Statistics for business and decision making

Prof. Ilaria Benedetti

04. Correlation among quantitative variables

Correlation

Measures the strength and direction of a relationship between two *metric* variables

For example, the relationship between weight and height or between consumption and price: it is not a “perfect” (deterministic) relationship, but we expect to find one *on average*

Correlation

Correlation measures to what extent two (or more) variables are related

Correlation expresses a relationship. This relationship can be:	**Positive** correlation indicates that the two variables move in the same direction	**Negative** correlation indicates that they move in opposite directions

The question is: do two variables “move together”?

Covariance

- Measures the co-movement of two variables *x* and *y* across observations
- Covariance estimate:

$$COV(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- For each observation *i*, a situation where both *x* and *y* are above or below their respective sample means increase the covariance value, while the situation where one of the variables is above the sample mean and the other is below decreases the total covariance.
- Contrarily to variance, covariance can assume both positive and negative values.
- If *x* and *y* always move in opposite direction, all terms in the summation above will be negative, leading to a large negative covariance. If they move always in the same direction there will be a large positive covariance.

Interpreting Covariance

- **Covariance** between two variables:

$Cov(x,y) > 0 \longrightarrow$ x and y tend to move in the same direction

$Cov(x,y) < 0 \longrightarrow$ x and y tend to move in opposite directions

$Cov(x,y) = 0 \longrightarrow$ x and y are independent

From covariance to correlation

- Covariance, like variance, depends on the measurement units.
- If one measures prices in dollars and consumption in ounces, a different covariance value is obtained as compared to the use prices of Euros and consumption in Kilograms, even if both situation refer exactly to the same goods and observations.
- Some form of normalization is needed to avoid the measurement unit problem
- The usual approach is **standardization**, which requires subtracting the mean and dividing by the standard deviation.

Correlation

- Considering the covariance expression, where the numerator is based already on differences from the means, all that is required is dividing by the sample standard deviations, for both *x* and *y*.

$$\boldsymbol{CORR(x,y) = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \cdot \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}} =$$

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

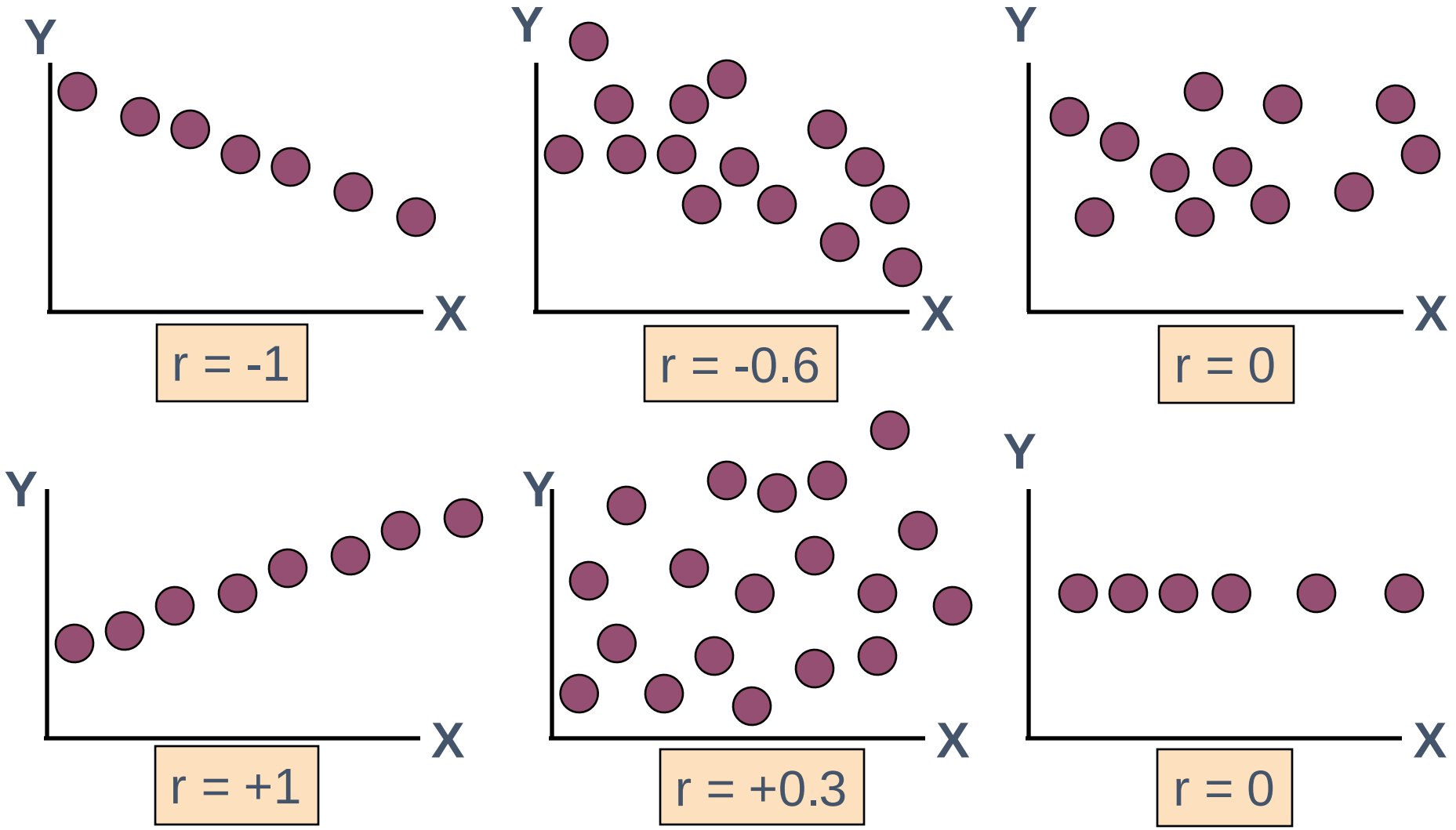
Correlation coefficient

- The correlation coefficient ***r*** gives a measure (in the range –1, +1) of the relationship between two variables
  - r=0 means *no correlation*
  - r=+1 means *perfect positive correlation*
  - r=-1 means *perfect negative correlation*
- Perfect correlation indicates that a *p%* variation in *x* **always** corresponds to a *p%* variation in *y*

Correlation and causation

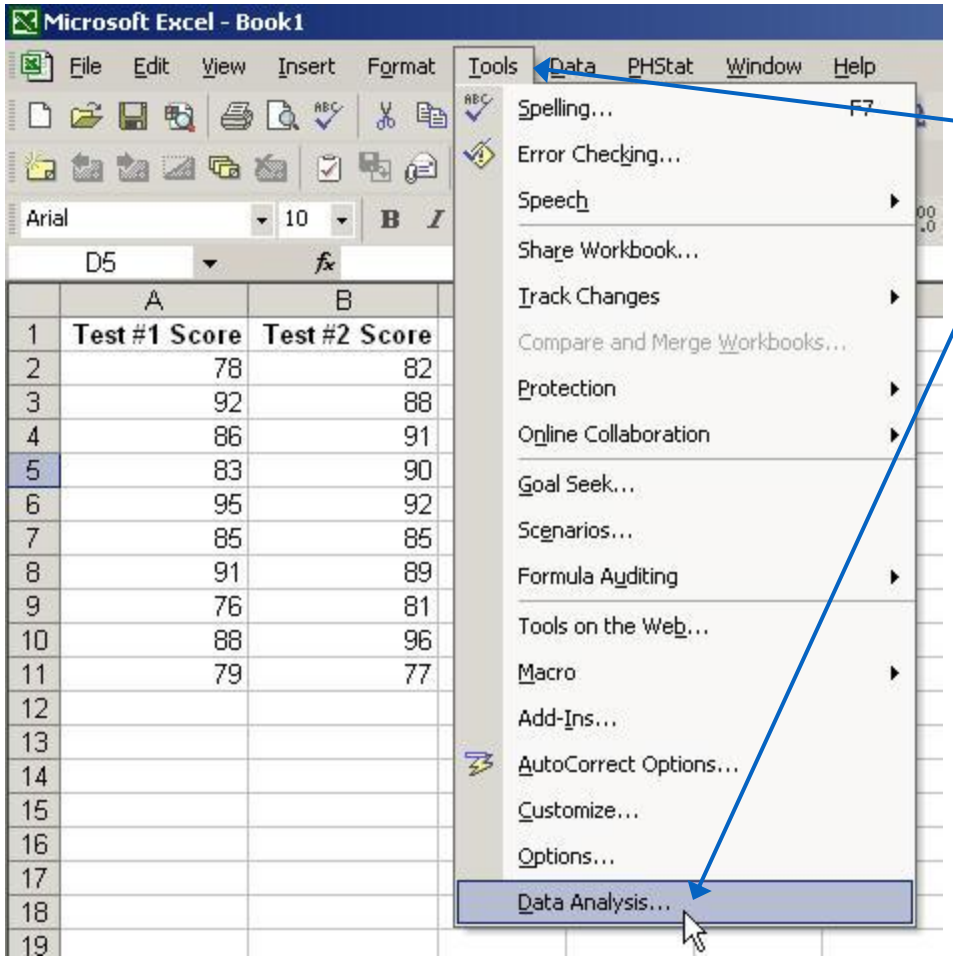
- Note that no assumption or consideration is made on *causality*,
- The existence of a positive correlation of *x* and *y* does not mean that is the increase in *x* which leads to an increase in *y*, but only that the two variable **move together**, to some extent.
- Thus, correlation is symmetric, so that $r_{xy} = r_{yx}$

Scatter Plots of Data with Various Correlation Coefficients

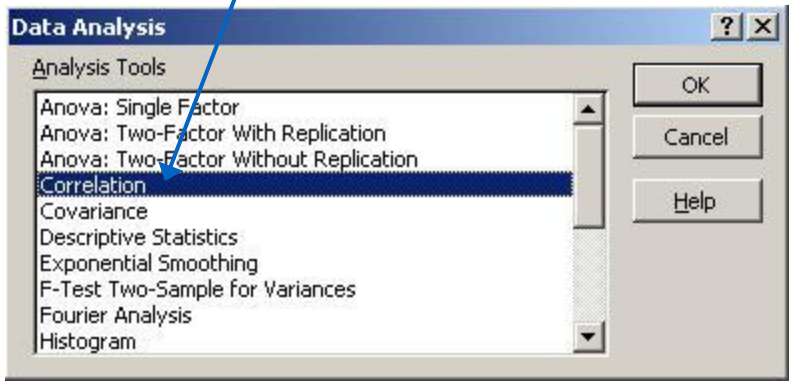


Using Excel to Find the Correlation Coefficient

Test#1	Test#2
78	82
92	88
86	91
83	90
95	92
85	85
91	89
76	81
88	96
79	77

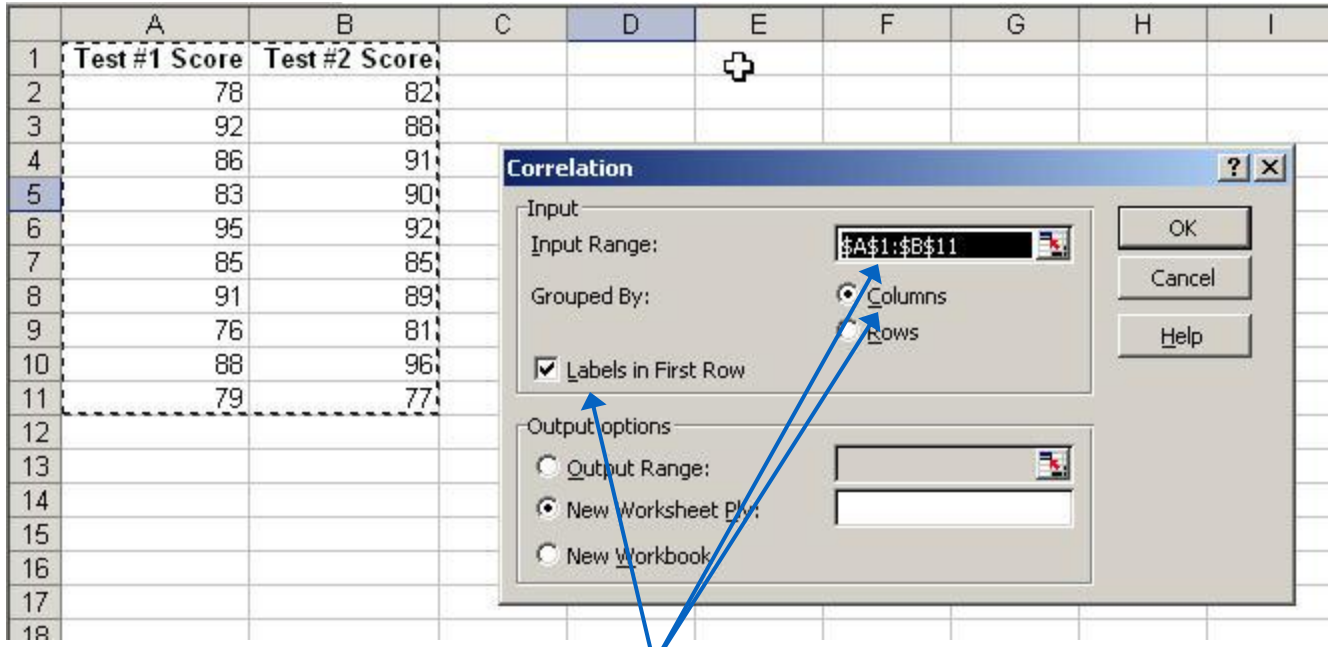


- Select Tools/Data Analysis
- Choose Correlation from the selection menu
- Click OK . . .



Using Excel to Find the Correlation Coefficient

(continued)

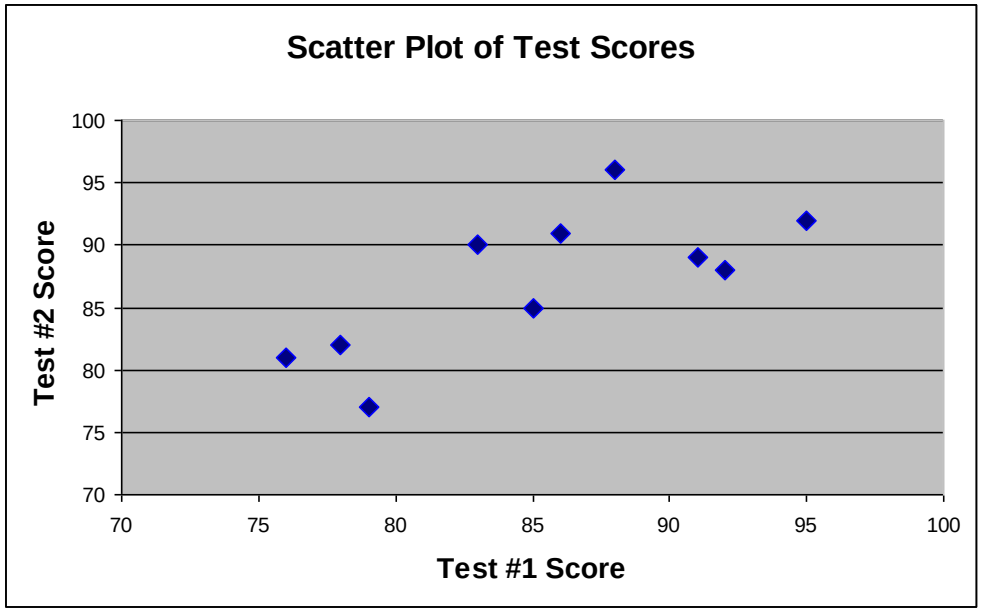


- Input data range and select appropriate options
- Click OK to get output

	A	B	C
1		*Test #1 Score*	*Test #2 Score*
2	Test #1 Score	1	
3	Test #2 Score	0.733243705	1
4			

Interpreting the Result

- r = .733
- There is a relatively strong positive linear relationship between test score #1 and test score #2



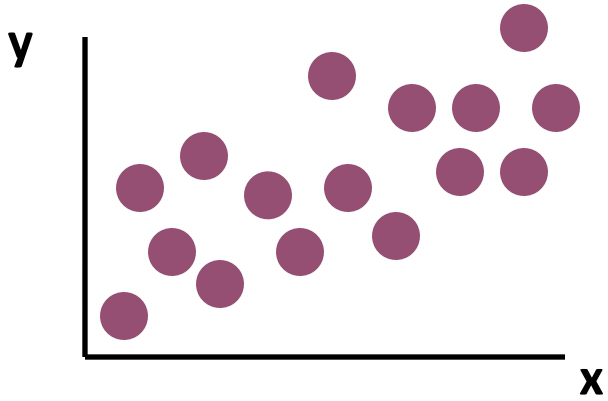
- Students who scored high on the first test tended to score high on second test

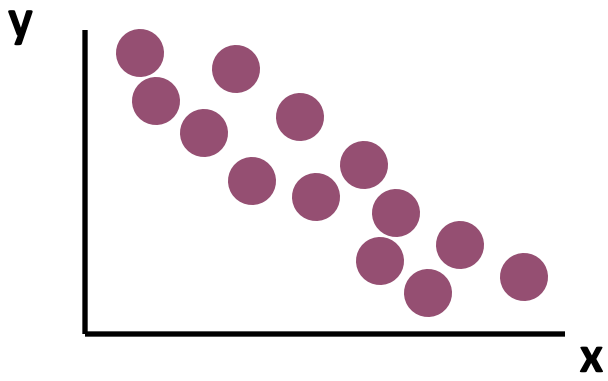
Scatter Plots and Correlation

- A scatter plot (or scatter diagram) is used to show the relationship between two variables
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Only concerned with strength of the relationship
  - No causal effect is implied

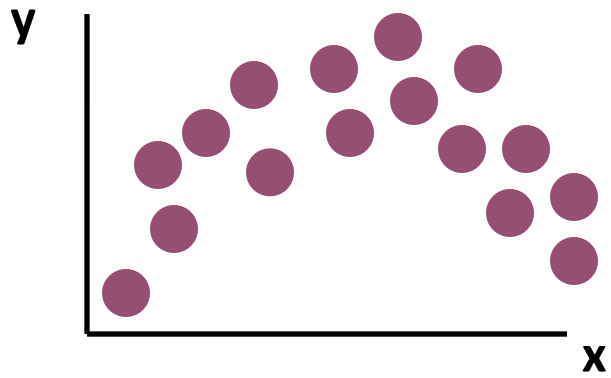
Scatter Plot Examples

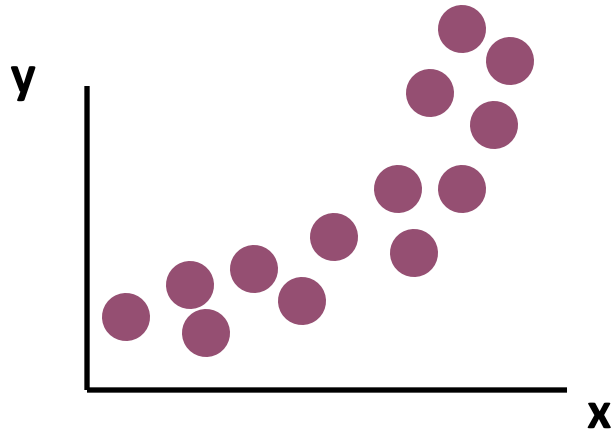
Linear relationships





Curvilinear relationships

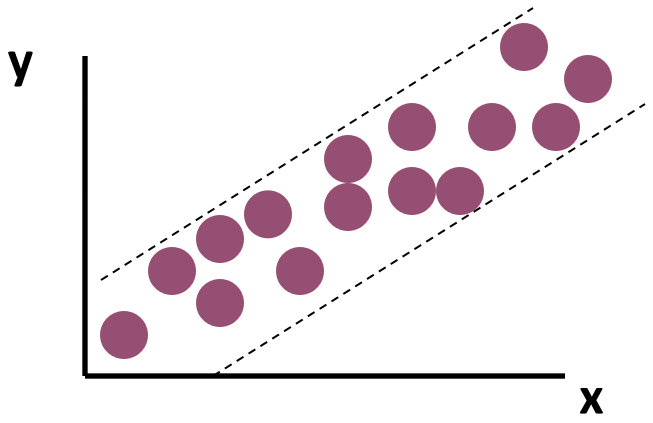


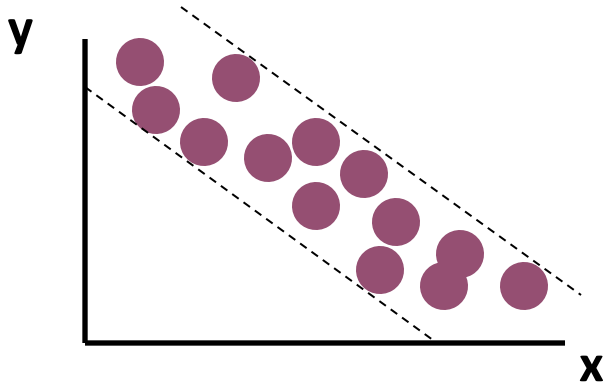


Scatter Plot Examples

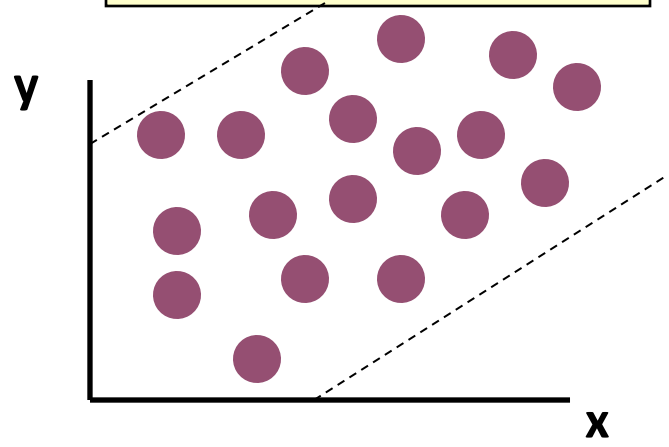
(continued)

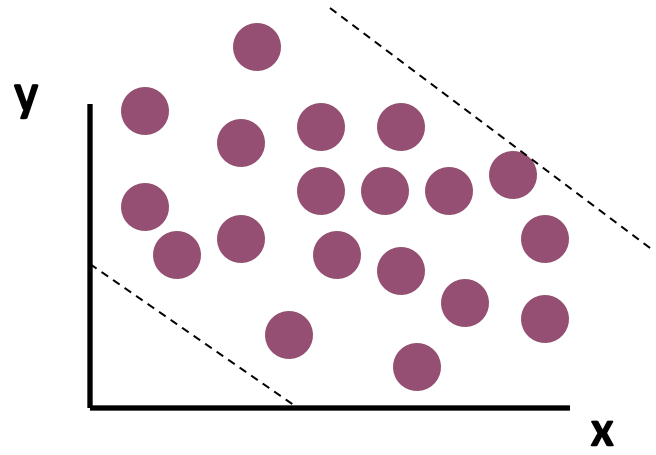
Strong relationships





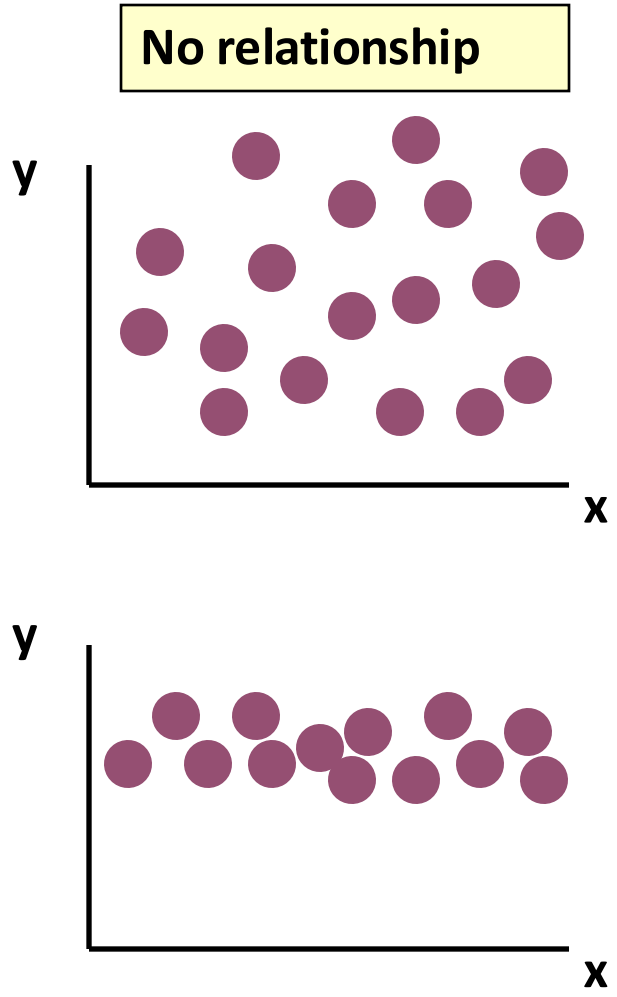
Weak relationships





Scatter Plot Examples

(continued)



Correlation Coefficient

(continued)

- The population correlation coefficient ρ (rho) measures the strength of the association between the variables
  - Unit free
  - Range between -1 and 1
  - The closer to -1, the stronger the negative linear relationship
  - The closer to 1, the stronger the positive linear relationship
  - The closer to 0, the weaker the linear relationship

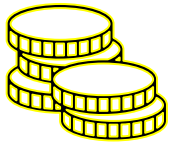
Correlation coefficient -example

We want to study the relationship among the web advertising expenditure and turnover (in tousand of dollars) for 8 retailers.

Is there a relationship between the two variables?

turnover	Web advert expenditure
y	x
35	8
49	9
27	7
33	6
60	13
21	7
45	11
51	12

Example.. Step by step

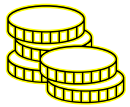


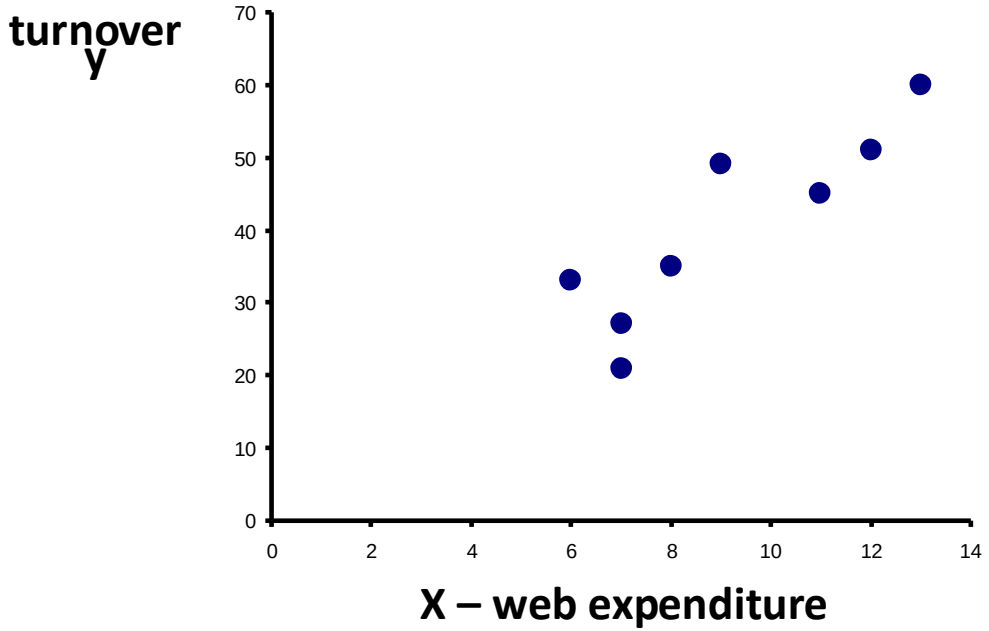
turnover	Web-adv expenditure					
y	x	$(y-\bar{y})$	$(x-\bar{x})$	$(y-\bar{y})*(x-\bar{x})$	$(y-\bar{y})^2$	$(x-\bar{x})^2$
35	8	-5.125	-1.125	5.766	26.266	1.266
49	9	8.875	-0.125	-1.109	78.766	0.016
27	7	-13.125	-2.125	27.891	172.266	4.516
33	6	-7.125	-3.125	22.266	50.766	9.766
60	13	19.875	3.875	77.016	395.016	15.016
21	7	-19.125	-2.125	40.641	365.766	4.516
45	11	4.875	1.875	9.141	23.766	3.516
51	12	10.875	2.875	31.266	118.266	8.266
$\bar{y}$ =40.125	$\bar{x}$ = 9.125			212.875	1230.875	46.875

co-variance(σ_{xy})=$^{212.875}/_{8} = 26.61$

Co-deviance

Calculation Example





$$\text{co-variance}(\sigma_{xy}) = \frac{212.875}{8} = 26.61$$

$$\sigma_y^2 = \frac{1230.875}{8} = 53.85 \qquad \sigma_y = 12.404$$

$$\sigma_x^2 = \frac{46.875}{8} = 5.85 \qquad \sigma_x = 2.42$$

$$\boldsymbol{\rho_{xy}} = \frac{26.61}{(12.404 * 2.42)} = 0.886$$

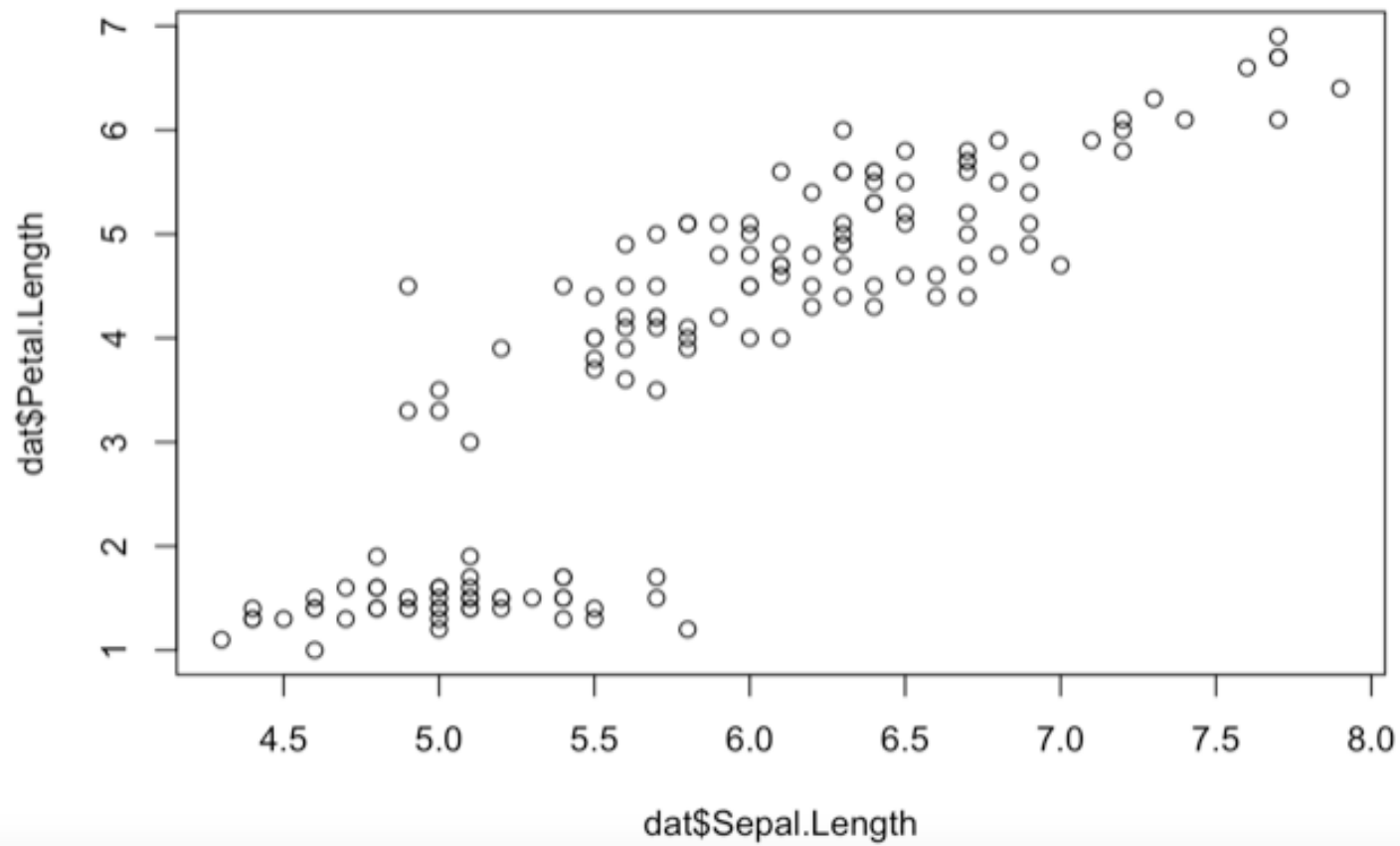
r = 0.886 → relatively strong positive linear association between x and y

Let's try with R !

Scatterplot

Scatterplots allow to check whether there is a potential link between two quantitative variables. For this reason, scatterplots are often used to visualize a potential correlation between two variables. For instance, when drawing a scatterplot of the length of the sepal and the length of the petal:

```
plot(dat$Sepal.Length, dat$Petal.Length)
```



Correlation

Another descriptive statistics is the correlation coefficient.

The correlation measures the *linear* relationship between two variables, and it can be computed with the `cor()` function:

```
cor(dat$Sepal.Length, dat$Sepal.Width)
```

```
## [1] -0.1175698
```