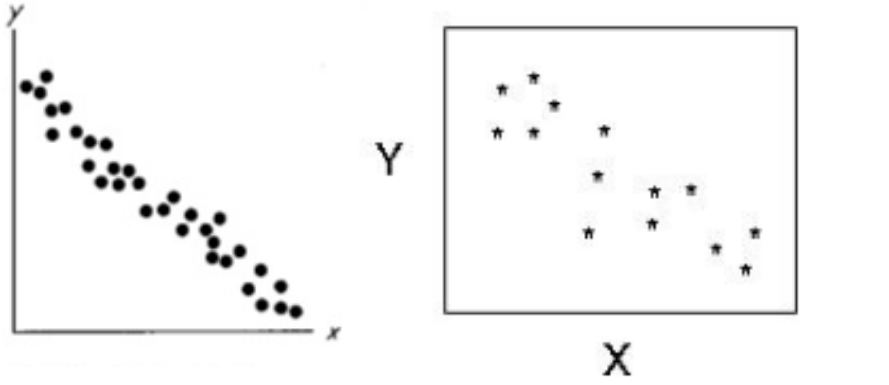
- If r < 0 than the association between X and Y is **negative** (inverse)

  for low value of X – corresponds high value of Y,

  for high level of X – corresponds low level of Y

X increase → Y decrease

X decrease → Y increase



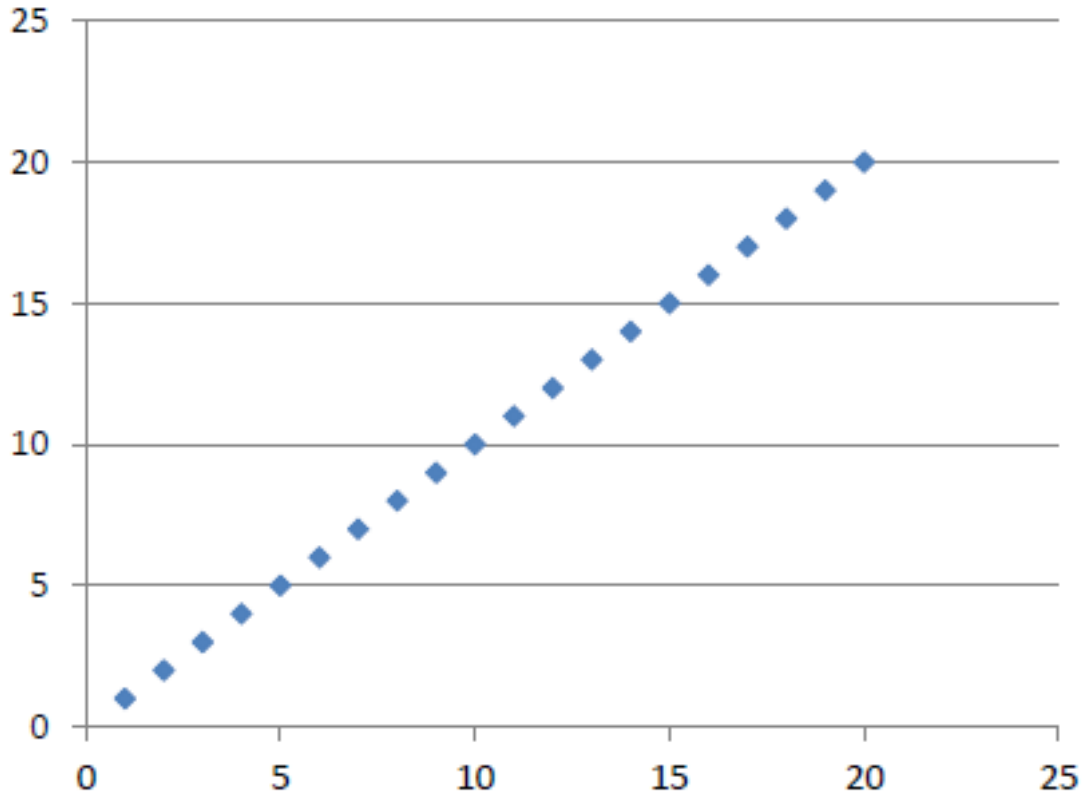
Strong correlation
r ≈ -1

Moderate correlation
r ≈ -0.5

!

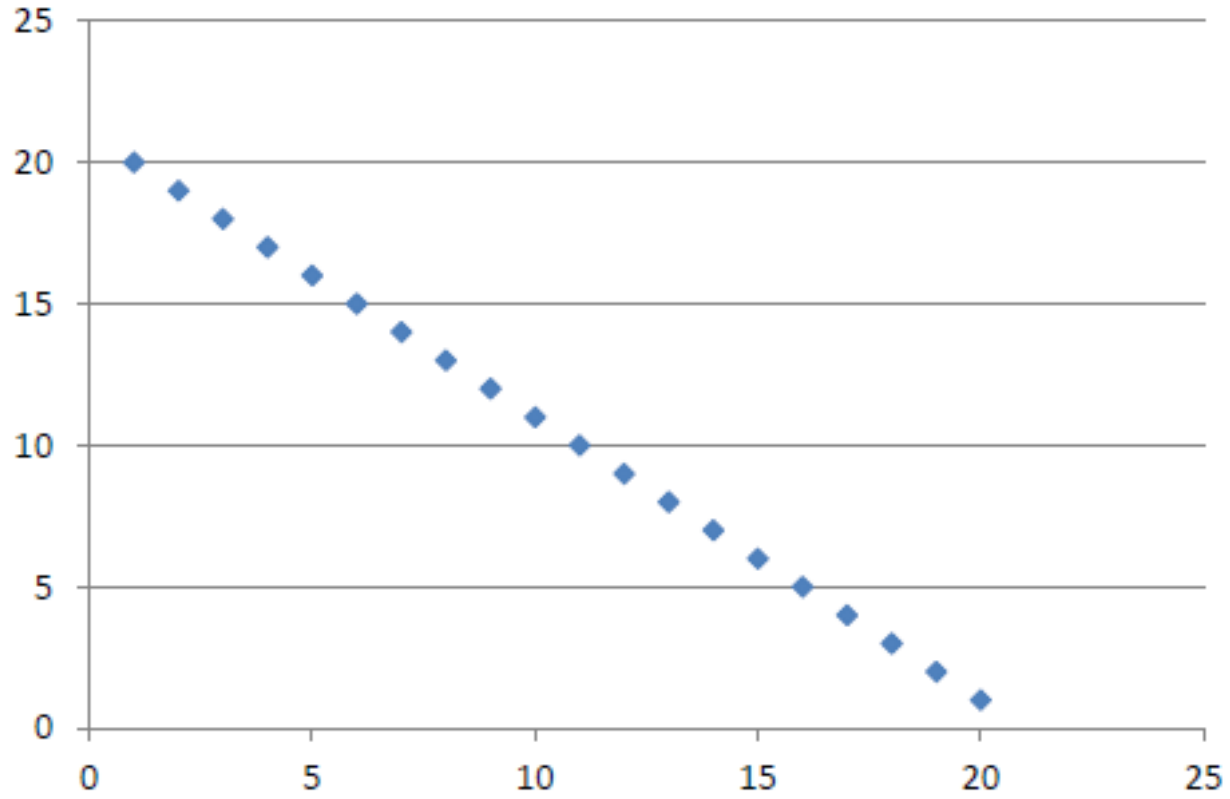
Perfect correlation, when all the dots are on a line → linear correlation

- r=1



Positive (direct)

r=-1



Negative (invers)

Statistically significant correlation

- T test for correlation
- Objective: to estimate the Pearson correlation coefficient in the population: ρ
- The estimation is based on a random sample of the population where r was calculated.

Significant statistically

- **H0 (null hyphotesis):**
  - the coefficient of correlation is not statistically significantly different than 0
- **H1 (alternative hyphotesis):**
  - the coefficient of correlation is statistically significantly different than 0

- The test parameter: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
- Result: p

- If p<0.05 then we reject the null hypothesis (H0) and accept the alternative hypothesis (H1): the correlation is statistically significant
- If p≥0.05 then we fail to reject the null hypothesis (H0): the correlation is not statistically significant

Examples

- Correlation between Sugar intake and Proportion of tooth decay surface r=0.76, p=0.001
- p<0.05 correlation between Sugar intake and Proportion of tooth decay surface is statistically significant
  - if we repeat the study the probability to not see a difference between 0 and r is very low

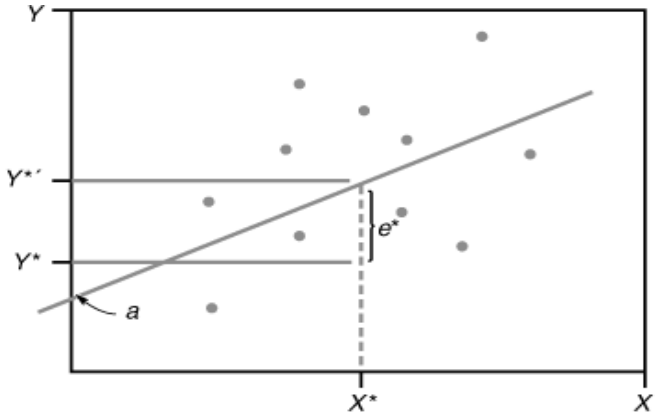
- Correlation between Sugar intake and Age r=0.08, p=0.85
- p>0.05 correlation between Sugar intake and Age is not statistically significant
  - if we repeat the study the probability to not see a difference between 0 and r is high

Regression

- Prediction = regression
  - modeling the relationship between Y and X by an equation

- univariate – based on one independent variable X
- multivariate – based on more than one independent variables X_i, i=2,n

Linear regression

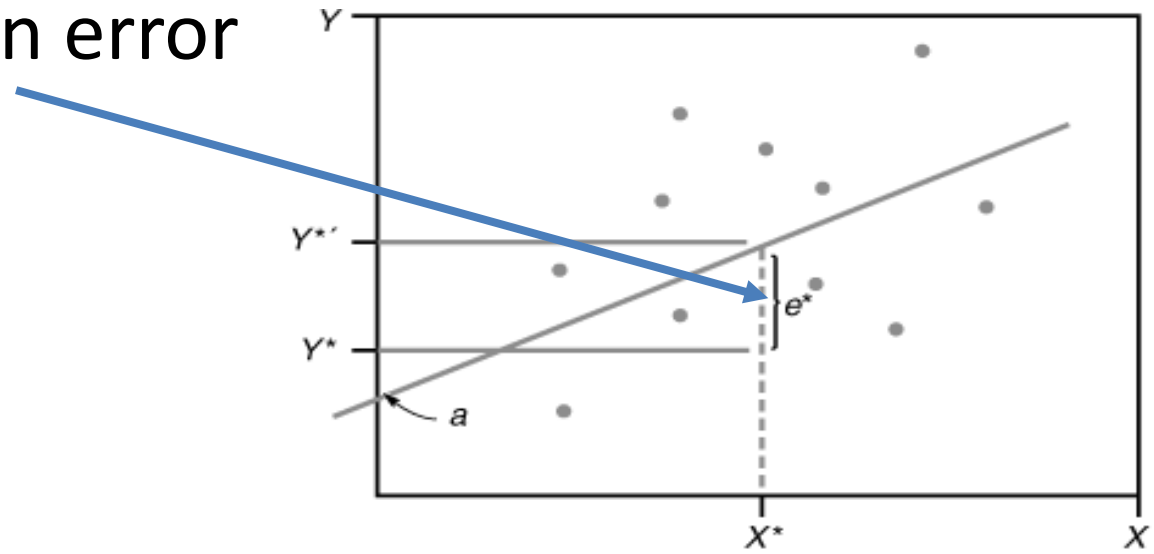


$$Y=aX+b$$

- a line obtained with the least squares method
  - minimize data point deviations
  - the best fit line for the data cloud
- The point where the line crosses the Y axis and is denoted by b
- The slope of the line with a

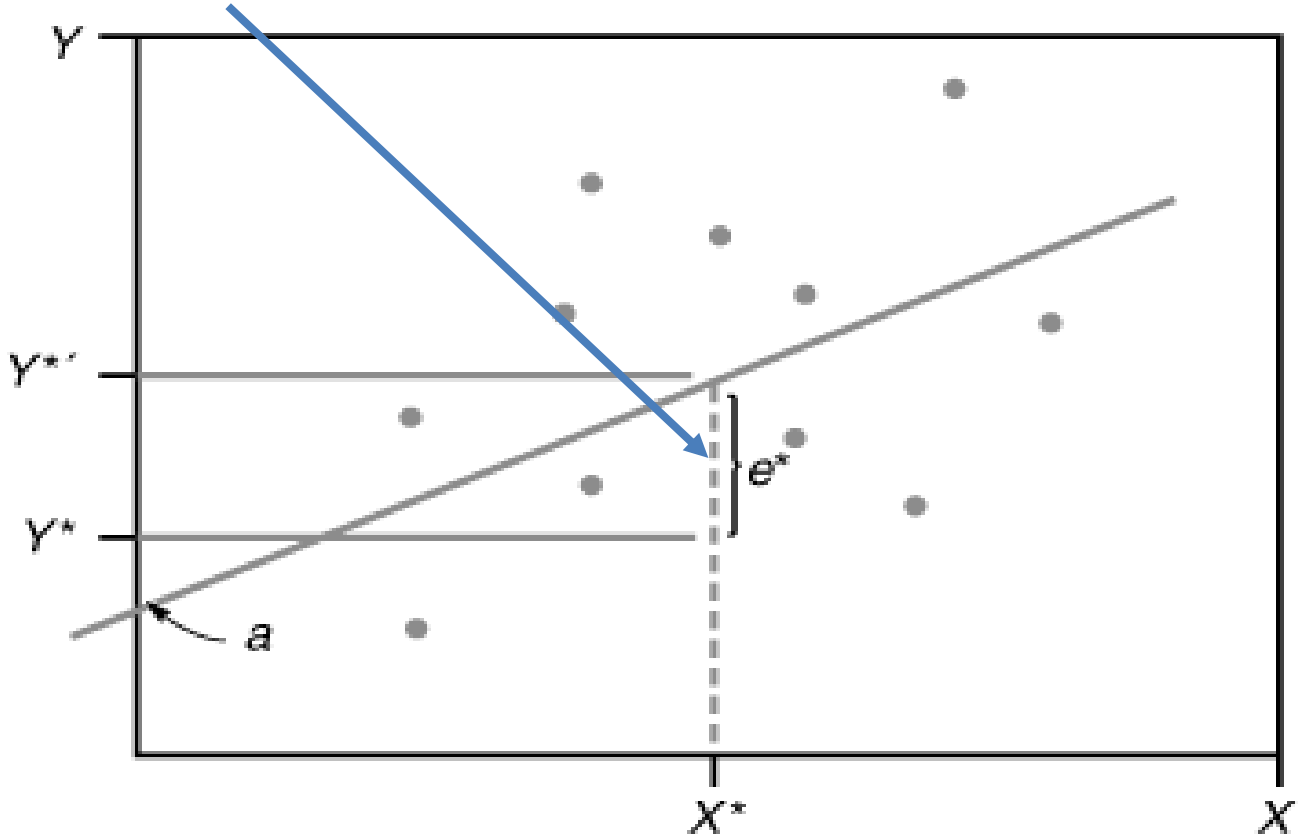
Linear regression (Y=aX+b)

- Y increases by **a** every time X increases by 1
- If **a** is negative, when X increases Y decrease
- If **a** is positive, when X increases, Y increases
- The linear model does not fit perfectly (not all dots are on the regression line) -> prediction error



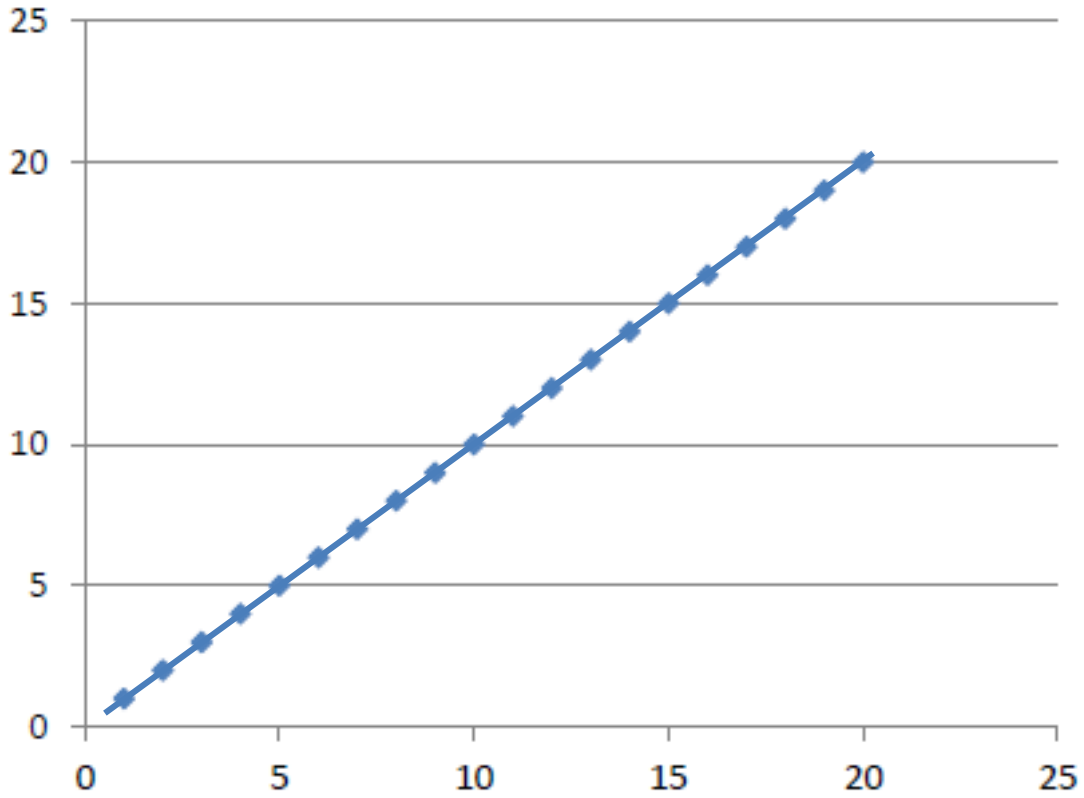
Linear regression (Y=aX+b)

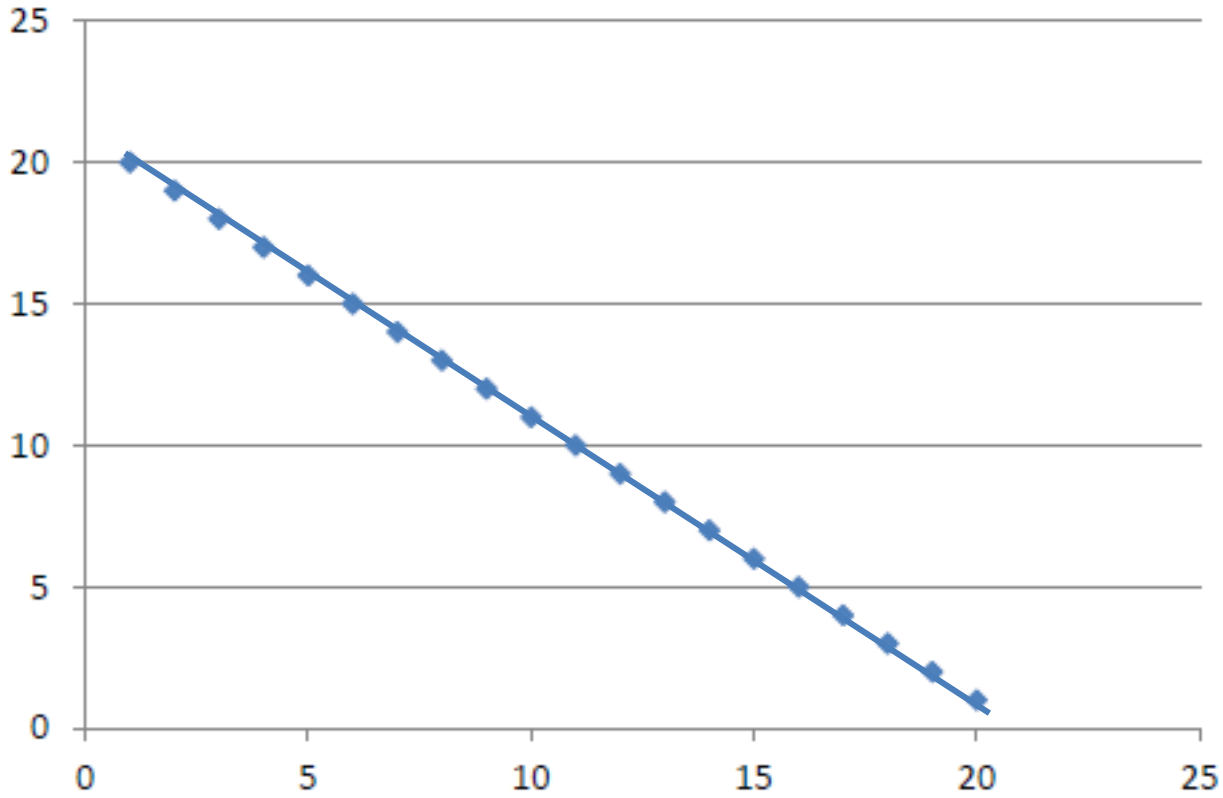
- The linear model does not fit perfectly (not all dots are on the regression line) -> prediction error



Perfect prediction – all the dots are on the line

-





- Formula for the coefficients of regression line:

$$b = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

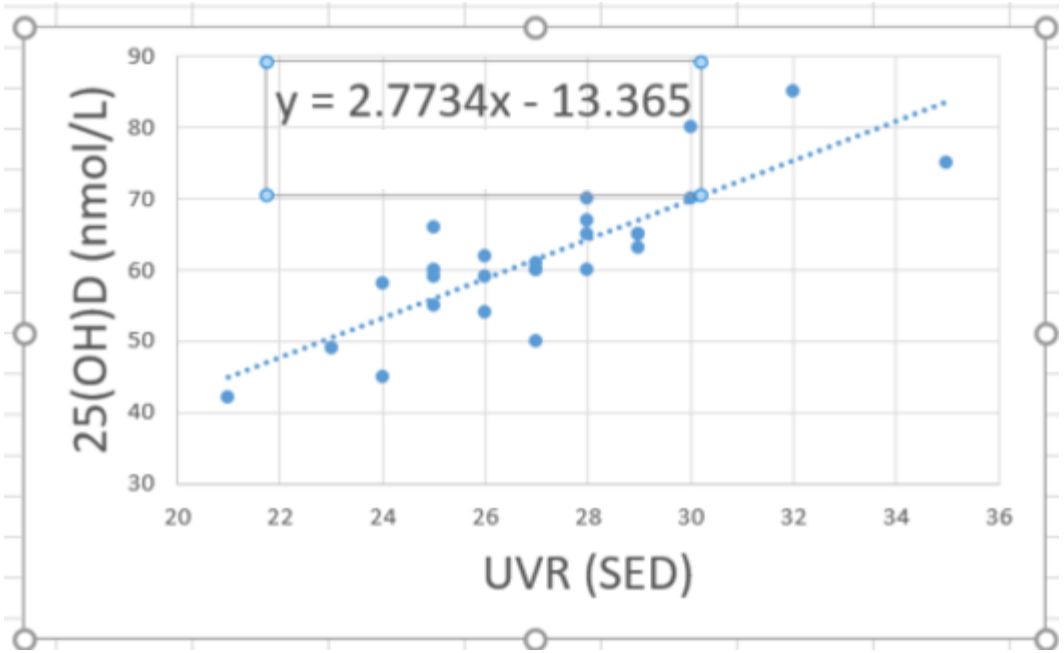
- two variables X=UVR, Y=25(OH)D
  - with two different measurement units: SED, nmol/L
- The dependent variable on the OY-axis –ex. 25(OH)D
- the independent variable on the OX-axis -ex. UVR

$$Y = ax + b$$

$$25(OH)D = 2.77UVR - 13.37$$

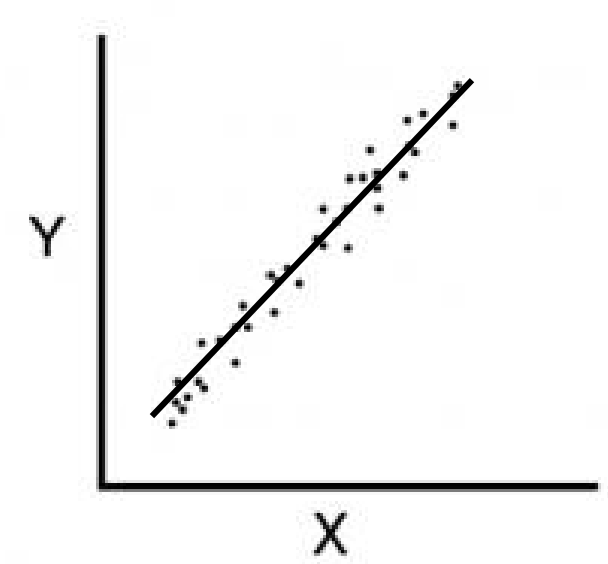
$$a=2.77$$

$$b=-13.37$$



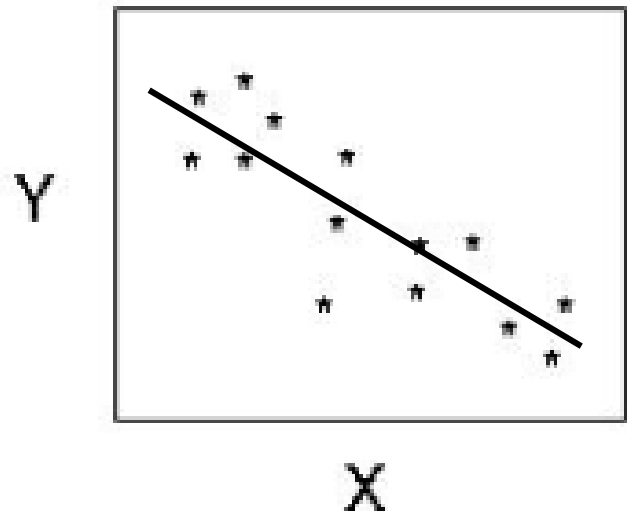
Correlation chart – XY Scatter

- two variables – with two different measurement units
- The dependent variable on the OY-axis and the independent variable on the OX-axis



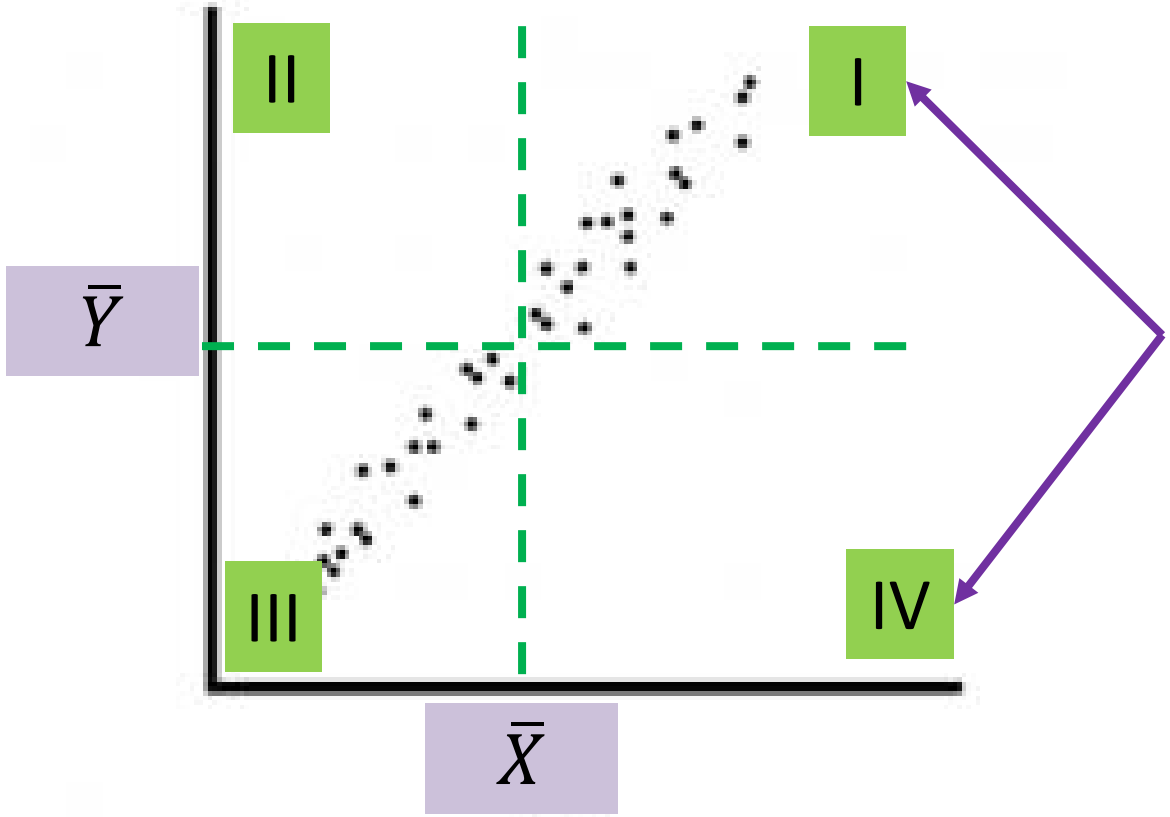
Ascending trend

$r>0, a>0$



Descending trend

$r<0, a<0$

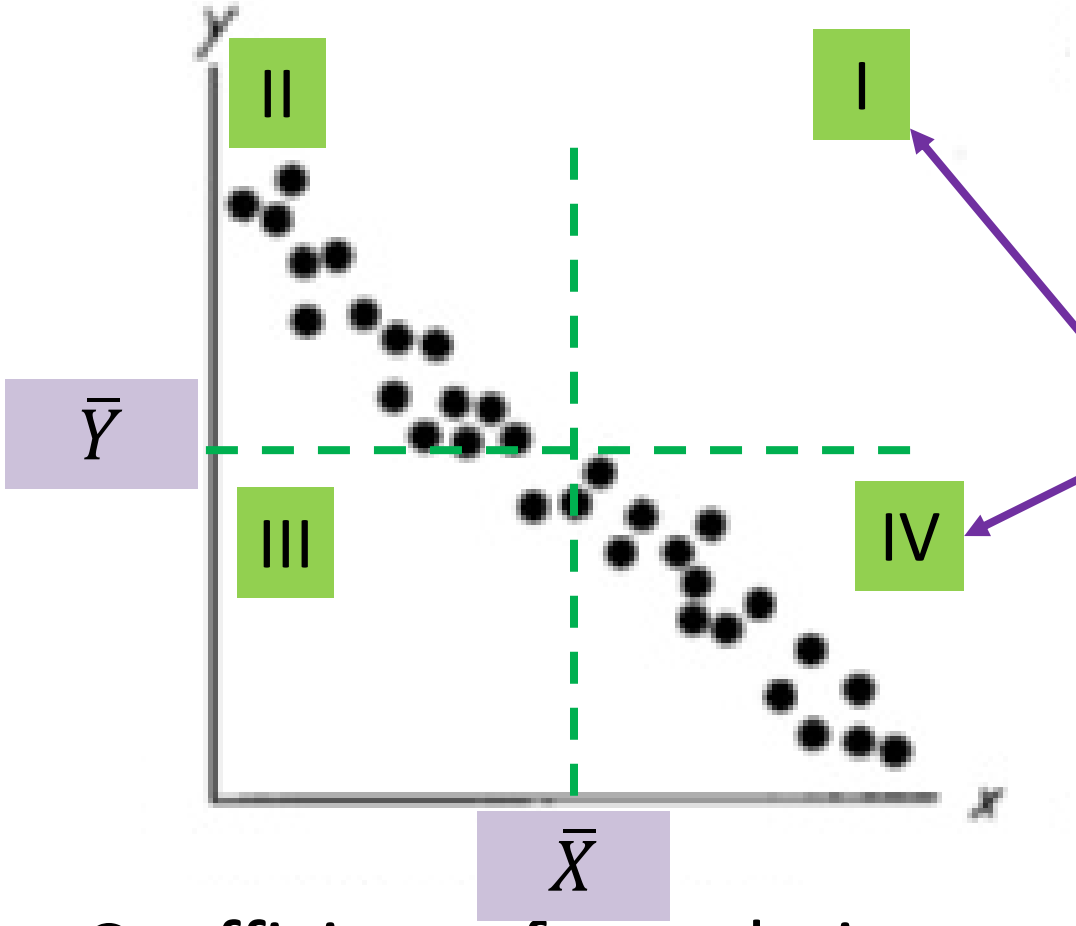


Quadrant given by the arithmetic means of X and Y

Coefficient of correlation r > 0

Ascending trend

More dots are in quadrants I and III then in quadrant II and IV



Quadrant given by the arithmetic means of X and Y

Coefficient of correlation $r < 0$

Descending trend

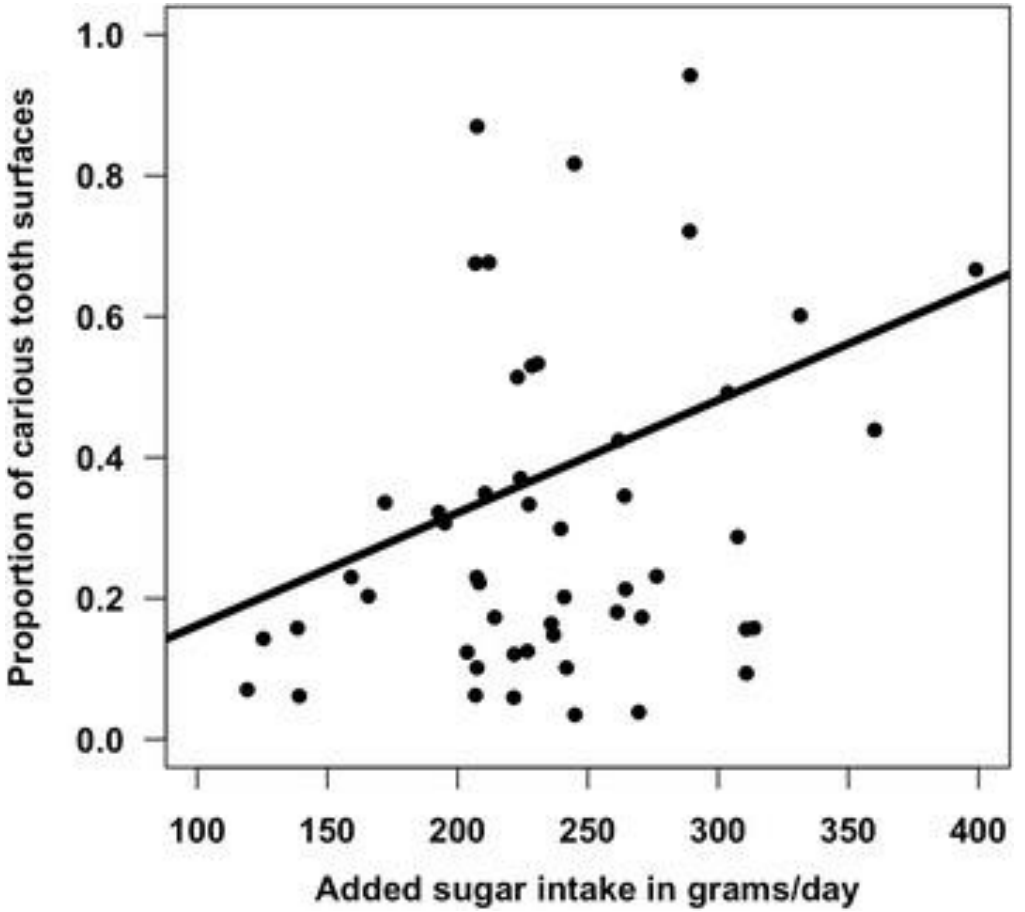
More dots are in quadrants II and IV then in quadrant I and III

Example – How we interpret r?

- Correlation between
  - Added sugar intake
  - Proportion of tooth decay surface

  r = 0.76

[Chi DL, Hopkins S, O'Brien D, Mancl L, Orr E, Lenaker D. Association between added sugar intake and dental caries in Yup'ik children using a novel hair biomarker. BMC Oral Health. 2015 Oct 9;15(1):121.]



Interpretation:

Positive correlation, ascending trend, direct proportional correlated

- for high level of sugar intake – corresponds high level of tooth decay surface
- Sugars intake increase → tooth decay surface increase

Dots are preponderant in quadrant I and III

Colton rules r in (0.75 to +1] →

between sugar intake and proportion of tooth decay surface there is a strong correlation

Example – How we interpret r?

- Correlatic
  - Added s

urface r = 0.76

[Chi DL, Hop
between adc
using a nove
9;15(1):121.

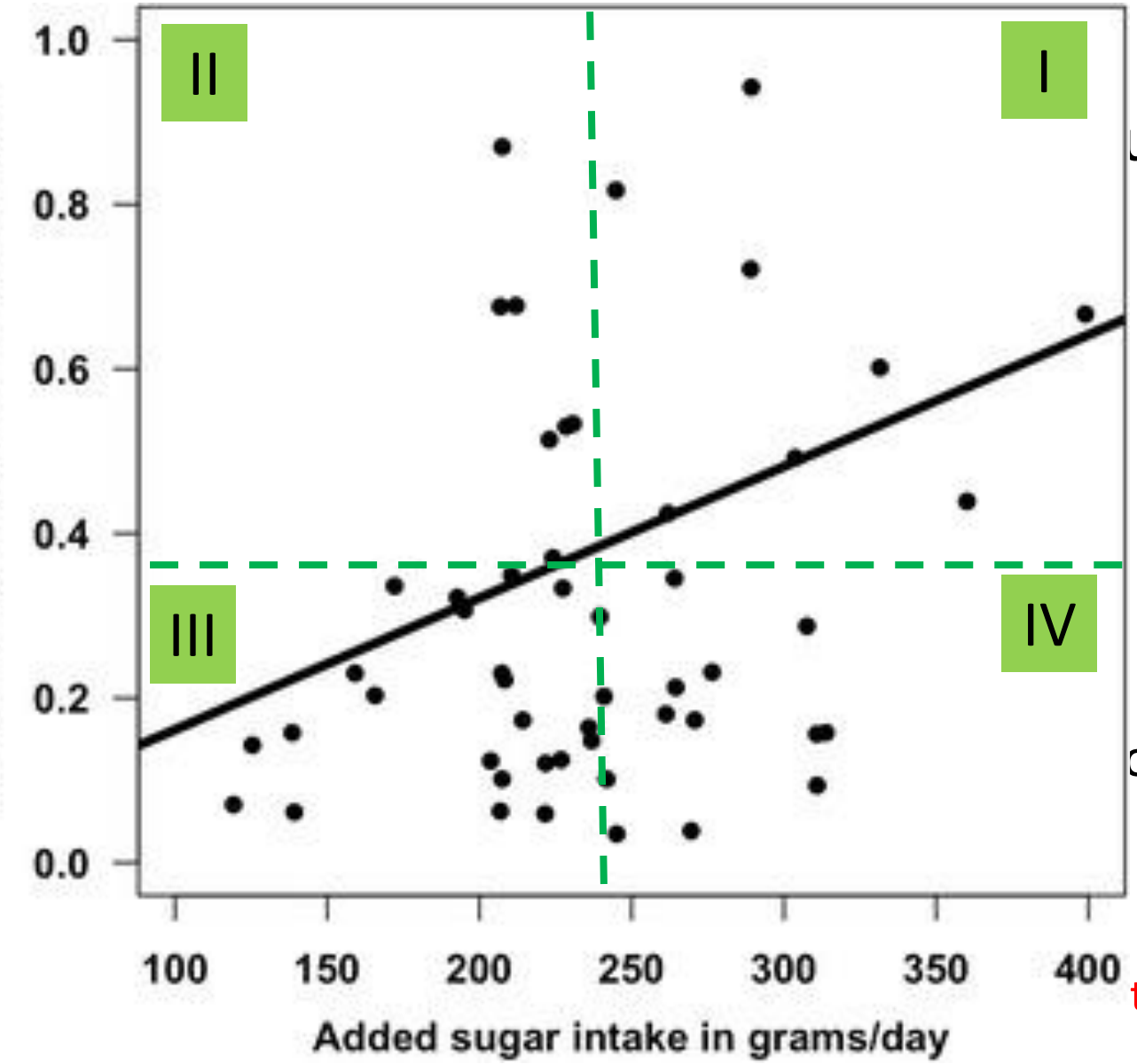
Interpretati

Positive cc

- for hig
- Sugga

Colton rules

between s

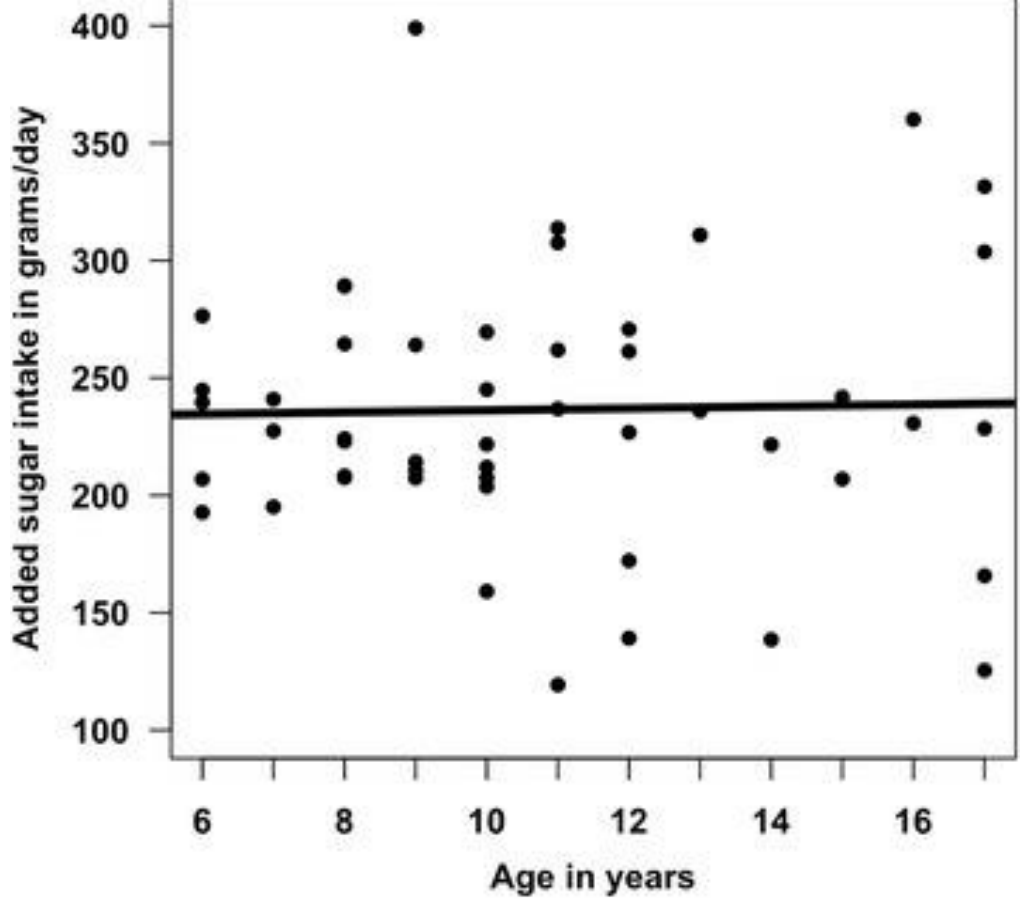


ooth decay surface

trong relationship

Example – How we interpret r?

- Correlation between
  - Sugar intake and Age r=0.08



Interpretation:

near 0

- no correlation between sugar intake and age
- Dots are equally dispersed in all four quadrant

Colton rules r in [-0.25 to +0.25] →

between sugar intake and age there is a no correlation

[Chi DL, Hopkins S, O'Brien D, Mancl L, Orr E, Lenaker D. Association between added sugar intake and dental caries in Yup'ik children using a novel hair biomarker. BMC Oral Health. 2015 Oct 9;15(1):121.]

Example – How we interpret r?

- Correlatio
  - Sugar int

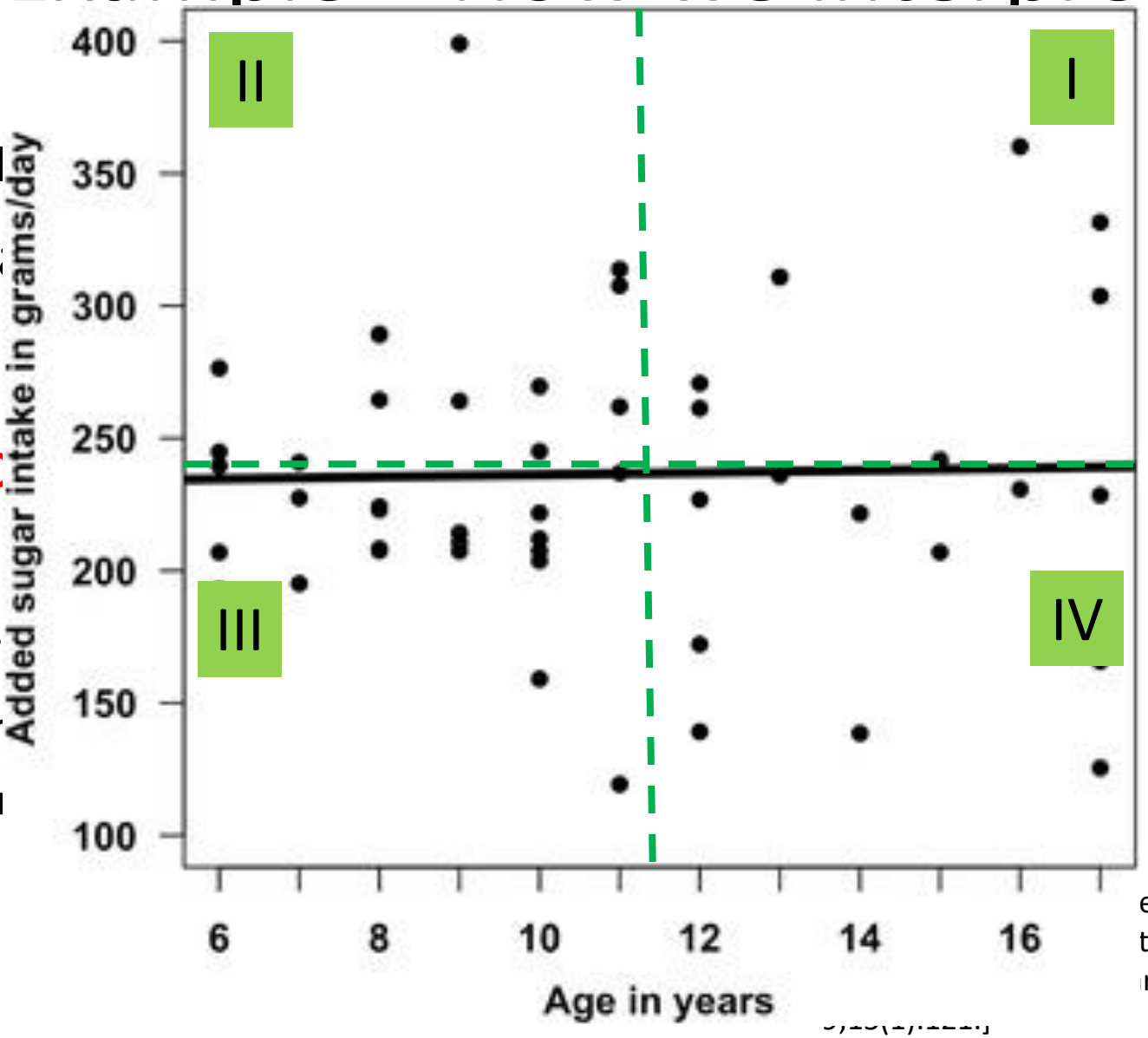
Interpretatic

near 0

- no corr

Colton rules r

between su

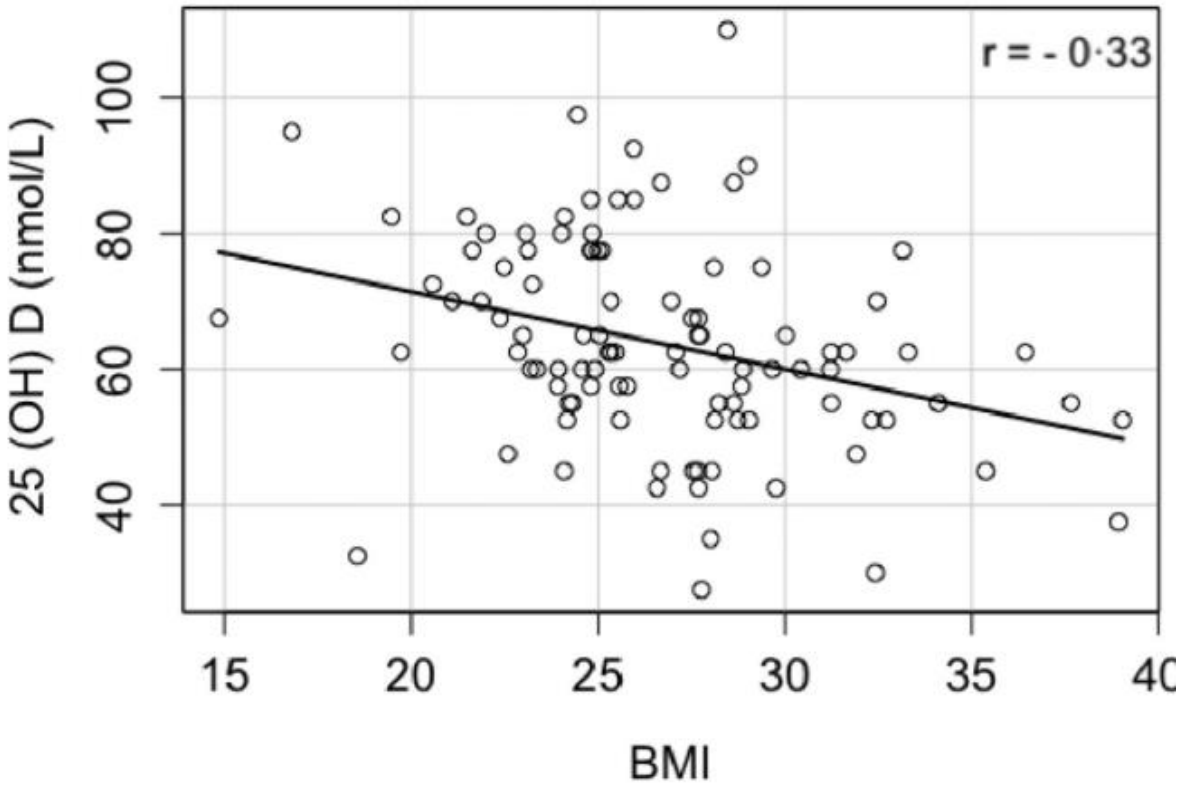


en D, Mancl L, Orr E, Lenaker D. Association
take and dental caries in Yup'ik children
rker. BMC Oral Health. 2015 Oct

Example – How we interpret r?

- Correlation between
  - Vitamin D2
  - Body mass index (BMI)

  r=-0.33



Interpretation:

Negative correlation, descending trend, indirect proportional correlated

- increase BMI – corresponds to smaller concentration of vitamin D2
- Dots are preponderant in quadrant II and IV

Colton rules r in (-0.50 to -0.25] →

between BMI and vitamin D2 there is a week correlation

Santana, Keila Valente de Souza de et al. Association between vitamin D status and lifestyle factors in Brazilian women: Implications of Sun Exposure Levels, Diet, and Health. eClinicalMedicine, Volume 47, 101400

Example – How we interpret r?

- Corre
  - Vita

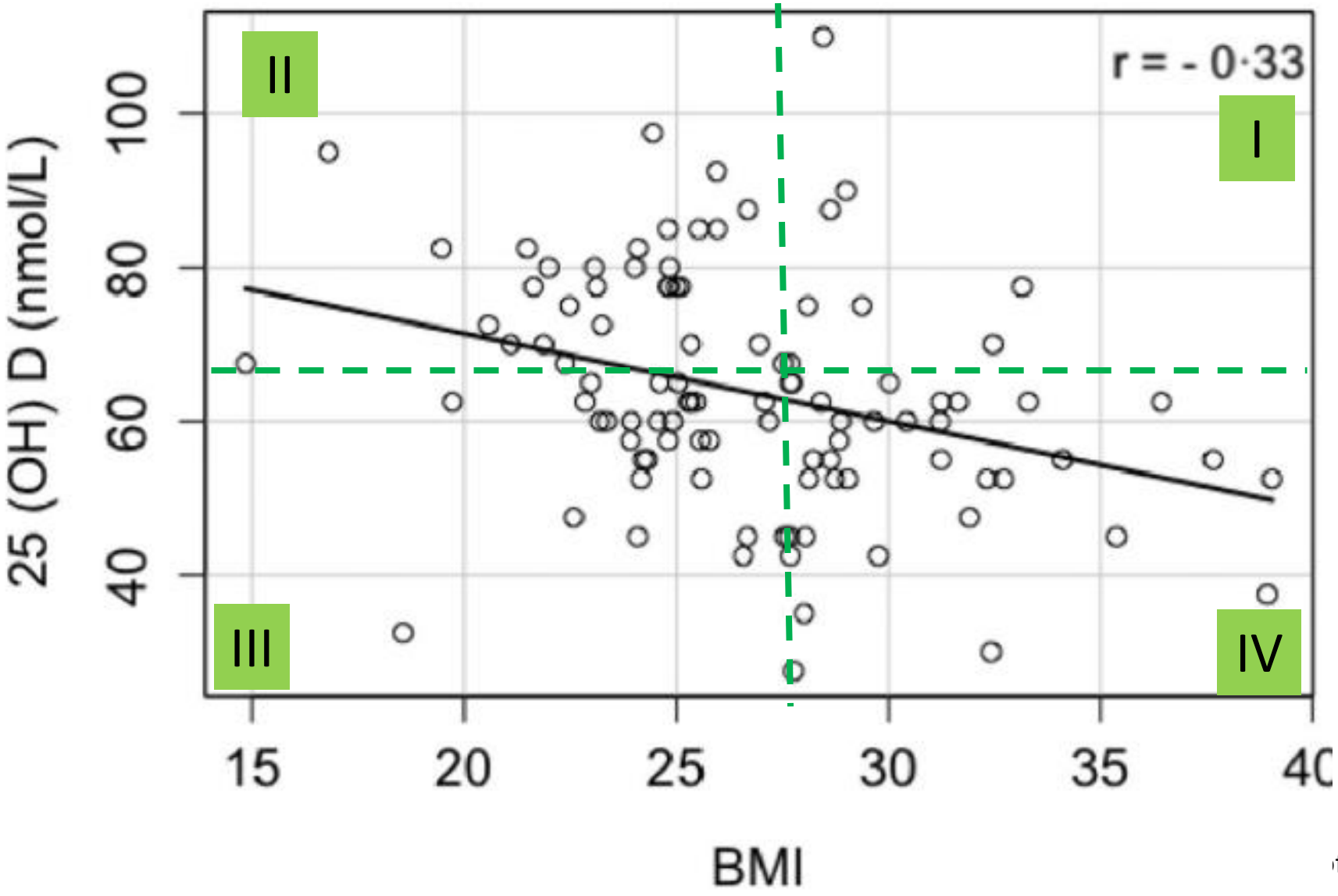
Interpre

Negat

- ir

Colton

betv



Santana, Keila Va f Sun Exposure Levels, Diet, and Health. eClinicalMedicine, Volume 47, 101400

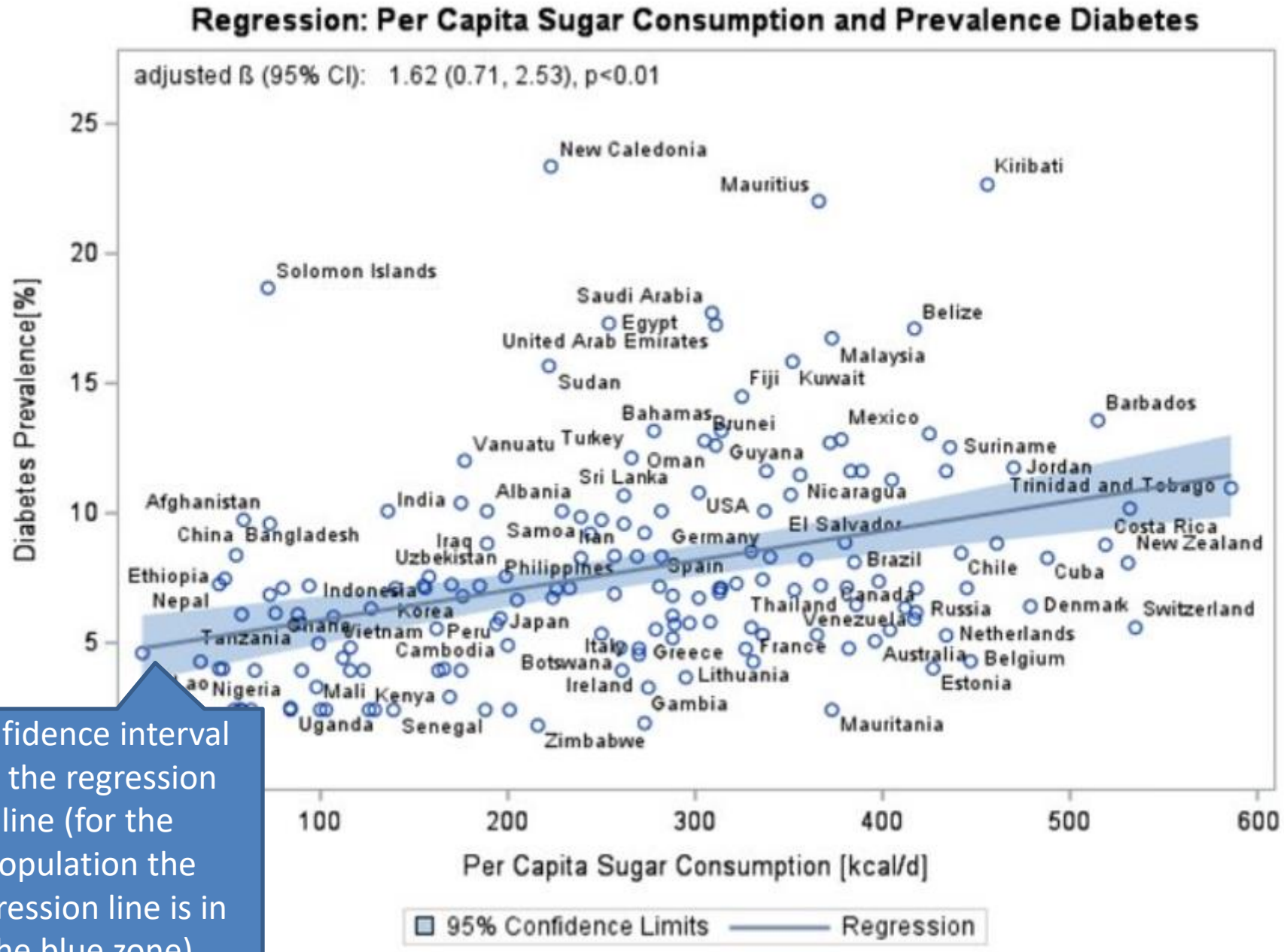
Linear regression

- Standard error
- If we repeat the study the regression line is different
- Variation of the regression line is called the standard error:

$$ES_{X,Y}=\sqrt{\frac{\sum_{i=1}^{n}(Y_{i-predicted}-Y_{i-observed})^2}{n-2}}$$

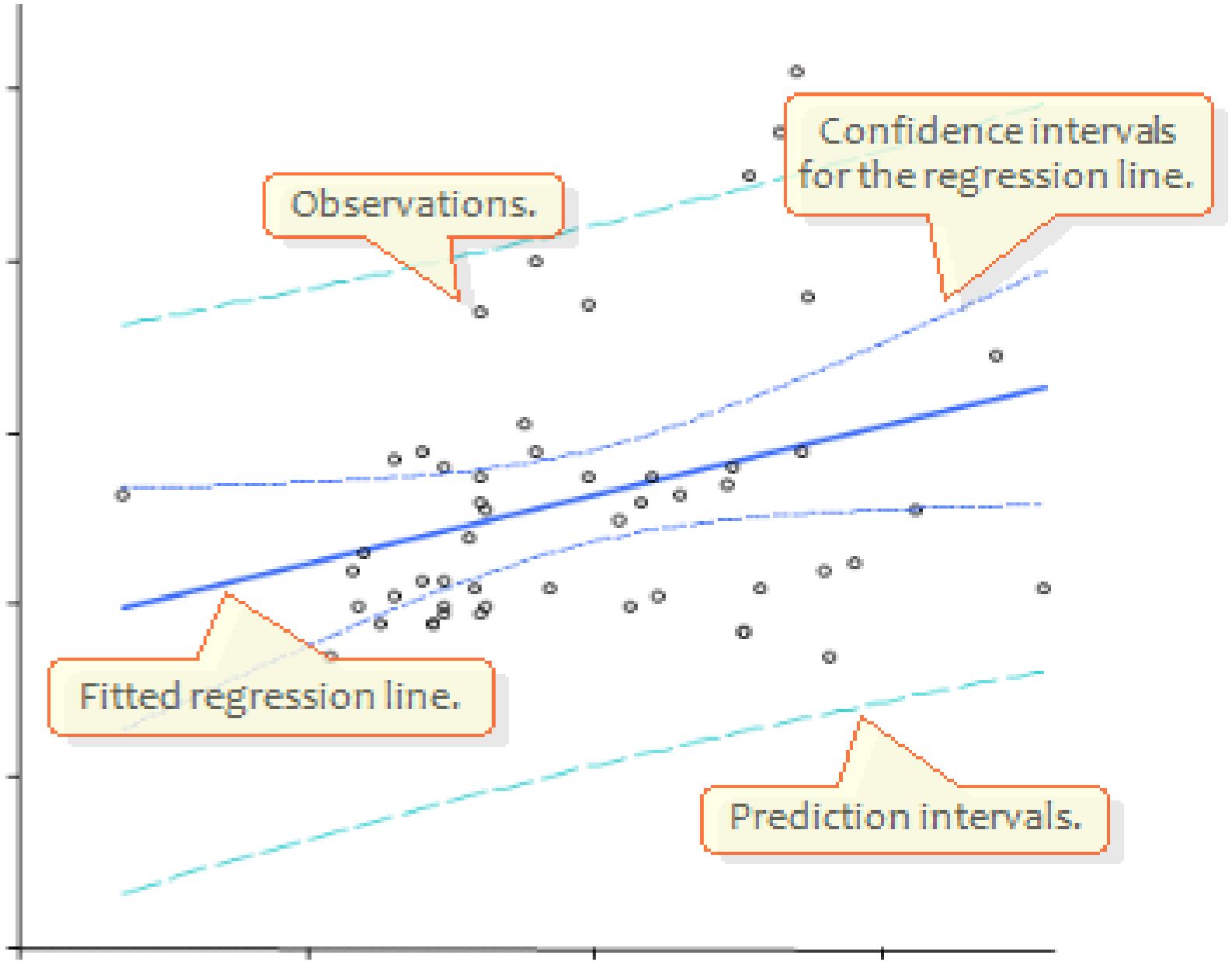
- Using standard error for calculating
  - Confidence interval of the regression coefficients: a, b

Lang A, Kuss O, Filla T, Schlesinger S. Association between per capita sugar consumption and diabetes prevalence mediated by the body mass index: results of a global mediation analysis. Eur J Nutr. 2021 Jun;60(4):2121-2129.



Confidence interval for the regression line (for the population the regression line is in the blue zone)

sugar consumption per capita (2007) and diabetes prevalence (2017) for 192 countries


Observations.
Confidence intervals for the regression line.
Fitted regression line.
Prediction intervals.

- Using standard error for
  - Testing the significance of each coefficient (t-test)
  - Predict the value of Y for an individual or for average

Model Coefficients - BMI (kg/m2)

Predictor	Estimate	SE	t	p
Intercept	24.6065	2.3932	10.282	< .001
Age (years)	0.0343	0.0386	0.888	0.377

BMI=0.03*AGE+24.61

b – was tatistically significant in prediction of BMI

a – was not statistically significant in prediction of BMI

  - AGE is not statistically significant in prediction of BMI

- Using standard error for
  - Testing the significance of each coefficient (t-test)
  - Predict the value of Y for an individual or for average

Model Coefficients - cGIM(mm)

Predictor	Estimate	SE	t	p
Intercept	0.67180	0.07250	9.27	< .001
Age (years)	0.00441	0.00117	3.77	< .001

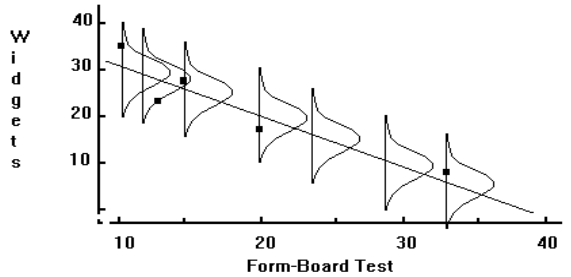
cGIM=0.004*AGE+0.67 – intima media thickness of carotid artery

b – was statistically significant in prediction of cGIM

a – was statistically significant in prediction of cGIM

  – AGE is statistically significant in prediction of cGIM

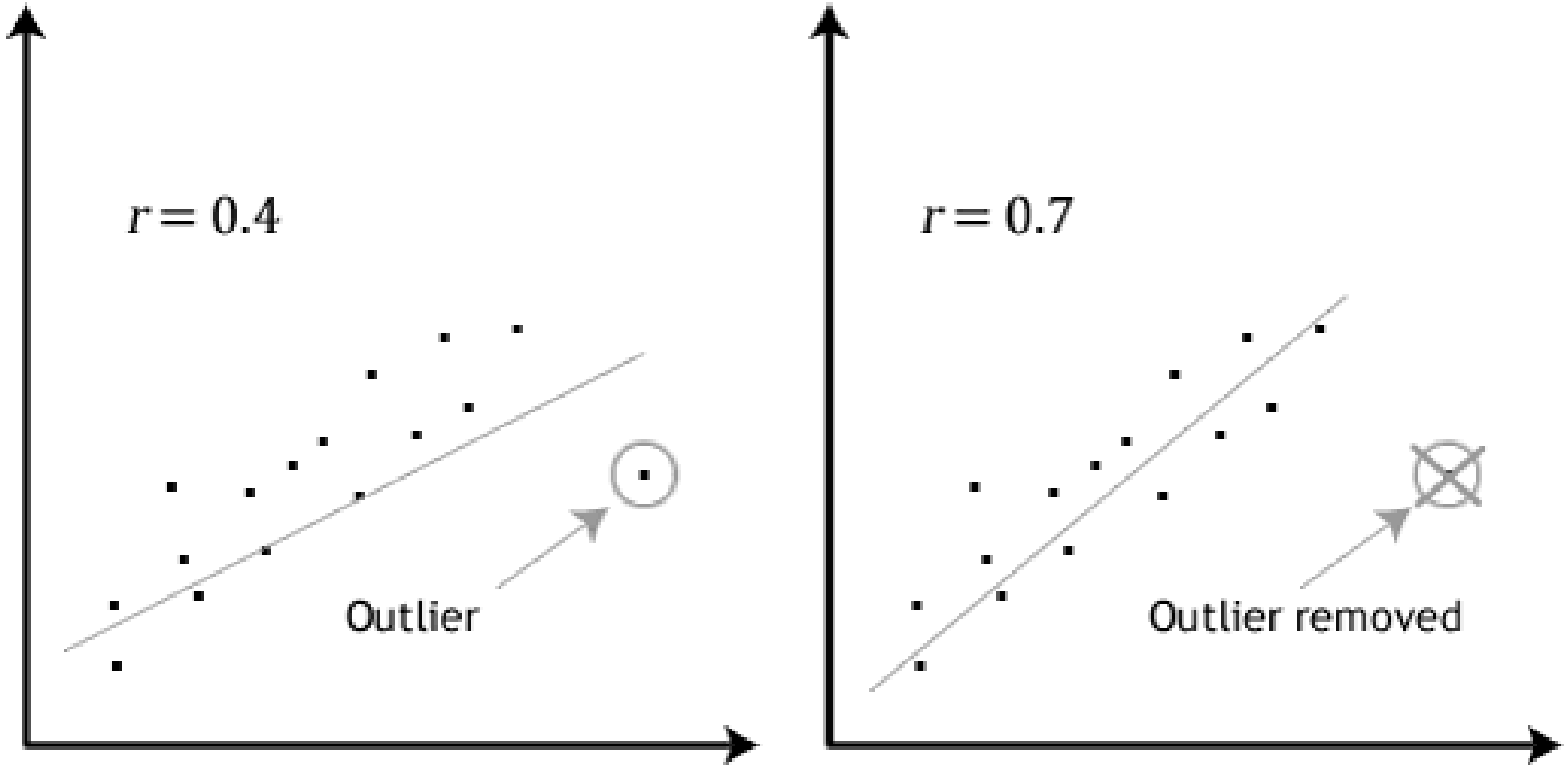
Requirements for linear correlation and for linear regression



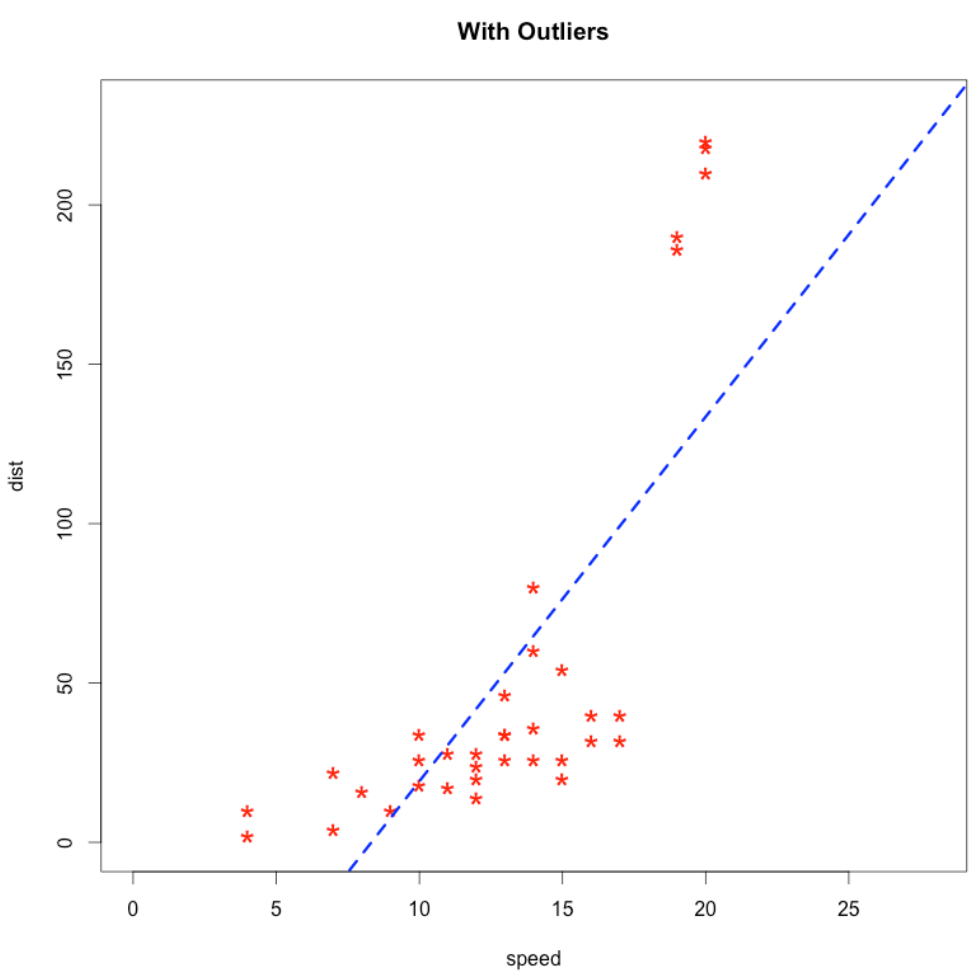
- No outliers for X and Y (normal distribution)
- Linear correlation
- Values of Y independent of each other
- Uniformity, homoscedasticity - equal variation of Y for each X

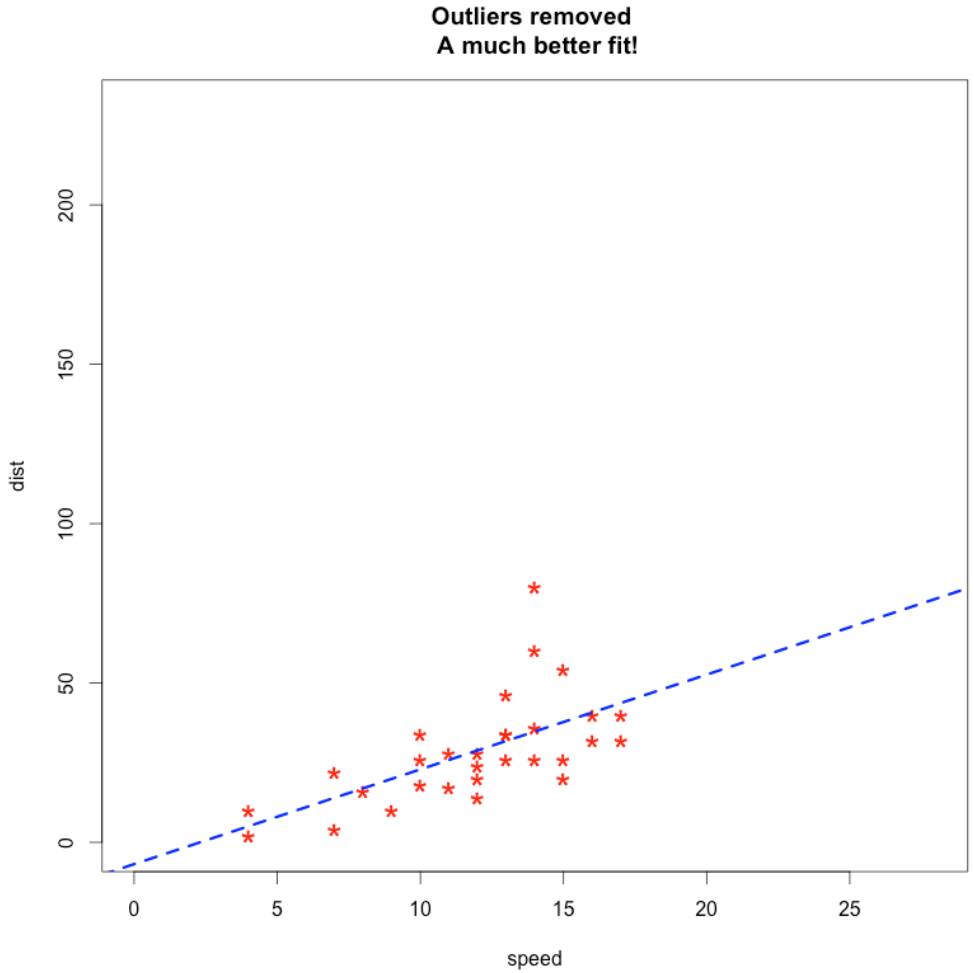
- Possible solutions in case that there are outliers:
  - Transformations: logarithm, normalization
  - Removing outlier cases from the analysis
  - Compute Spearman correlation coefficient

The effect of outliers on r – Pearson coefficient of correlation



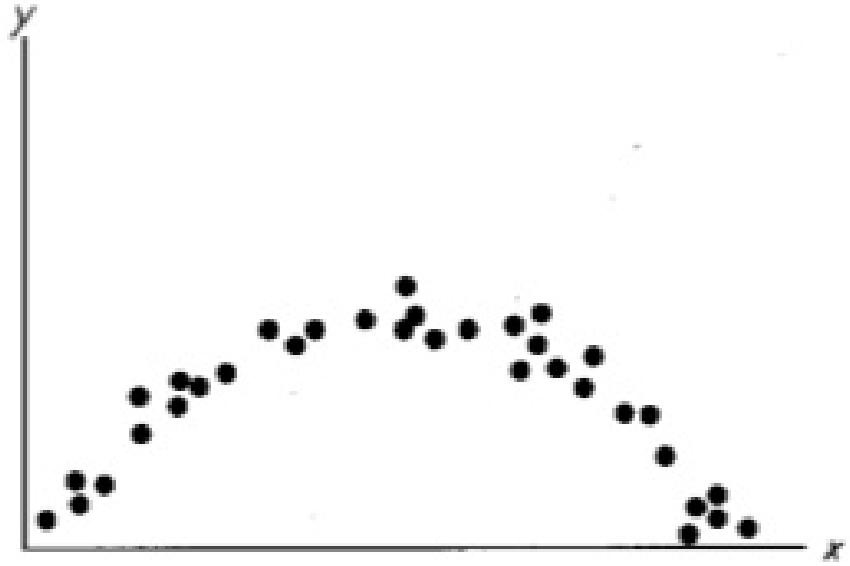
The effect of outliers on r – Pearson coefficient of correlation





Correlation coefficient and linear regression

- The coefficient is independent of the **measuring unit**, linear regression not
- When weight is measured in **kg or gram**, the regression line has a different coefficient, while the correlation is the same
- p - **The significance** of the correlation coefficient and the significance of the regression coefficient is the same
- The correlation coefficient and the coefficient of the regression line have the **same sign**



- There is **no linear relationship** between X and Y –
  - Spearman correlation coefficient will appreciate better the correlation

Spearman correlation coefficient

- The ranks for two quantitative variable X and Y: R_X and R_Y

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(R_{X,i} - \overline{R_X}\right)\left(R_{Y,i} - \overline{R_Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(R_{X,i} - \overline{R_X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(R_{Y,i} - \overline{R_Y}\right)^2}}$$

- It can be calculated for two ordinal variables also

The Spearman correlation coefficient

- indicates the association between X and Y
- Always between -1 and 1

$$r \in [-1,1]$$

As $|\ r\ |$ approaches 1 the association is greater

As $|\ r\ |$ approaches 0 the association is less

! We are talking about **no** linear relationship

The Spearman correlation coefficient

- If r>0 than the association between X and Y is **positive** (direct, but not linear)
- If r<0 than the association between X and Y is **negative** (inverse, but not linear)
- Colton rules apply

Statistically significant correlation

- T test for correlation
- Objective: to estimate the Spearman correlation coefficient in the population: ρ
- The estimation is based on a random sample of the population where r was calculated.

Significant statistically

- **H0 (null hyphotesis):**
  - the Spearman coefficient of correlation is not statistically significantly different than 0
- **H1 (alternative hyphotesis):**
  - the Sperman coefficient of correlation is statistically significantly different than 0

- Result: p

- If $p<0.05$ then we reject the null hypothesis (H0) and accept the alternative hypothesis (H1): the correlation is statistically significant
- If $p\geq0.05$ then we fail to reject the null hypothesis (H0): the correlation is not statistically significant

Correlation Matrix

		BMI (kg/m2)
BMI (kg/m2)	Spearman's rho	—
	df	—
	p-value	—
Waist circumferences (cm)	Spearman's rho	0.626 ***
	df	98
	p-value	< .001

Note. * p < .05, ** p < .01, *** p < .001

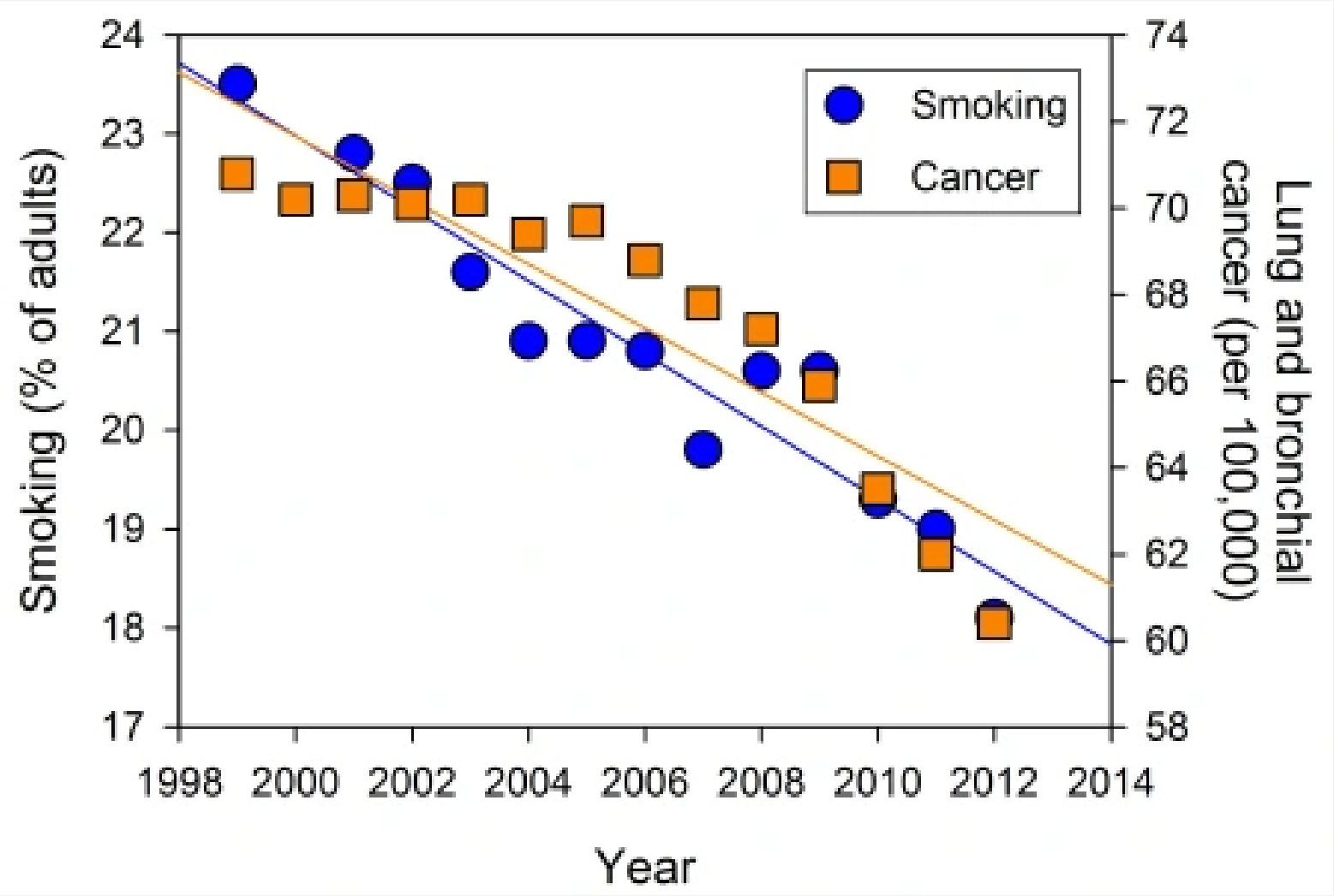
- p<0.05, we reject H0 and accept H1: the Spearman coefficient of correlation between waist circumference and BMI is statistically significant

Causation is not causation

- Causation = one event is the result of the occurrence of the other event

- Correlation ≠ not automatically mean causation

- To demonstrate that X is the cause of Y
  - to demonstrate that X is correlated with Y
  - the occurrence of X is before the occurrence of Y
  - if we remove X, Y will be removed/improved also



Smoking and lung/bronchial cancer rates (data via the CDC). P < 0.0001

- A demonstrated causation
- Lung cancer and smoking

Dependent

- if we suppose that X is the cause of Y, then Y is **dependent** from X

decrease prevalence of smoking → decrease prevalence of lung cancer

- X = independent variable (prevalence of smoking)
- Y = dependent variable (prevalence of lung cancer)

Scenario

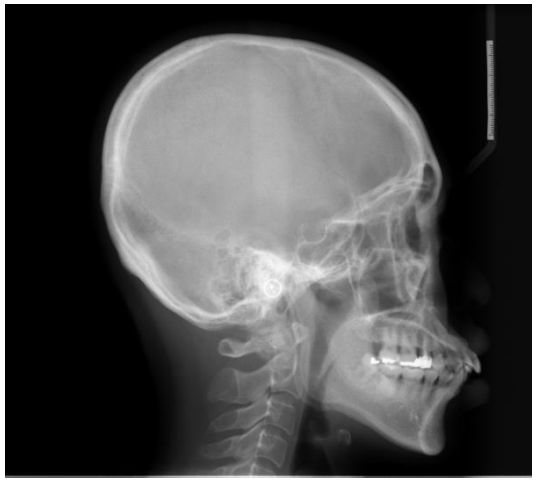
Class II Division I Malocclusion

- Skeletal Class II malocclusion is a dentofacial deformity caused by a growth disorder of the bones frequently associated with mandibular retrusion relative to upper facial structures and with airways problems.
- “The aim of this study was to investigate the association between upper airways measurements with the length of the mandible”.

[Silva NN, Lacerda RH, Silva AW, Ramos TB. Assessment of upper airways measurements in patients with mandibular skeletal Class II malocclusion. Dental Press J Orthod. 2015 Oct;20(5):86-93.]

Scenario

- A total of 80 lateral cephalograms from 80 individuals aged between 10 and 17 years old were assessed. 40 radiographs of Class I malocclusion individuals were matched by age with 40 radiographs of individuals with mandibular Class II malocclusion.
- mandibular length (Xi-Pm, Co-Gn and Go-Me)
- mandibular position (facial depth and SNB).



[Silva NN, Lacerda RH, Silva AW, Ramos TB. Assessment of upper airways measurements in patients with mandibular skeletal Class II malocclusion. Dental Press J Orthod. 2015 Oct;20(5):86-93.]

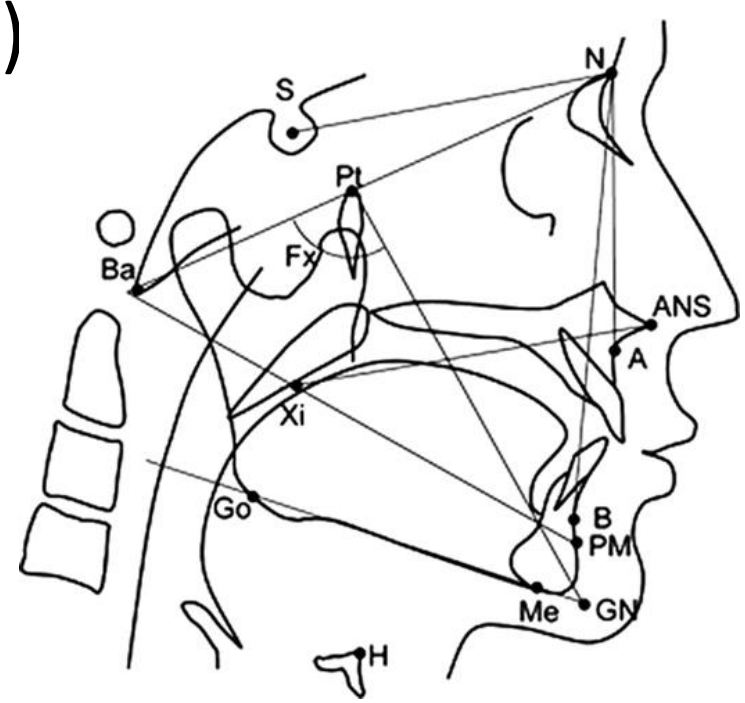


Table 4

[Silva NN, Lacerda RH, Silva AW, Ramos TB. Assessment of upper airways measurements in patients with mandibular skeletal Class II malocclusion. Dental Press J Orthod. 2015 Oct;20(5):86-93.]

- Correlation between upper airways measurements and mandibular length, position as well as direction of mandibular growth in both groups.

Measures	r	P	%	Interpretation
Oropharynx				
Xi-Pm	0.31	0.004	9.6%	Significant, positive and moderate correlation
Co-Gn	0.24	0.02	5.7%	Significant, positive and weak correlation
Go-Me	0.13	0.23	1.6%	There was no correlation between variables
Facial depth	0.21	0.06	4.4%	There was no correlation between variables
SNB	0.37	0.001	13.6%	Significant, positive and moderate correlation

- Thank you!