ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

Συσχέτιση μεταξύ δυο ποσοτικών μεταβλητών – Γραμμική εξάρτηση

Κωνσταντίνος Ι. Μπουγιούκας, PhD

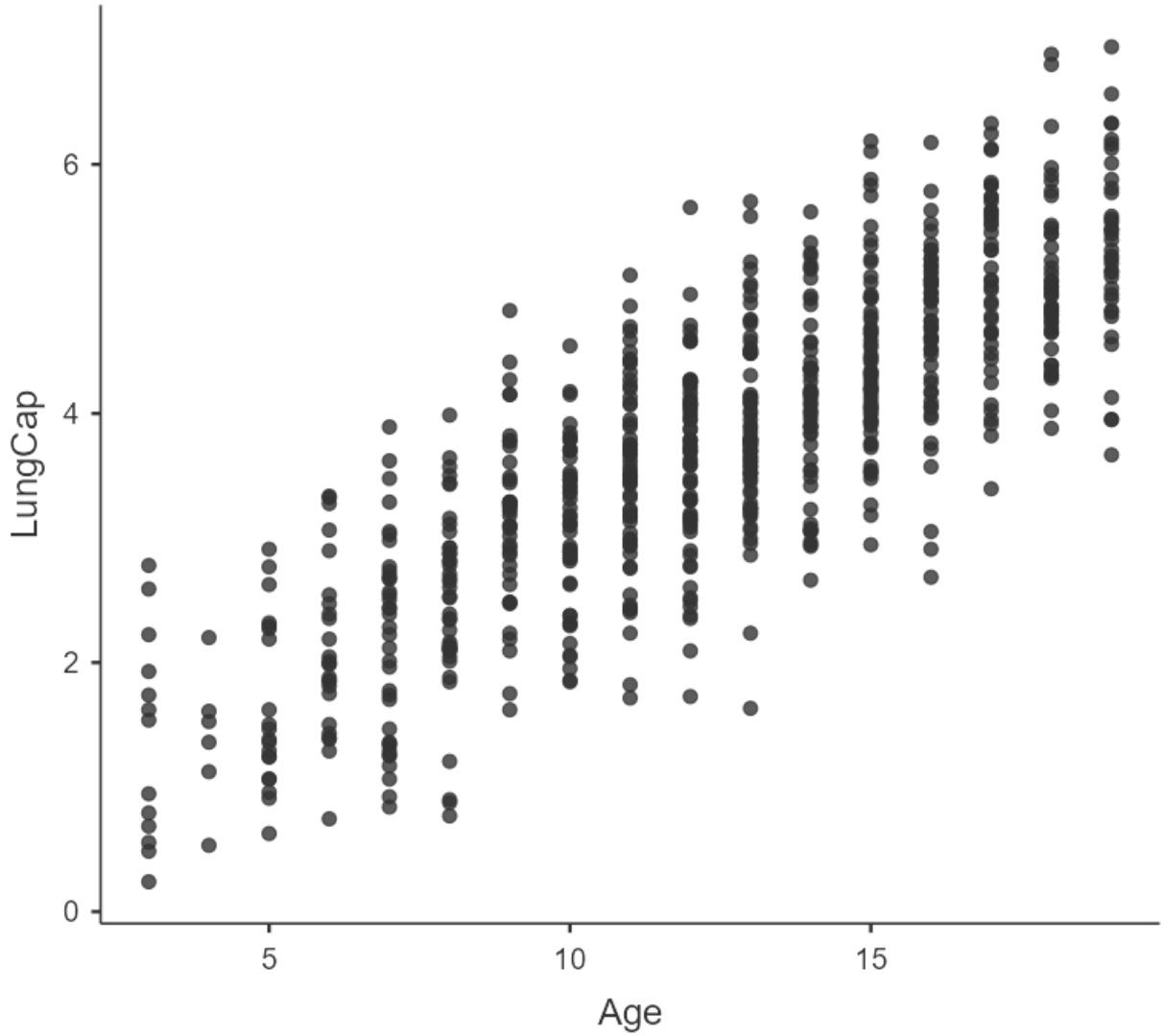
Συσχέτιση δυο ποσοτικών μεταβλητών

Ερευνητικό ερώτημα

Έστω ότι μια έρευνα προσπαθεί να απαντήσει αν συσχετίζεται η **ηλικία** με την **χωρητικότητα των πνευμόνων**.

Συσχέτιση μεταξύ των μεταβλητών **δεν** συνεπάγεται αιτιολογική σχέση μεταξύ των μεταβλητών

Ερευνητικό ερώτημα- Scatterplot



Συσχέτιση ποσοτικών μεταβλητών

⇨Συντελεστής συσχέτισης **Pearson *r***

(εφαρμόζεται όταν τα δεδομένα στις μεταβλητές ακολουθούν την κανονική κατανομή και υπάρχει γραμμική σχέση)

⇨ Συντελεστής συσχέτισης **Spearman r_s**

(εφαρμόζεται συνήθως όταν τα δεδομένα δεν ακολουθούν την κανονική κατανομή και όταν υπάρχουν ακραίες τιμές)

Συντελεστές συσχέτισης

Ποσοτικοποίηση της **κατεύθυνσης** και της **ισχύος** της σχέσης των δυο ποσοτικών μεταβλητών

Έχουν τιμές: -1 ≤ r ≤ +1 (*χωρίς μονάδες*)

- **-1** τέλεια αρνητική συσχέτιση
- **+1** τέλεια θετική συσχέτιση
- **= 0** δεν υπάρχει συσχέτιση

Έλεγχος Υποθέσεων

H_0: ρ = 0

H_1: ρ≠ 0

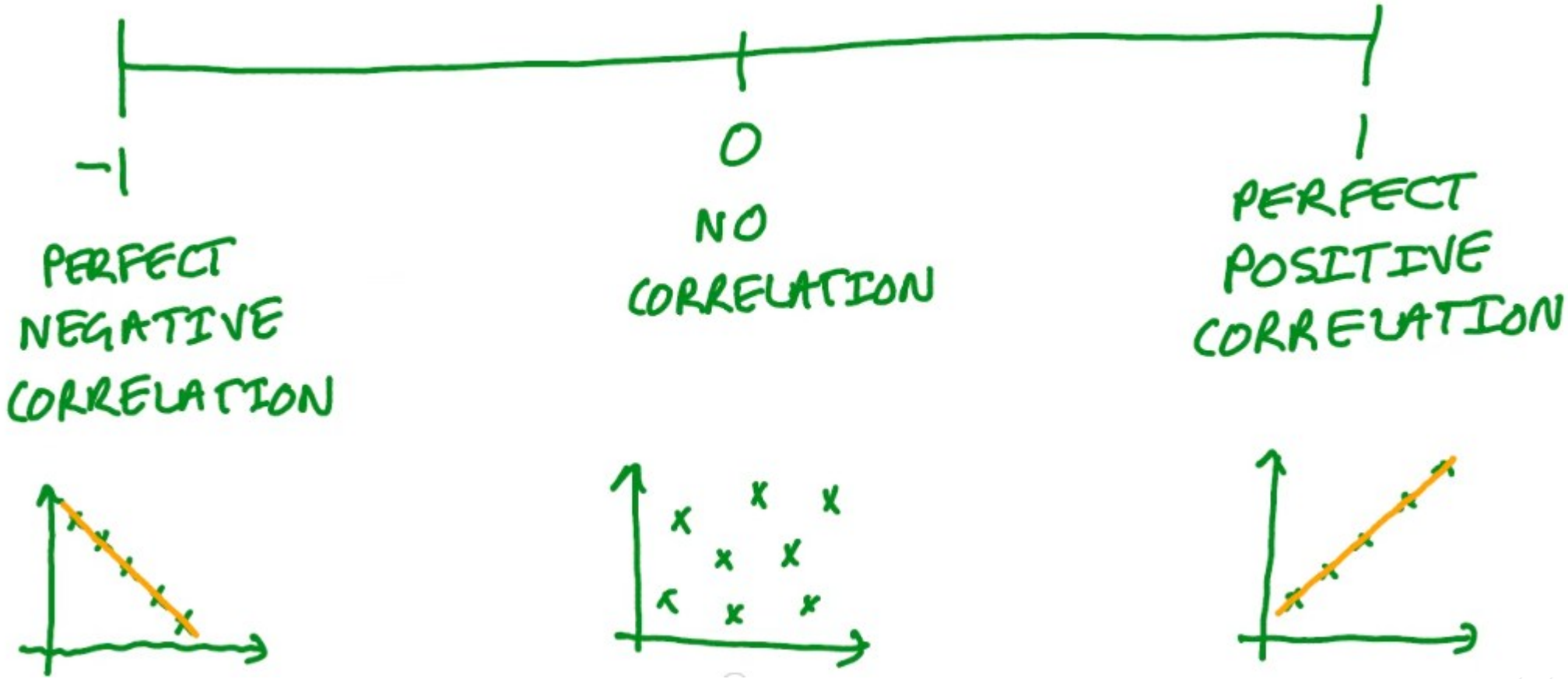
Συντελεστής συσχέτισης του Pearson r (Pearson's correlation coefficient)

- Δείχνει την κατεύθυνση και την ισχύ μιας **γραμμικής** συσχέτισης
- Οι δυο ποσοτικές μεταβλητές (X, Y) πρέπει να ακολουθούν την **κανονική κατανομή**
- Οι **ακραίες τιμές** μπορεί να επηρεάσουν σημαντικά τον συντελεστή του Pearson
- Περιορίζεται η χρήση του όταν υπάρχουν **υπο-ομάδες**

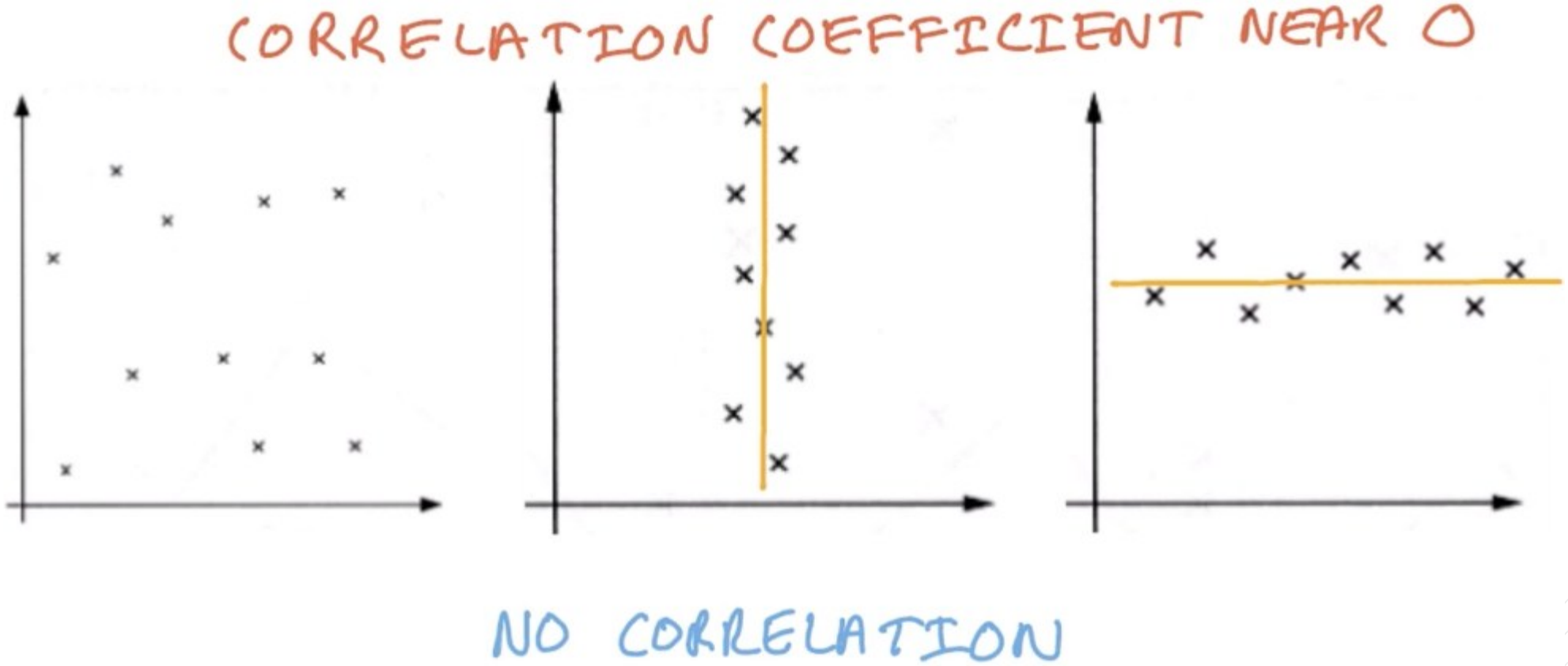
Βαθμός (ένταση) της συσχέτισης δυο ποσοτικών μεταβλητών

$-0.5 < r \leq -0.3$ ή $0.3 \leq r < 0.5$	**ασθενής** συσχέτιση
$-0.7 < r \leq -0.5$ ή $0.5 \leq r < 0.7$	**μέτρια** συσχέτιση
$-0.8 < r \leq -0.7$ ή $0.7 \leq r < 0.8$	**ισχυρή** συσχέτιση
$-1.0 \leq r \leq -0.8$ ή $0.8 \leq r \leq 1.0$	**πολύ ισχυρή** συσχέτιση

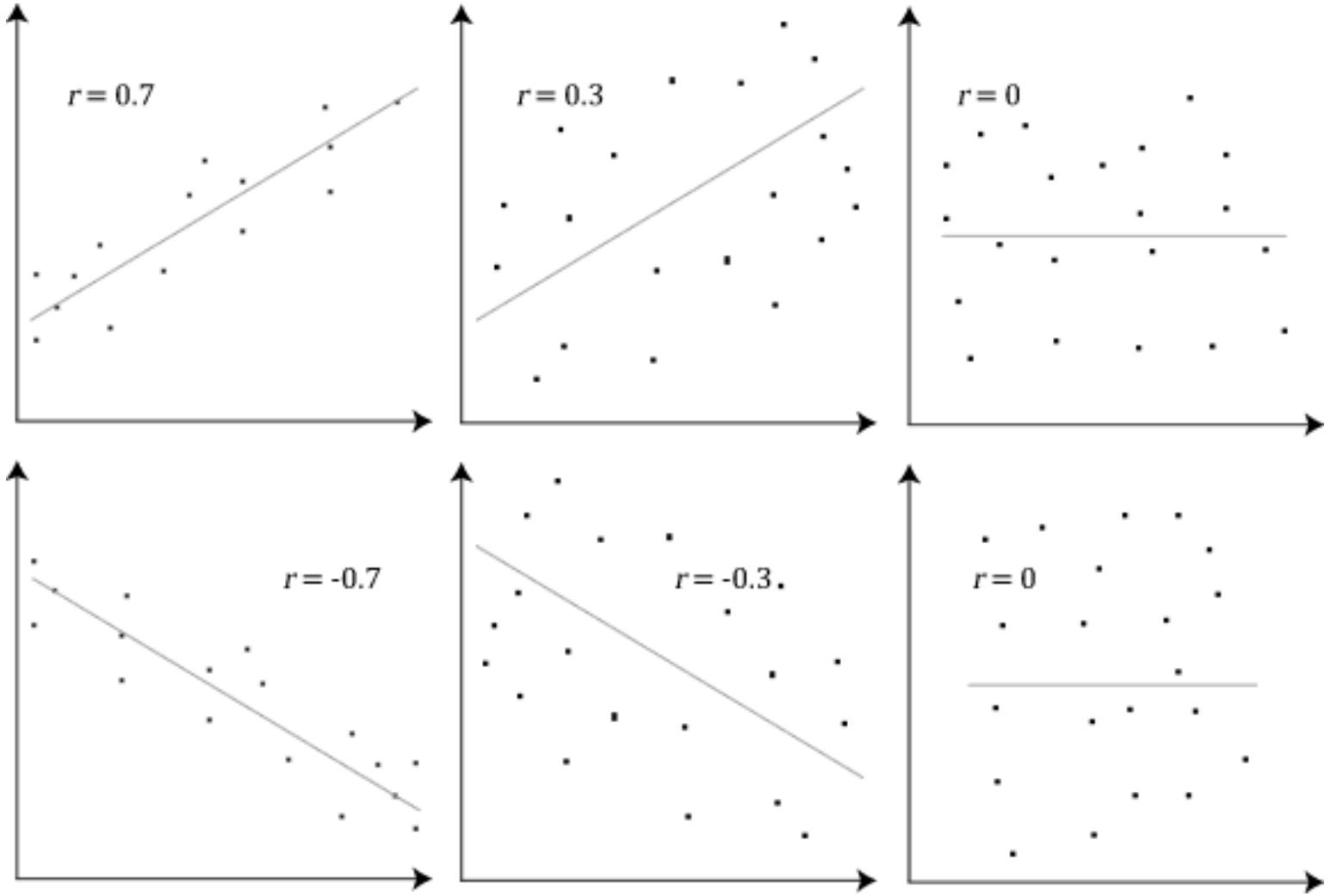
Συντελεστής συσχέτισης του Pearson r (Pearson's correlation coefficient)



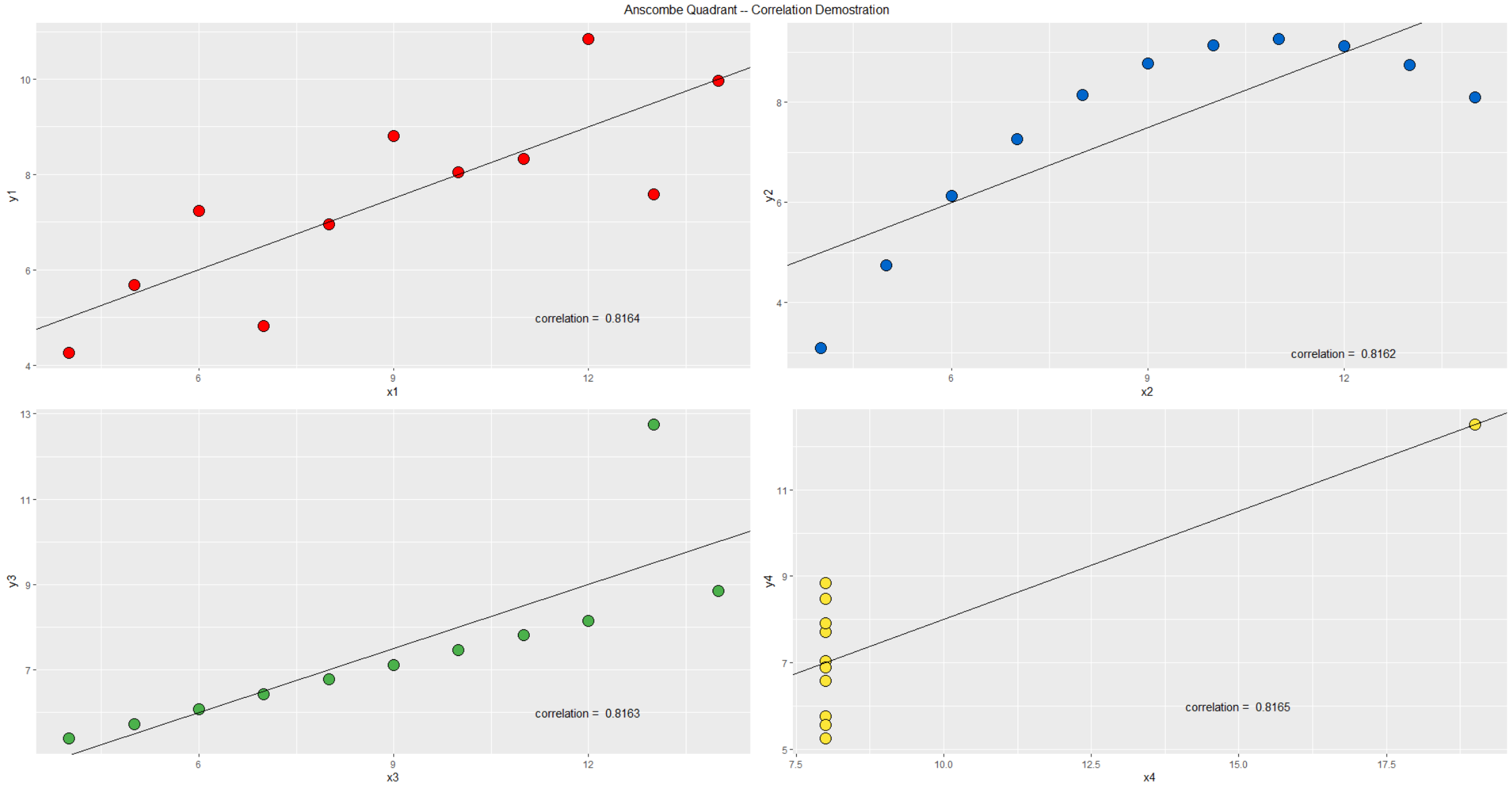
Συντελεστής συσχέτισης του Pearson r (Pearson's correlation coefficient)



Συντελεστής συσχέτισης του Pearson r (Pearson's correlation coefficient)

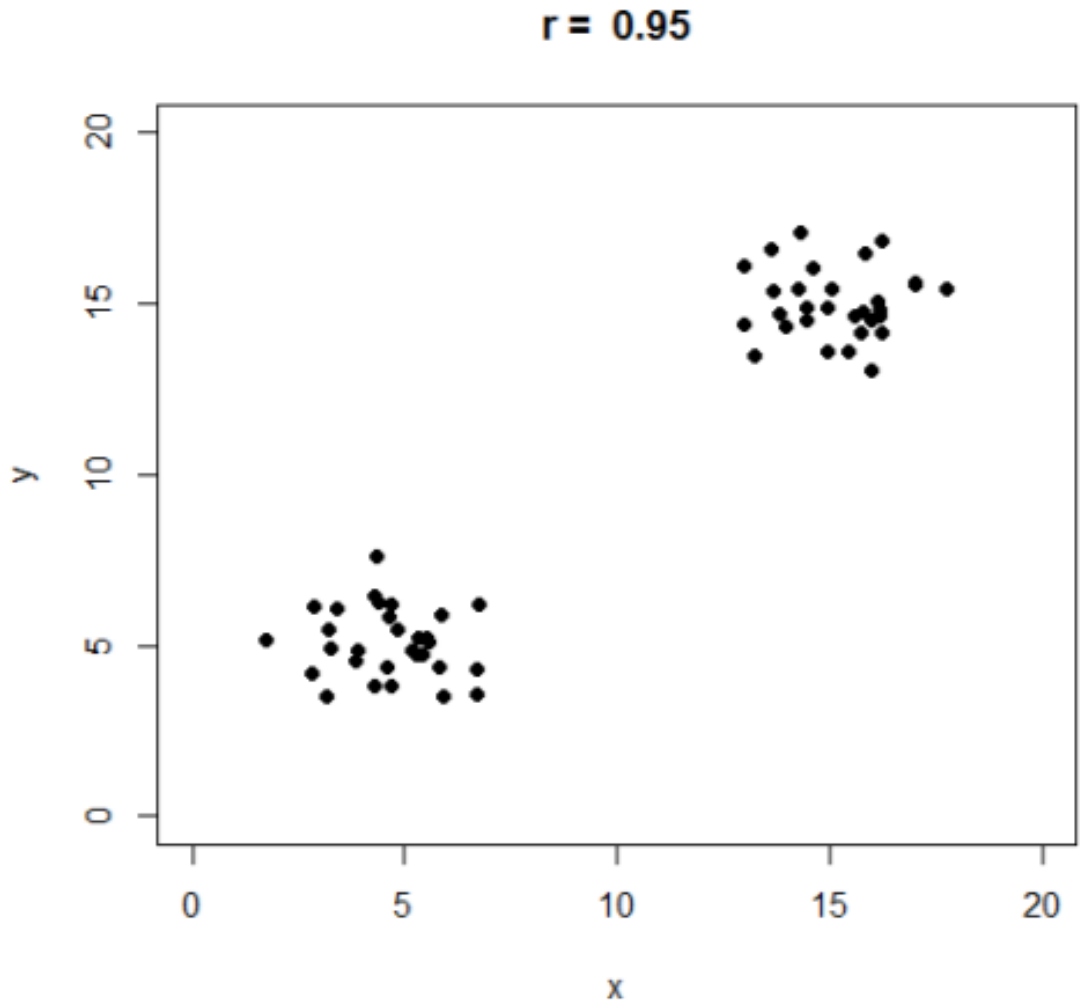


Ο συντελεστής συσχέτισης του Pearson r δεν είναι πάντα χρήσιμος



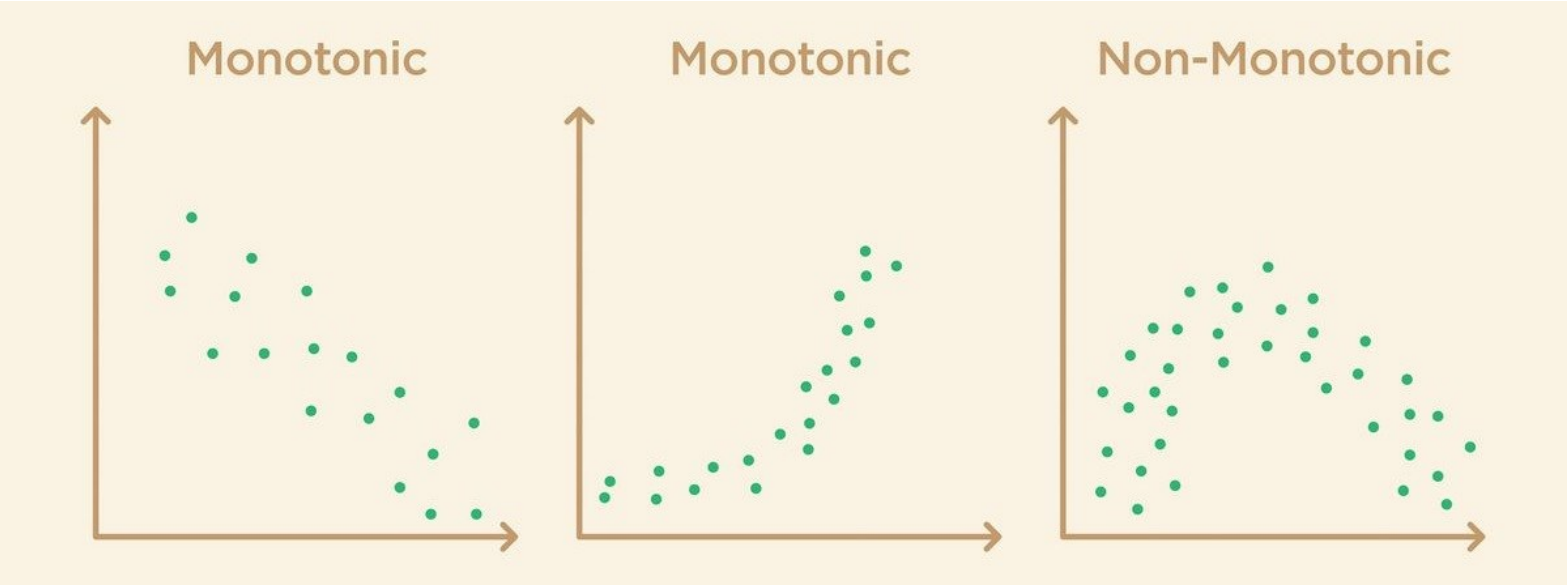
Ο συντελεστής συσχέτισης του Pearson r δεν είναι πάντα χρήσιμος

Περιορίζεται η χρήση του συντελεστή Pearson r όταν υπάρχουν **διακριτές υπο-ομάδες.**



Συντελεστής συσχέτισης του Spearman r_s

- Η **μη παραμετρική** εκδοχή του συντελεστή Pearson. Βασίζεται στις τάξεις (ranks) των μετρήσεων.
- Μετράει την κατεύθυνση και την ισχύ μιας **μονοτονικής σχέσης**
- Είναι **πιο ανθεκτικός** στην ύπαρξη ακραίων τιμών
- Χρησιμοποιείται και σε **διατάξιμες μεταβλητές**



Ερώτηση 1

Ο συντελεστής συσχέτισης του Pearson:

- ✗ (α) Εφαρμόζεται σε μη τυχαία δείγματα
- ✗ (β) Έχει μονάδες μέτρησης
- ✗ (γ) Χρησιμοποιείται όταν υπάρχουν ακραίες τιμές
- ✓ (δ) Προϋποθέτει η σχέση μεταξύ των ποσοτικών μεταβλητών να είναι γραμμική

Ερώτηση 2

Για τον συντελεστή συσχέτισης του Spearman ισχύει:

- ✗ (α) -0.1 ≤ r ≤ +0.1
- ✓ (β) -1 ≤ r ≤ +1
- ✗ (γ) λαμβάνει μόνο θετικές τιμές
- ✗ (δ) έχει μονάδες μέτρησης

Ερώτηση 3

Ο συντελεστής συσχέτισης του Spearman :

- ✓ (α) Εφαρμόζεται καλά όταν οι μεταβλητές έχουν μια μονοτονική σχέση
- ✗ (β) Δεν μπορεί να υπολογιστεί όταν η σχέση είναι γραμμική
- ✗ (γ) Προϋποθέτει οι μεταβλητές να είναι κανονικές
- ✗ (δ) Εφαρμόζεται καλά όταν τα δεδομένα έχουν την μορφή U

Ερώτηση 4

Τιμή του συντελεστής συσχέτισης του Pearson r ίση με:

✗ (α) -1 δείχνει τέλεια θετική συσχέτιση

✗ (β) 0.5 δείχνει ότι δεν υπάρχει συσχέτιση

✗ (γ) 1 δείχνει ότι δεν υπάρχει συσχέτιση

✓ (δ) -1 δείχνει τέλεια αρνητική συσχέτιση

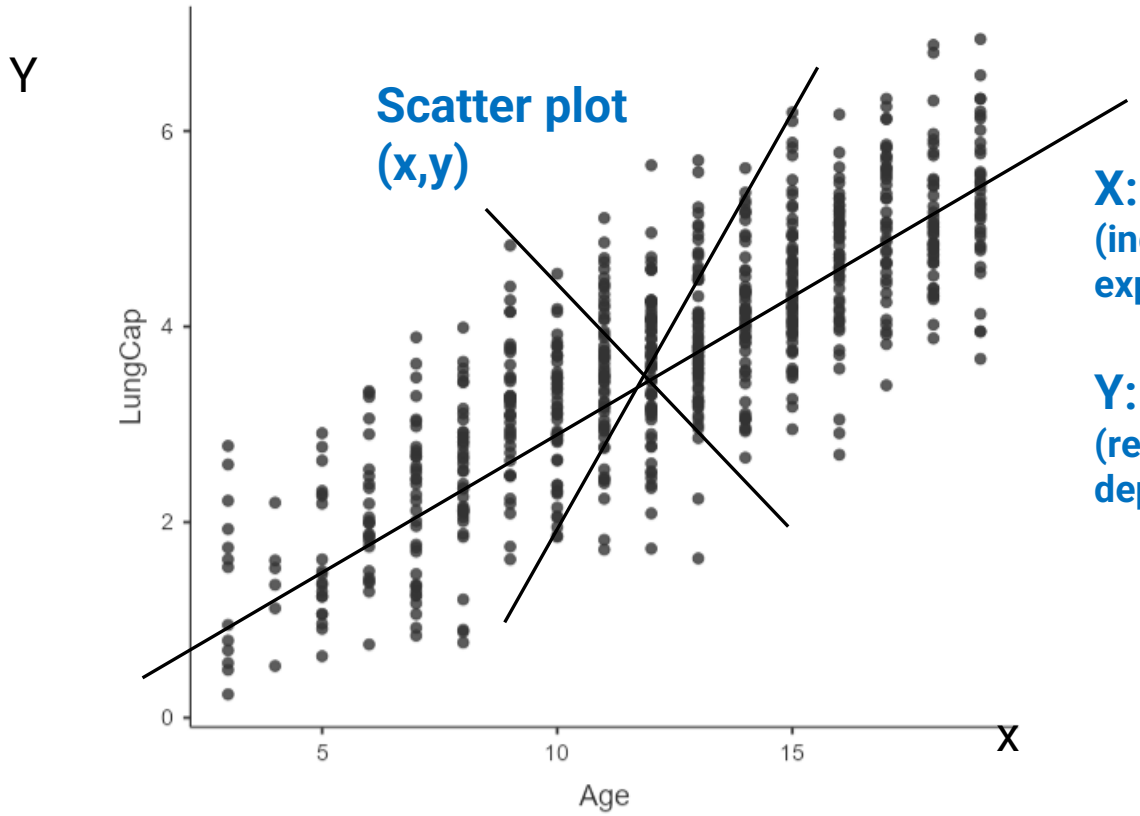
Γραμμική Εξάρτηση

Απλή γραμμική εξάρτηση

Εξαρτημένη μεταβλητή → $y_i = a + \beta * x_i + \varepsilon_i$ $i=1,2,\ldots,n$

Ανεξάρτητη μεταβλητή

Random error



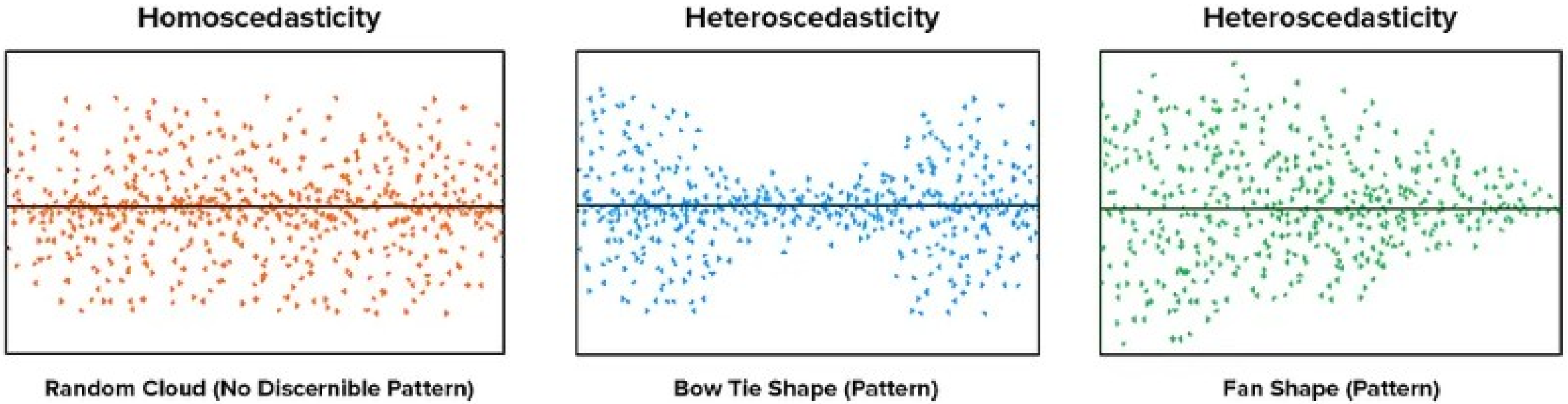
X: Age
(independent or explanatory variable)

Y: LungCap
(response or dependent variable)

Απλή γραμμική εξάρτηση – Βασικές Προϋποθέσεις (Basic Assumptions)

- Γραμμική εξάρτηση των μεταβλητών Χ και Υ
- Τα σφάλματα να είναι ανεξάρτητα (δηλ. να μην υπάρχει αυτοσυσχέτιση)
- Κανονική κατανομή των σφαλμάτων
- Ομοσκεδαστικότητα (homoscedasticity) των σφαλμάτων
- Να μην υπάρχουν επηρεάζουσες παρατηρήσεις (influential points)

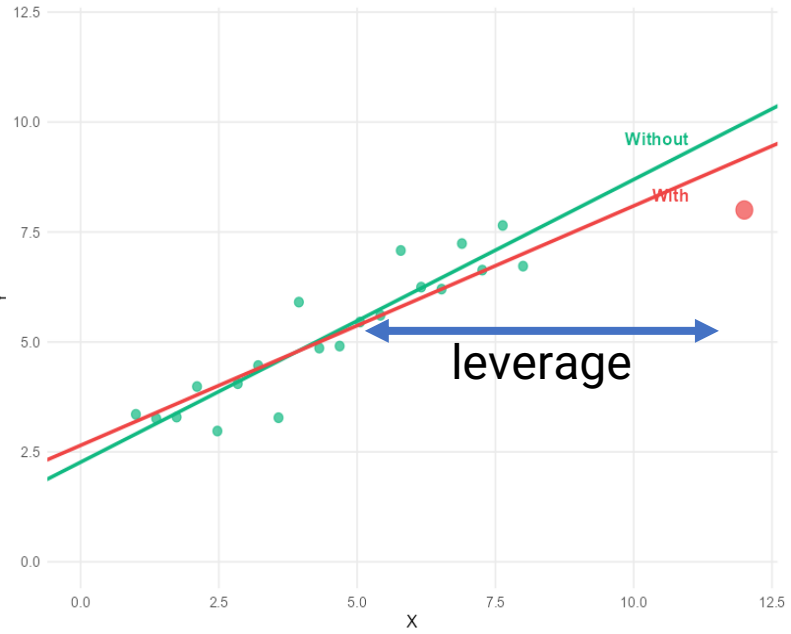
Ομοσκεδαστικότητα (homoscedasticity) των σφαλμάτων



Homoscedasticity vs Heteroscedasticity

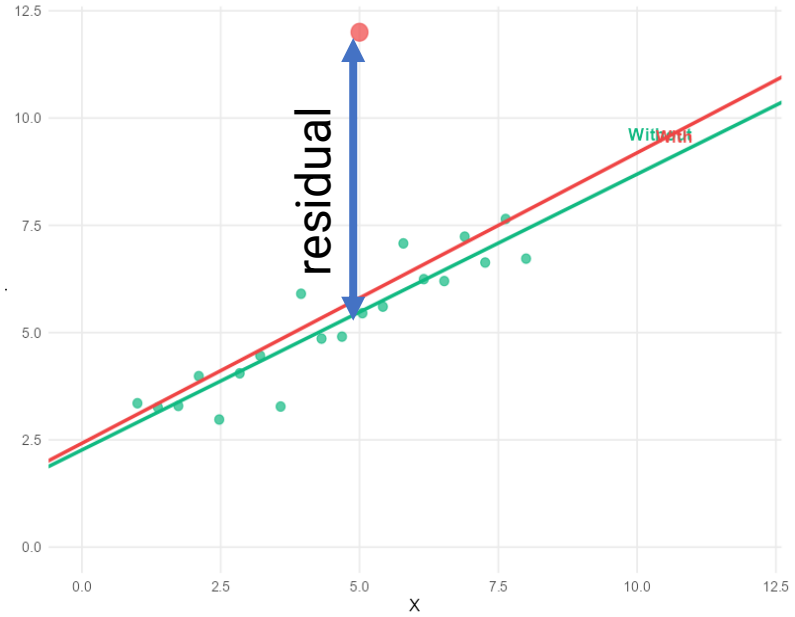
Επηρεάζουσες παρατηρήσεις (influential points)

High leverage



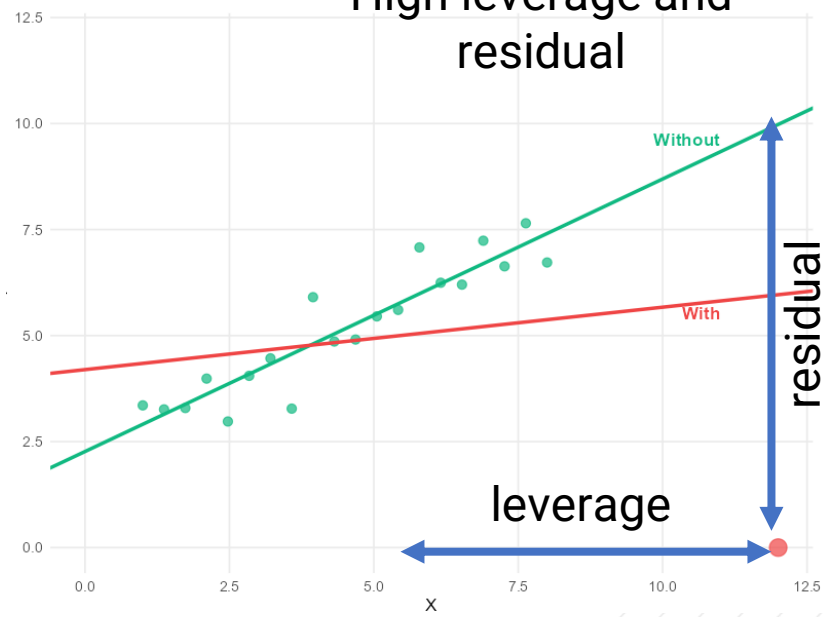
The leverage h (x-value) of the red point is much higher than the mean of X of the observations. However, the red point does not influence the model very much (not influential).

High residual



The residual (y-value) of the red point is much higher than the mean of Yof the observations. However, the red point does not influence the model very much (not influential).

Influential Point
High leverage and residual

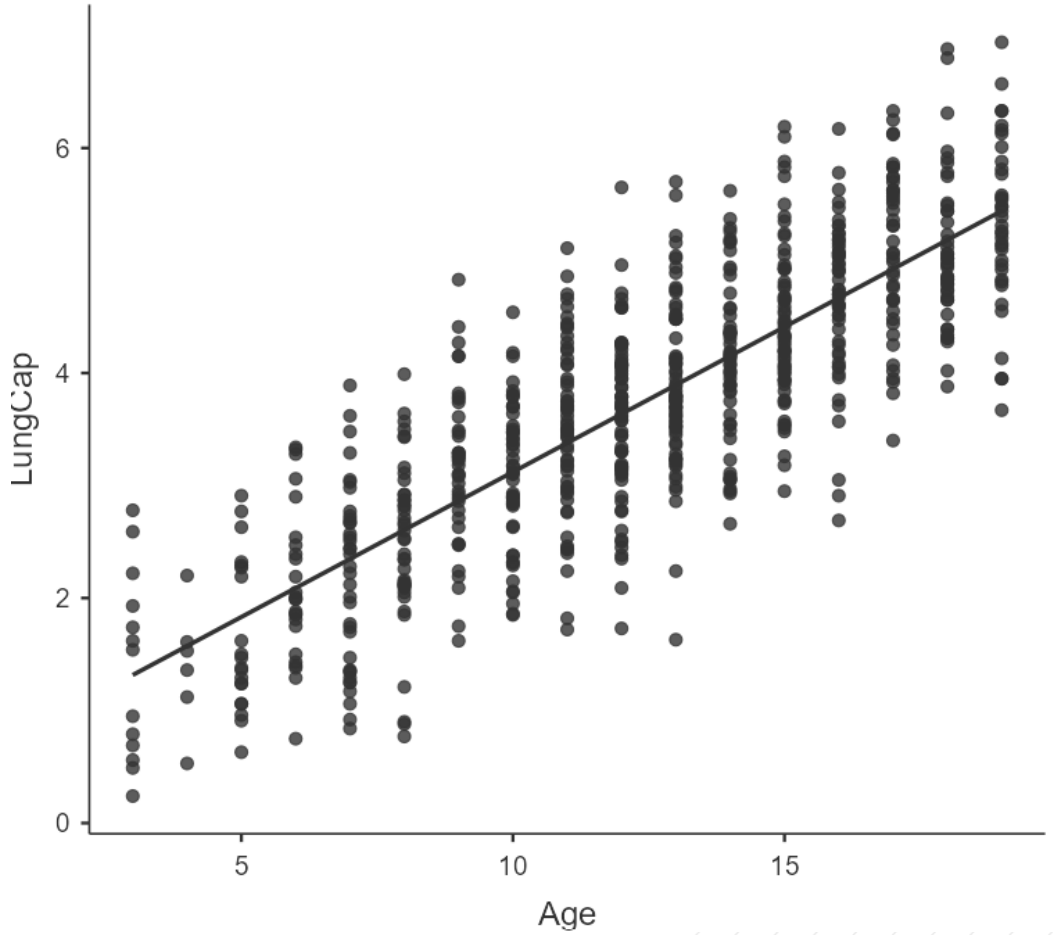


The red point has both high leverage and a large residual, and therefore it has a strong influence on the model (influential point).

Απλή γραμμική εξάρτηση – Έλεγχος Υποθέσεων

$$\hat{y} = \alpha + \beta * x$$

- **Ho: β=0** (no association)
- **H1: β≠0** (there is association)



Απλή γραμμική εξάρτηση – Ποσοτική ανεξάρτητη μεταβλητή

$$\hat{LungCap} = 0.54 + 0.26 * Age$$

Αν αυξηθεί η ηλικία ενός ατόμου κατά ένα χρόνο π.χ. από τα 14 χρόνια στα 15 χρόνια, αυξάνεται **κατά μέσο** όρο η χωρητικότητα των πνευμόνων του κατά **0.26 L** (95%CI: 0.24 έως 0.27, p <0.001).

Απλή γραμμική εξάρτηση – Αξιολόγηση του μοντέλου

Συντελεστής προσδιορισμού:

$$R^2 = \frac{explained\ variation}{total\ variation}$$

(R^2 : 0 to 1)

Μέτρο '**καλής προσαρμογής**' του γραμμικού μοντέλου στα δεδομένα.

Κοντά στο 1 ⇨ μεγάλο ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής ερμηνεύεται από την ανεξάρτητη μεταβλητή του μοντέλου.

Π.χ. για την περίπτωση της ποσοτικής μεταβλητής της ηλικίας (Age) βρίσκουμε $R^2 = 0.67$ και σημαίνει ότι 67% διακύμανσης της εξαρτημένης μεταβλητής (LungCap) ερμηνεύεται από την ανεξάρτητη μεταβλητή του μοντέλου (Age).