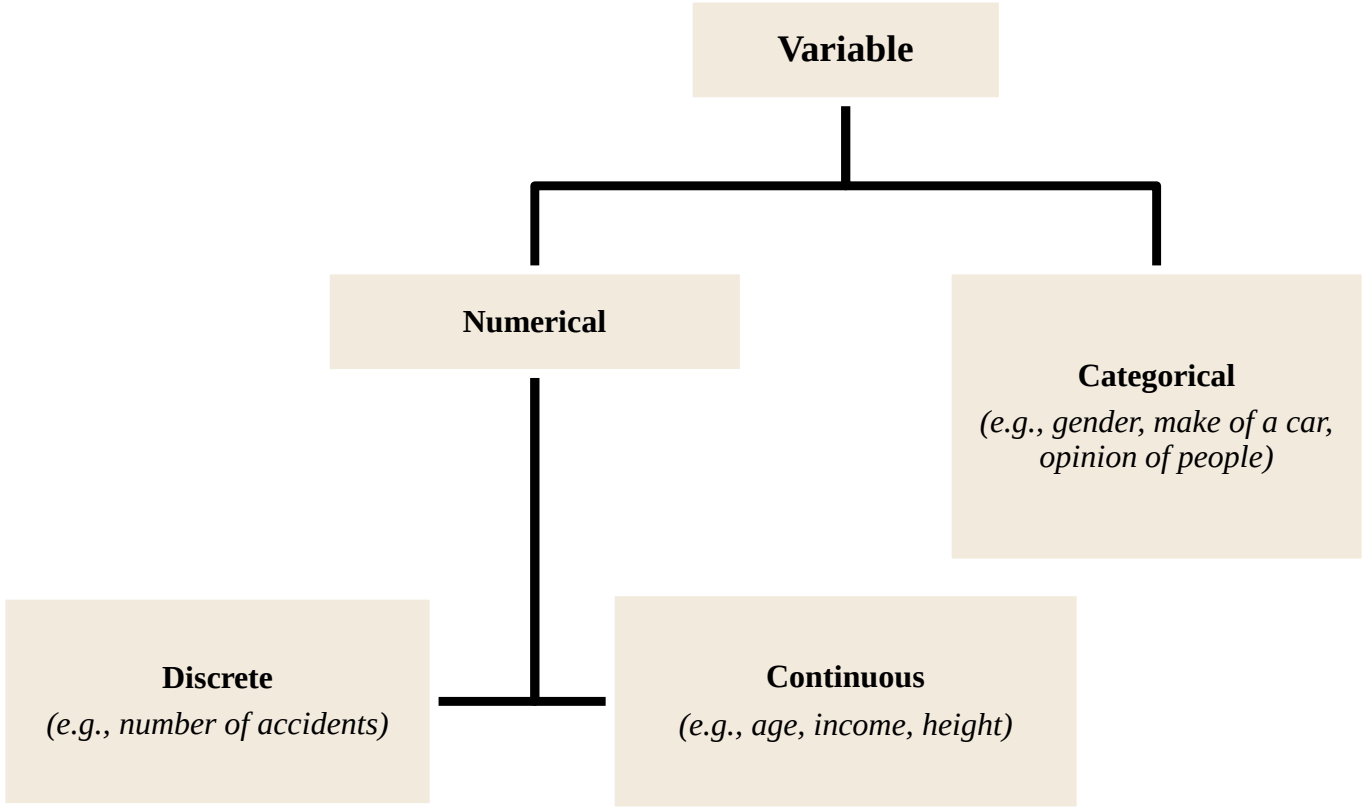
Measures of association

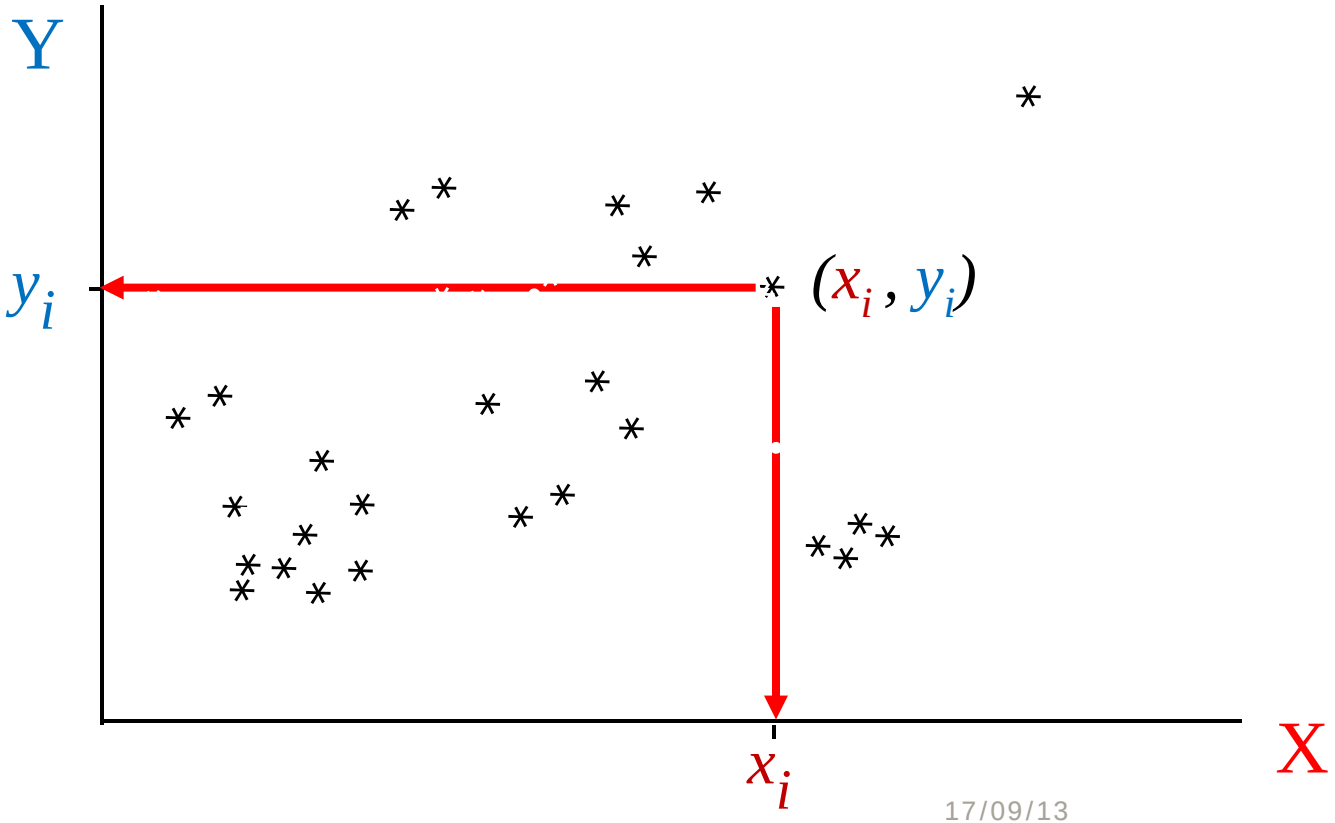
Types of variables



Numerical

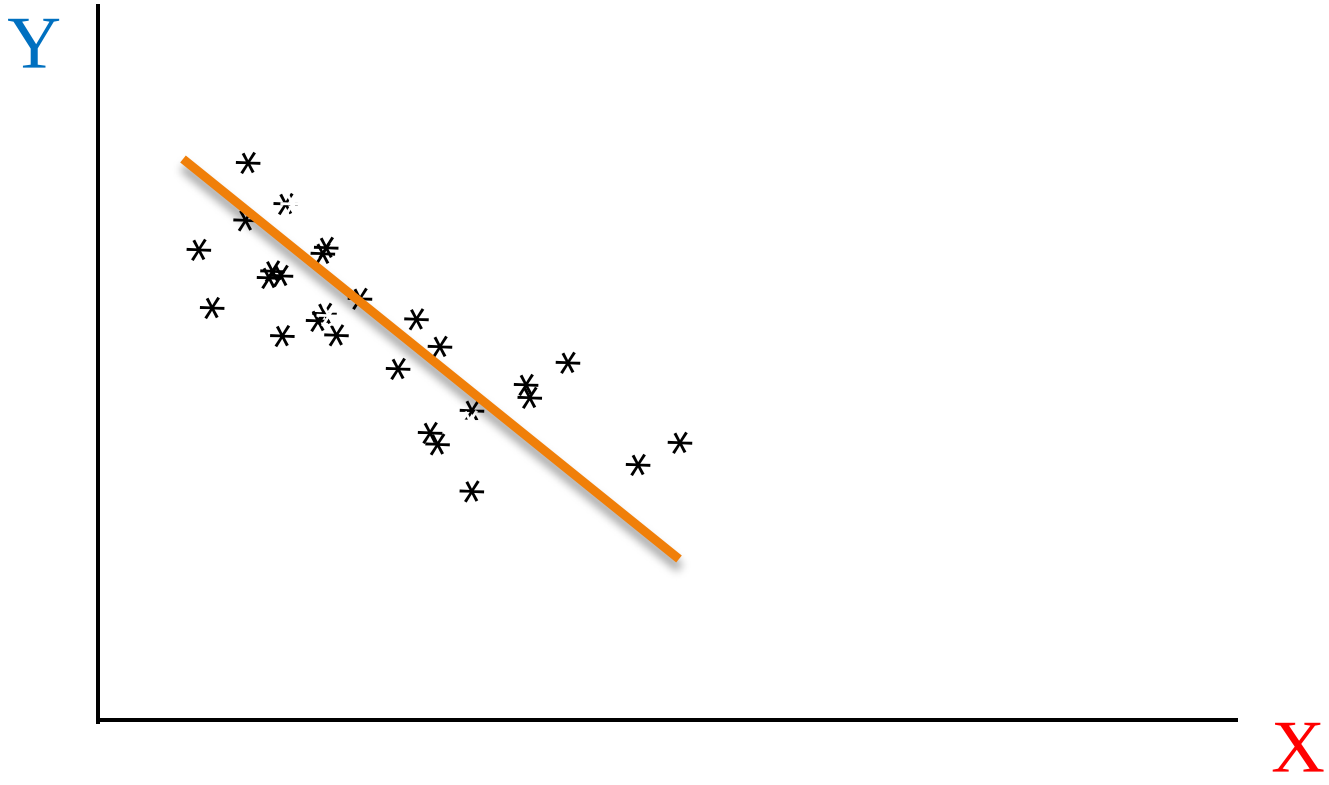
Numerical: Scatter plots

Explore relationship between two quantitative variables (X and Y)



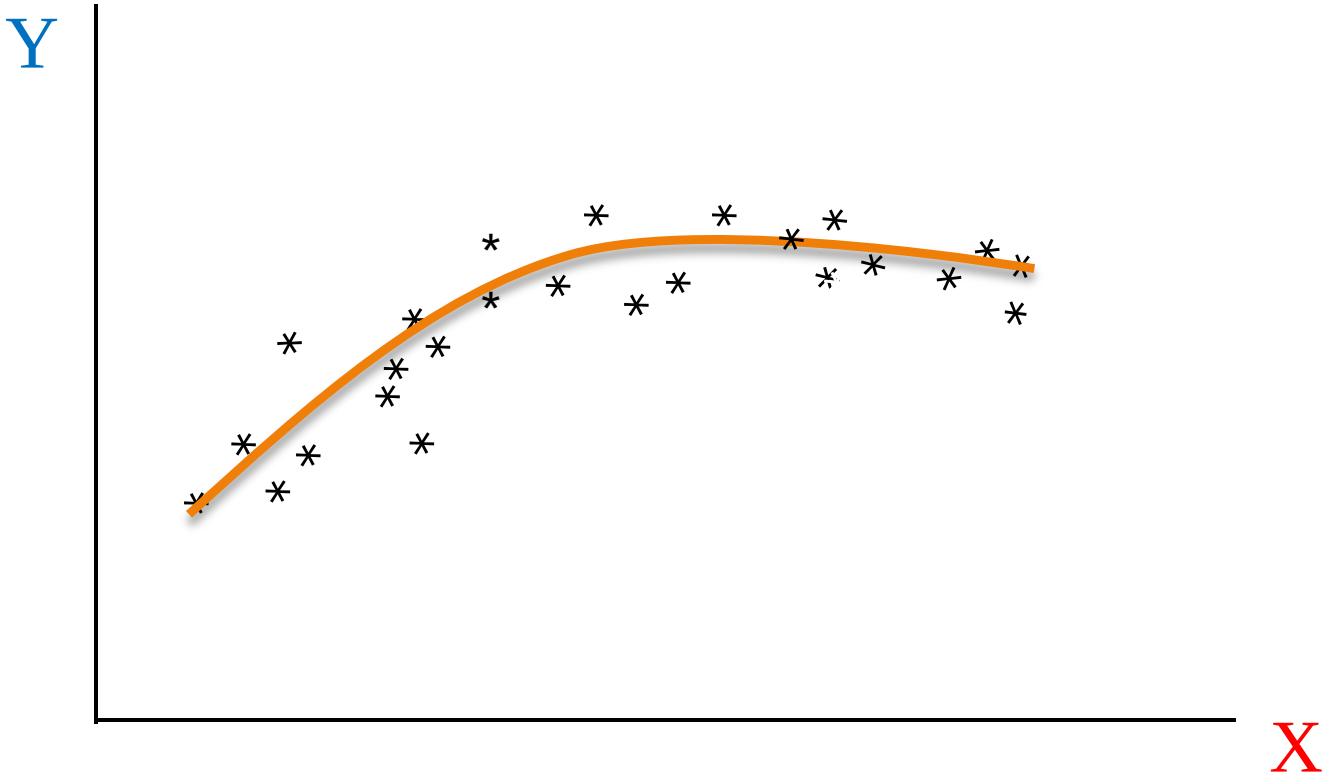
A linear relationship

- Linear pattern; moving on the X axis to the right, the Y values change in the almost same direction with approximately same rate



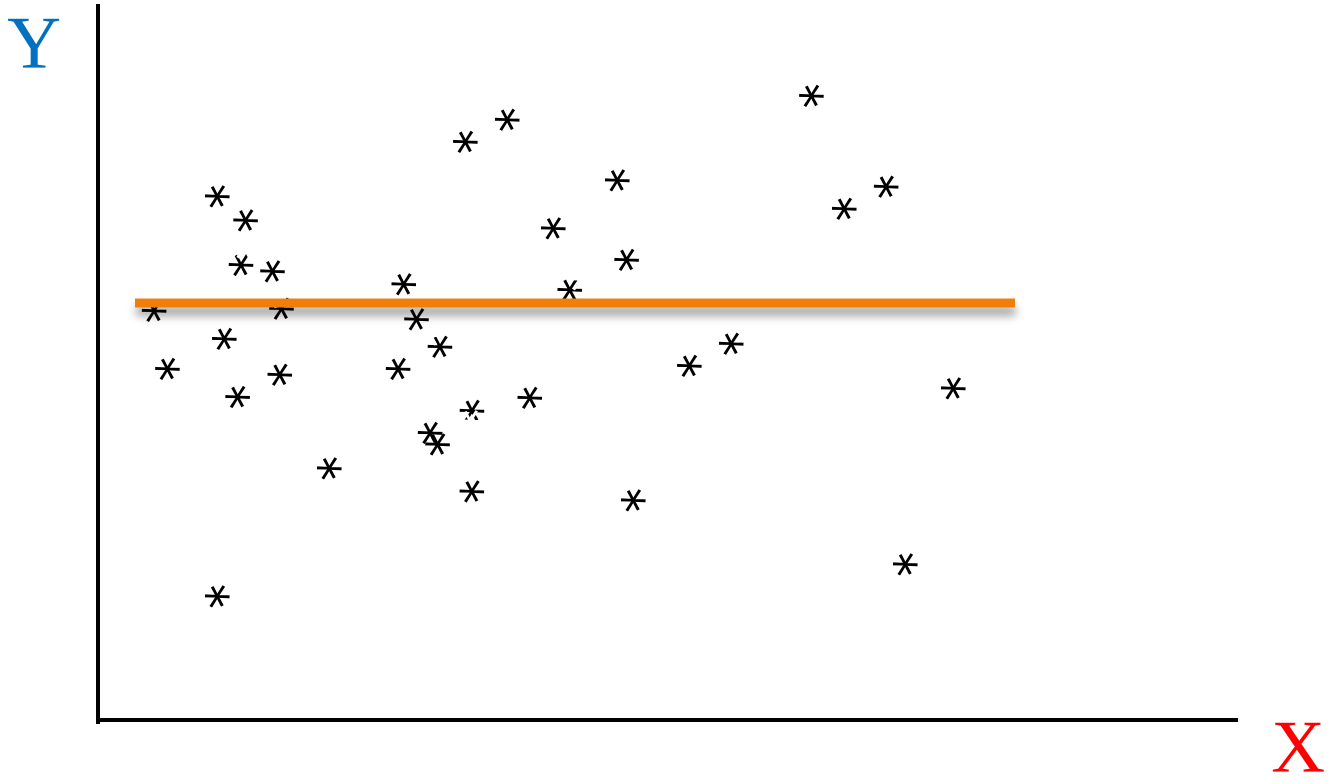
Non-linear relationship

- Non-Linear pattern



No relationship

- No pattern



When linearly related, how closely X and Y change together?

- Correlation: Measure of how 'close' two quantitative variables are to being *linearly related*
-
- Correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

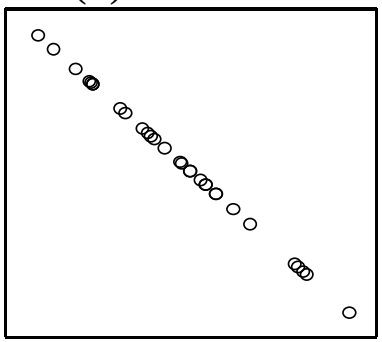
r estimates the correlation between X and Y in the population (denoted by ρ) from sample observations (x_i, y_i)

$$(-1 < r < +1)$$

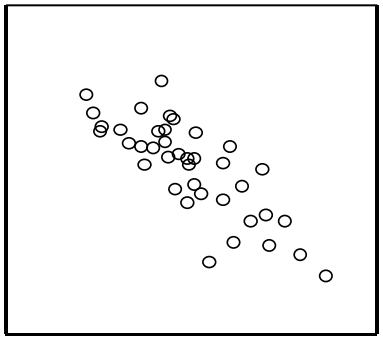
- If $r = 1$, then X and Y have a perfect *positive linear* relationship.
- If $r = -1$, then X and Y have a perfect *negative linear* relationship.
- If $r = 0$, then there is *no linear* relationship between X and Y.

Negative

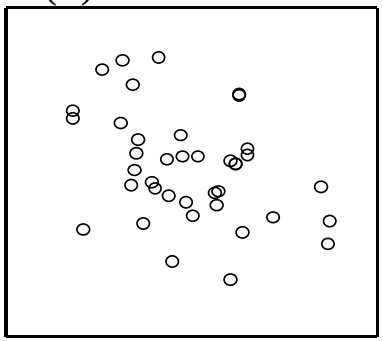
(a) $r = -1$



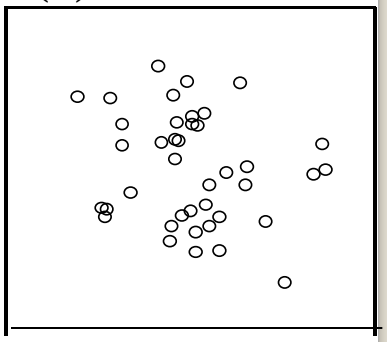
(b) $r = -0.8$



(c) $r = -0.4$



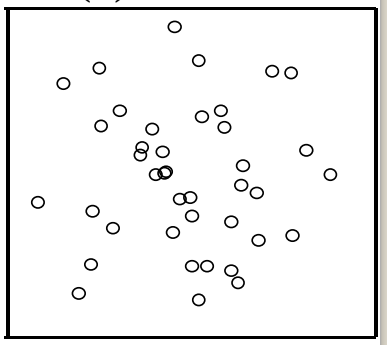
(d) $r = -0.2$



Perfect correlation

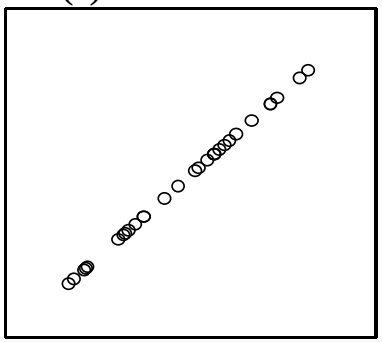
Becoming weaker

(e) $r = 0$

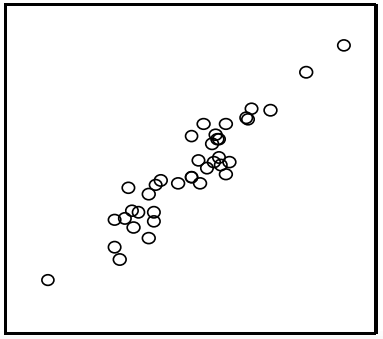


Positive

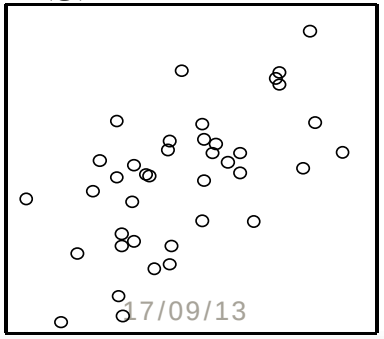
(i) $r = +1$



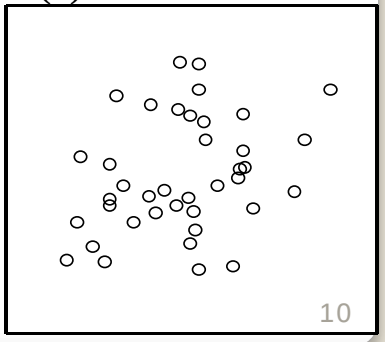
(h) $r = +0.95$



(g) $r = +0.6$



(f) $r = +0.3$

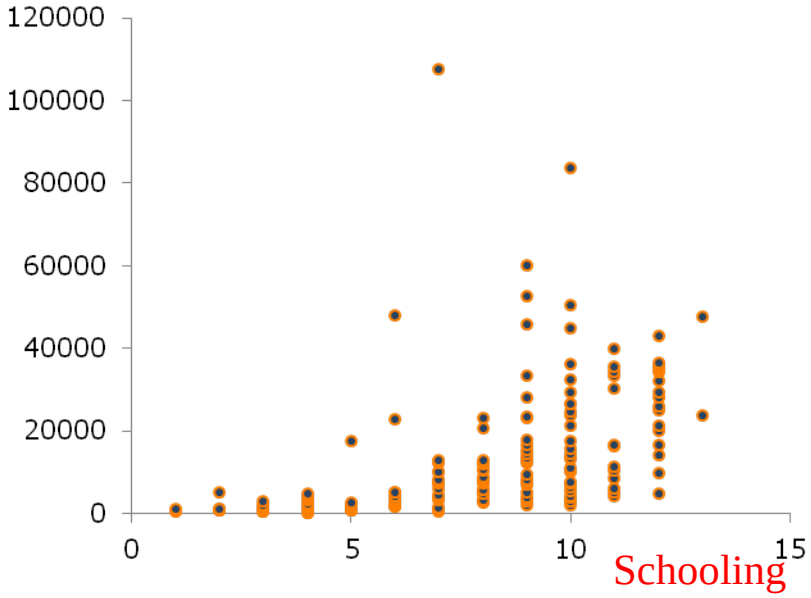


Transformations

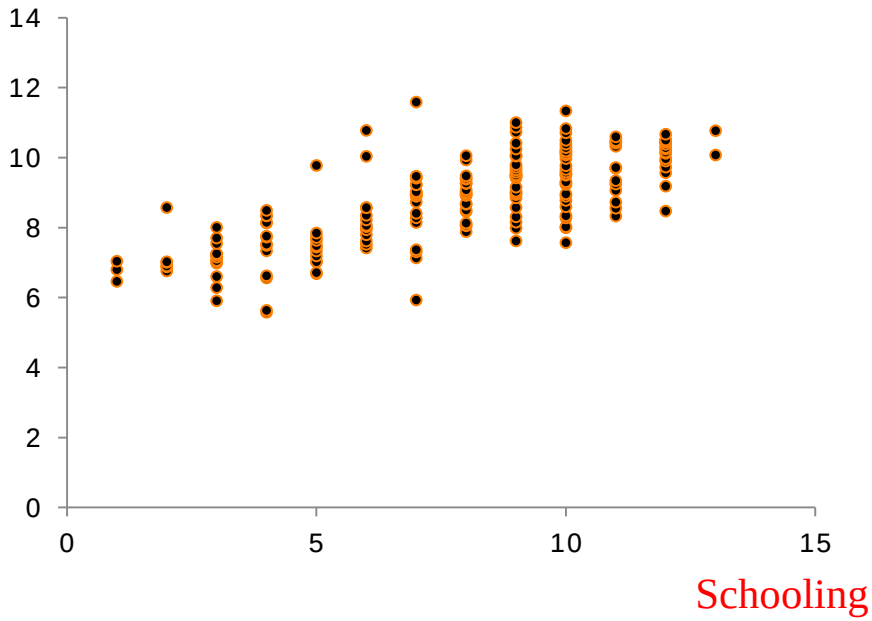
- In some problems either or both of Y and X can be replaced by transformations so that the scatter plots better describe the relationship

- Logarithmic transformation is the most frequently practiced

GNI



Ln(GNI)



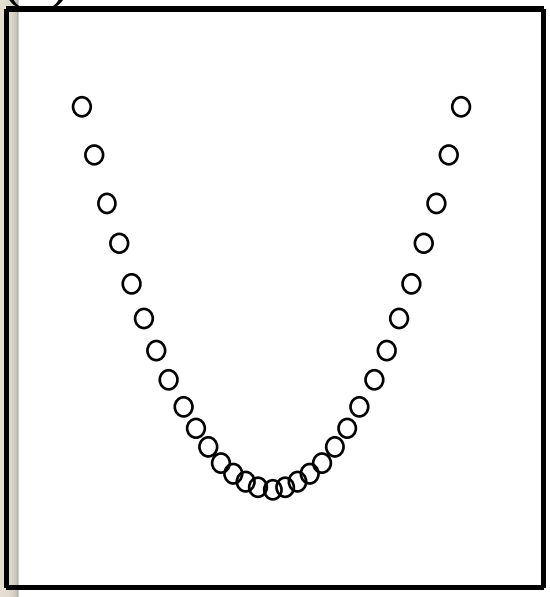
Data source: Average years of schooling and GNI for 187 countries (2011, HDI dataset,UNDP)

No relationship

r=0

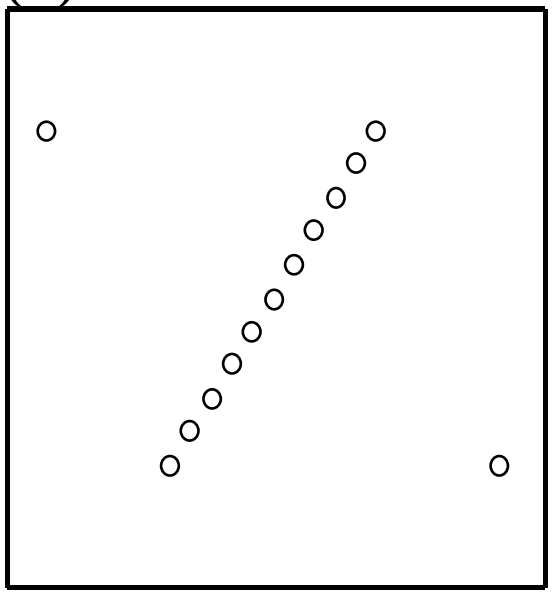
r=0

(a)



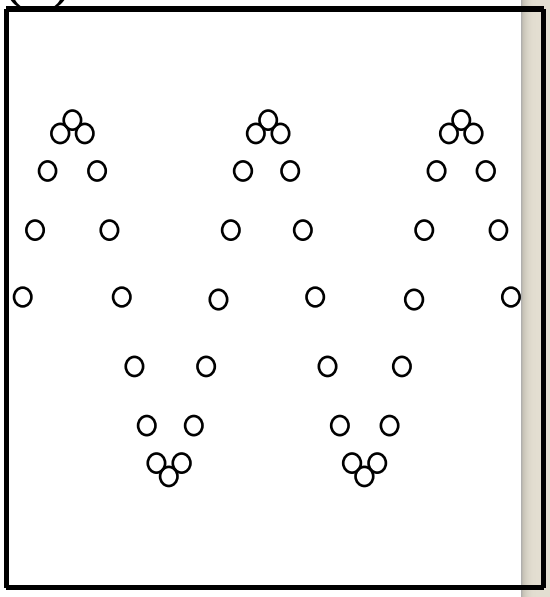
r=0

(b)



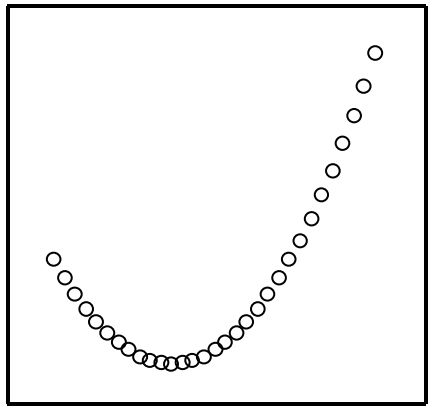
r=0

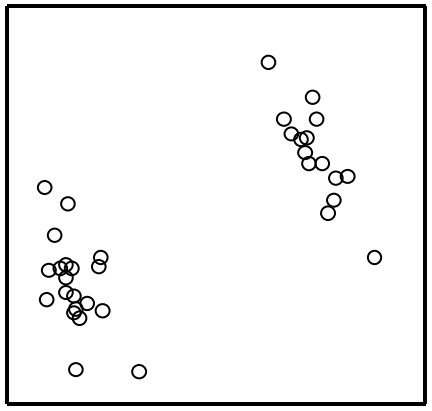
(c)

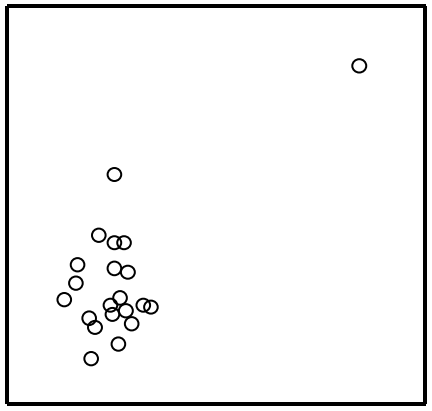


High relationship

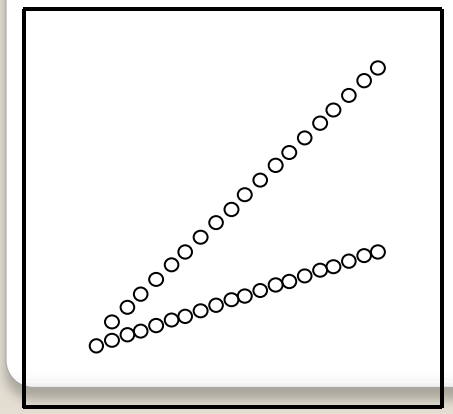
High value of r

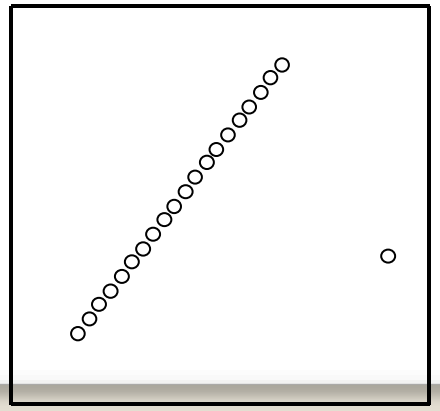


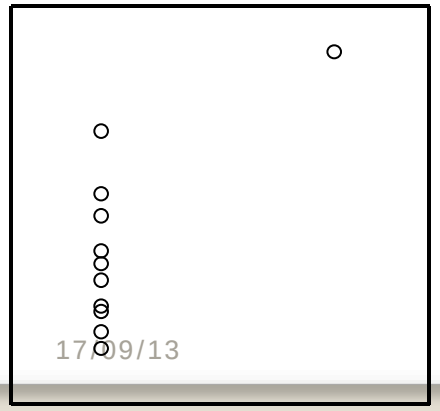




r=0.7 for all







while interpreting

- Correlation coefficients make sense only for studying linear relationships.
- When interpreting values of r, also look at scatter plot.

Cause-and-effect relationship

- Strong linear relationships do not necessarily mean a cause-and-effect relationship
- Suspect! a third variable might have linked X and Y which we have not measured or even thought about.

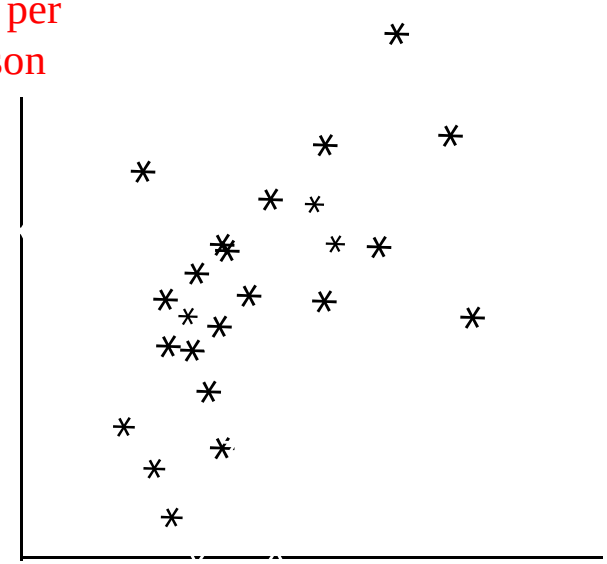
Example

SO what?

Lengthen lives of people in Zambia by shipping TV sets to them?!!!!!!

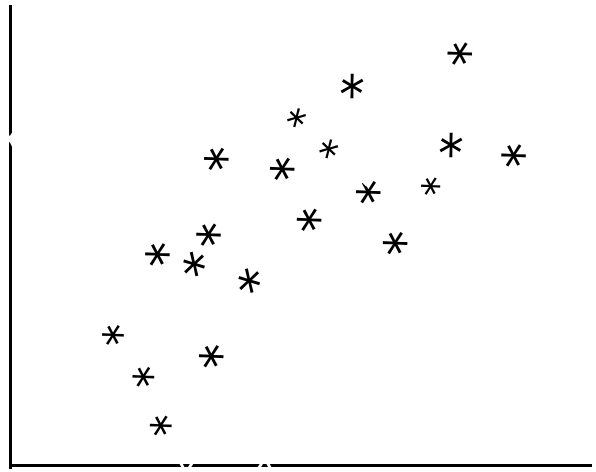
Example

of TV sets per person



GDP per capita

Life expectancy



GDP per capita

NO!! The higher the income, the more TV per person as well as better health services and longer life

Categorical

Contingency tables

sex	opinion		Row total
	Agree	disagree	
M	50	70	120
F	50	30	80
Column total	100	100	200

Use Chi-square test to check the hypothesis of:

Opinion does not depend on sex

In this case: we reject the hypothesis