Απλή γραμμική εξάρτηση (Simple linear regression)

Γιώτα Τουλούμη
Καθηγήτρια Βιοστατιστικής και Επιδημιολογίας
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής
Ιατρική Σχολή, ΕΚΠΑ
gtouloum@med.uoa.gr

Βάνα Σύψα
Καθηγήτρια Επιδημιολογίας και Ιατρικής Στατιστικής
Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής
Ιατρική Σχολή, ΕΚΠΑ
vsipsa@med.uoa.gr

Συσχέτιση και εξάρτηση

Παράδειγμα: Έστω ότι επιθυμούμε να διερευνήσουμε τη σχέση ανάμεσα στο ανάστημα πατέρα και παιδιού. Αυτό μπορεί να μελετηθεί με 2 τρόπους

- **Διερεύνηση του βαθμού συμμεταβολής των 2 μεταβλητών → <u>Συσχέτιση</u>**
  - Υπάρχει γραμμική συσχέτιση; Θετική ή αρνητική;

- **Διερεύνηση της εξάρτησης του αναστήματος παιδιού (εξαρτημένη μεταβλητή) από ανάστημα πατέρα (ανεξάρτητη μεταβλητή) → <u>Εξάρτηση</u>**
  - Εκτός από το αν υπάρχει θετική ή αρνητική γραμμική συσχέτιση (ή όχι συσχέτιση) επιτρέπει να απαντήσουμε π.χ. αν ένας πατέρας είναι 10cm ψηλότερος από έναν άλλον, πόσο θα διαφέρει το ανάστημα των παιδιών τους κατά μέσο όρο

Συσχέτιση και εξάρτηση

- **Συσχέτιση:**
  Μέτρο του βαθμού (της δύναμης) της γραμμικής σχέσης
  Διερεύνηση συσχέτισης μεταξύ δύο μεταβλητών – δεν υπάρχει ενδιαφέρον (δυνατότητα) διερεύνησης αιτιολογικής σχέσης

- **Εξάρτηση ή Παλινδρόμηση:**
  Μέθοδος για την διερεύνηση των μεταβολών των τιμών της μιας **ποσοτικής** μεταβλητής (εξαρτημένη) συναρτήσει των μεταβολών των τιμών της άλλης (ανεξάρτητη)
  Σχέση αίτιου-αιτιατού:
  αίτιο→ανεξάρτητη μεταβλητή
  αιτιατό→ εξαρτημένη μεταβλητή

Συσχέτιση και εξάρτηση

Η διάκριση μεταξύ συσχέτισης και εξάρτησης (παλινδρόμησης) είναι περισσότερο εννοιολογική και λιγότερο στατιστική

•Εάν μας ενδιαφέρει **η ένταση** της σχέσης των δύο μεταβλητών, αρκεί η συσχέτιση (correlation coefficient)

•Εάν μας ενδιαφέρει η μελέτη **της εξάρτησης** της μιας μεταβλητής από την άλλη (εξαρτημένη μεταβλητή-ανεξάρτητη μεταβλητή) τότε επιλέγουμε την παλινδρόμηση (εξάρτηση)

Εξαρτημένη και ανεξάρτητη μεταβλητή

1. **Εννοιολογικά, ως ανεξάρτητη μεταβλητή θα πρέπει να επιλέγεται αυτή που αποτελεί αιτιολογικό/προγνωστικό παράγοντα για τα επίπεδα της άλλης μεταβλητής**

   - Π.χ. ηλικία και επίπεδα συστολικής αρτηριακής πίεσης

   - Η ηλικία επηρεάζει τα επίπεδα συστολικής πίεσης

     Συστολική πίεση → **Εξαρτημένη (Υ)**

     Ηλικία → **Ανεξάρτητη (Χ)**

Εξαρτημένη και ανεξάρτητη μεταβλητή

2. **Τρόπος επιλογής των μεταβλητών:**

   Η **εξαρτημένη** μεταβλητή πρέπει να έχει επιλεγεί με τυχαίο τρόπο

   - π.χ. αν η επιλογή των ηλικιών δεν έχει γίνει τυχαία αλλά με τέτοιο τρόπο ώστε να περιλαμβάνεται στο δείγμα ικανός αριθμός ατόμων από όλες τις ηλικιακές ομάδες → η ηλικία δεν μπορεί να είναι εξαρτημένη μεταβλητή – θα αποτελέσει την ανεξάρτητη μεταβλητή

Προϋποθέσεις

- **Συσχέτιση:**

Τα δύο ποσοτικά μεγέθη να κατανέμονται κανονικά και να έχουν επιλεγεί τυχαία

- **Εξάρτηση:**

Το εξαρτημένο μέγεθος να **κατανέμεται κανονικά** (για κάθε συγκεκριμένη τιμή του ανεξάρτητου) **και να έχει επιλεγεί τυχαία**

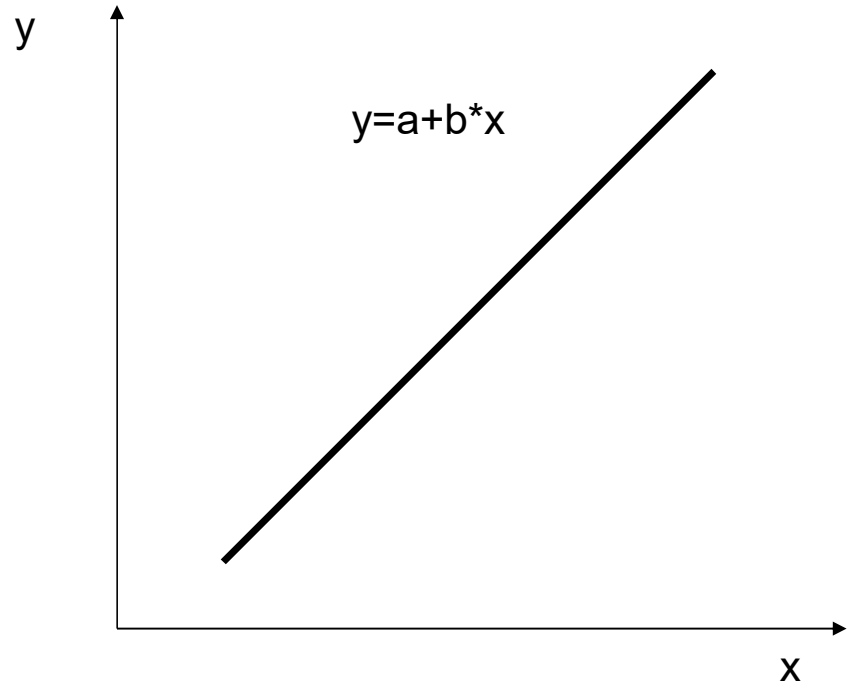
Απλή γραμμική εξάρτηση (simple linear regression)

- Στην απλή εξάρτηση διερευνάται η σχέση μιας εξαρτημένης μεταβλητής με **μία** μόνο ανεξάρτητη μεταβλητή.

- Ποια είναι η λογική της;
  - Προσπαθούμε να εκτιμήσουμε την ευθεία που χαρακτηρίζει «καλύτερα» τη σχέση μεταξύ της εξαρτημένης και ανεξάρτητης μεταβλητής
  - Με βάση αυτή τη γραμμή μπορούμε να βρούμε κάθε τιμή της εξαρτημένης που αντιστοιχεί σε συγκεκριμένη τιμή της ανεξάρτητης.

Εξετάζουμε τη γραμμική σχέση

Εξίσωση ευθείας

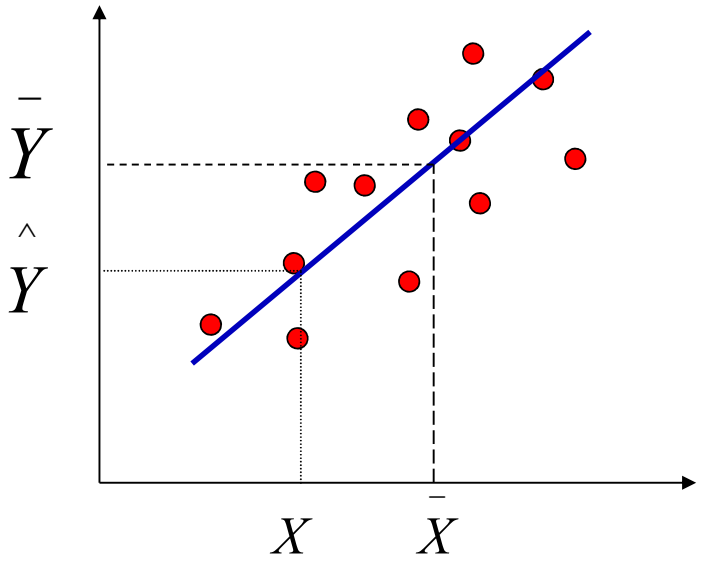


- b = κλίση (slope) της ευθείας
- a=σταθερά (intercept) της ευθείας → το σημείο που τέμνει η ευθεία τον άξονα y στο x=0

Εξίσωση ευθείας

$$\hat{Y} = a + bX$$

$\hat{Y}$: η τιμή της Y που αντιστοιχεί σε ορισμένη τιμή της X με βάση τη γραμμή εξάρτησης



Μπορεί επίσης να γραφτεί ως εξής:

$$\hat{Y} = a + b(X - \bar{X}) + b\bar{X}$$

Προσθαφαιρώ την ίδια ποσότητα

$$\hat{Y} = (a + b\bar{X}) + b(X - \bar{X})$$

$$\hat{Y} = \bar{Y} + b(X - \bar{X})$$

Εξίσωση απλής γραμμικής εξάρτησης

$$\hat{Y} = a + bX$$

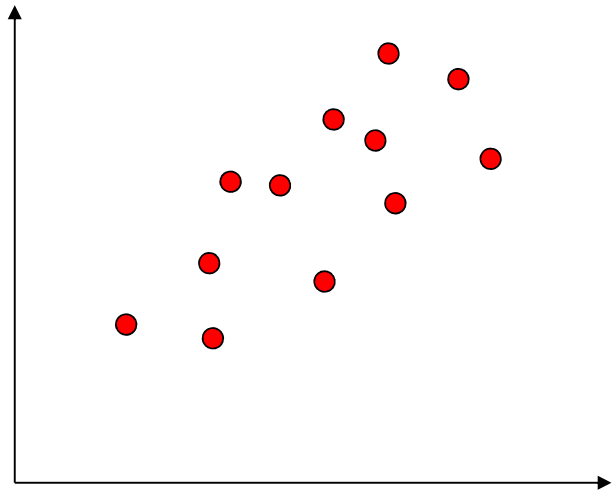
ή $\hat{Y} = \overline{Y} + b(X - \overline{X})$

- $\overline{X}, \overline{Y}$: οι μέσες τιμές των 2 μεταβλητών με βάση τα δεδομένα μας
- $\hat{Y}$: η τιμή της Y που αντιστοιχεί σε ορισμένη τιμή της X με βάση τη γραμμή εξάρτησης
- b: ο συντελεστής κλίσης της γραμμής εξάρτησης (regression coefficient)

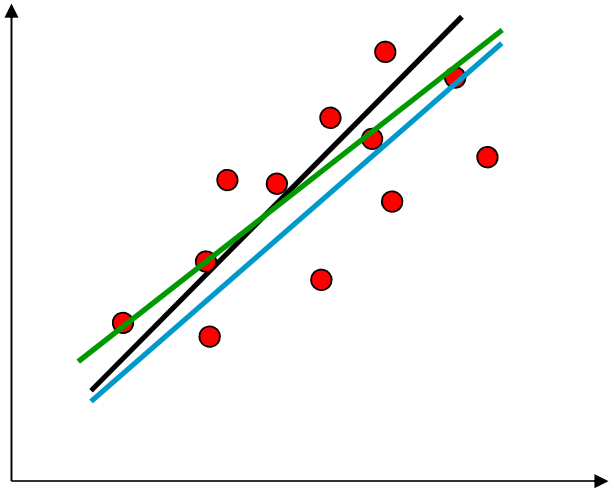
Πώς θα βρω το b;

Μέθοδος εκτίμησης: Ευθεία ελαχίστων τετραγώνων (least squares)

- Αρχικά, τα δεδομένα έχουν την παρακάτω μορφή:

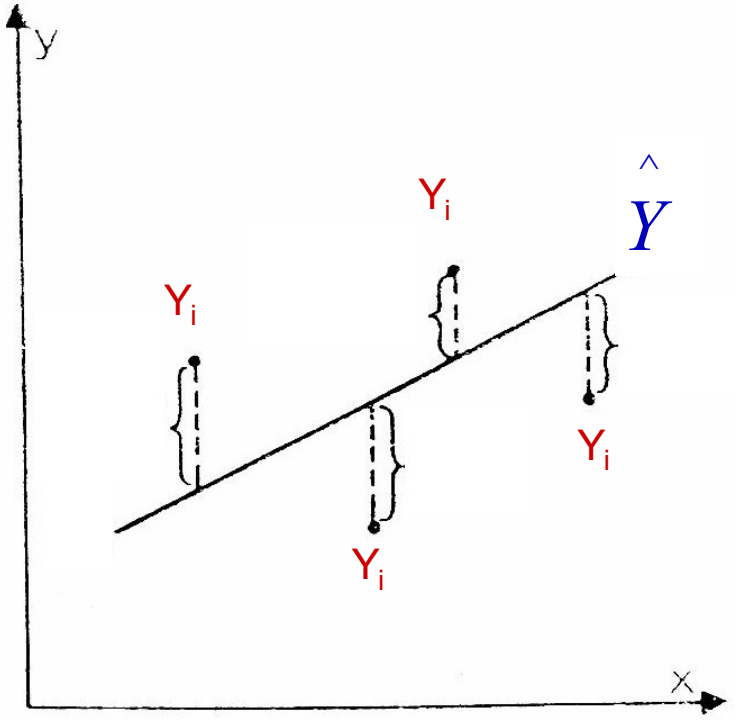


- Ποια ευθεία αναπαριστά καλύτερη τη σχέση αυτή;



Μέθοδος εκτίμησης: Ευθεία ελαχίστων τετραγώνων (least squares)

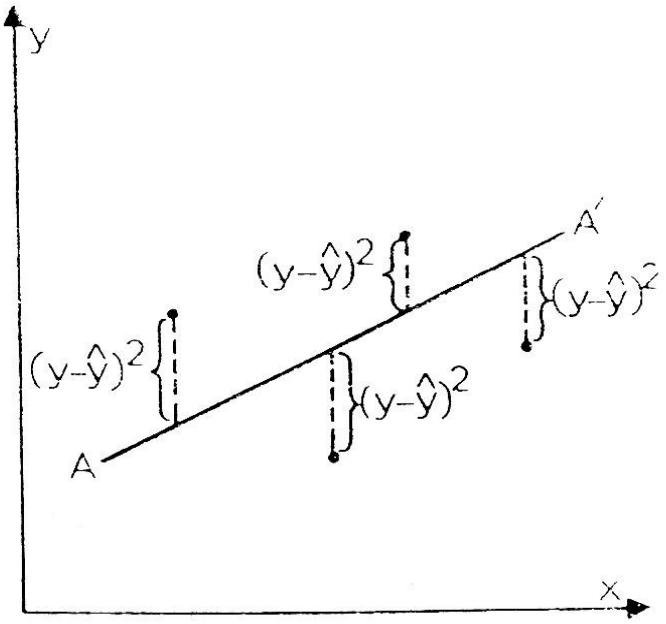
- Η «καλύτερη» ευθεία είναι αυτή από την οποία οι παρατηρούμενες τιμές Y_i απέχουν τη μικρότερη απόσταση (μικρή απόσταση μεταξύ του πραγματικού Y_i και του $\hat{Y}$ που εκτιμά η ευθεία



Μέθοδος εκτίμησης: Ευθεία ελαχίστων τετραγώνων (least squares)

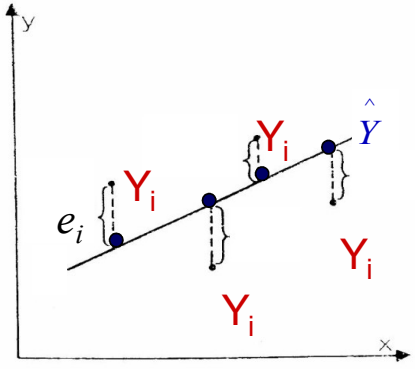
- Η ευθεία εξάρτησης της Y από τη X που εφαρμόζει καλύτερα στα δεδομένα είναι αυτή που ελαχιστοποιεί το άθροισμα των τετραγώνων:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



Y_i και $\hat{Y}_i$

- Για δεδομένο X_i, στην εξίσωση ευθείας αντιστοιχεί η τιμή $\widehat{Y}_i$
- Ορίζουμε ως υπόλοιπο (residual) e_i τη διαφορά μεταξύ της παρατηρούμενης τιμής Y_i και της εκτιμώμενης $\widehat{Y}_i$ από την εξίσωση ευθείας



$$\widehat{Y}_i - Y_i = e_i$$

π.χ. αν Y_i: ύψος παιδιού,

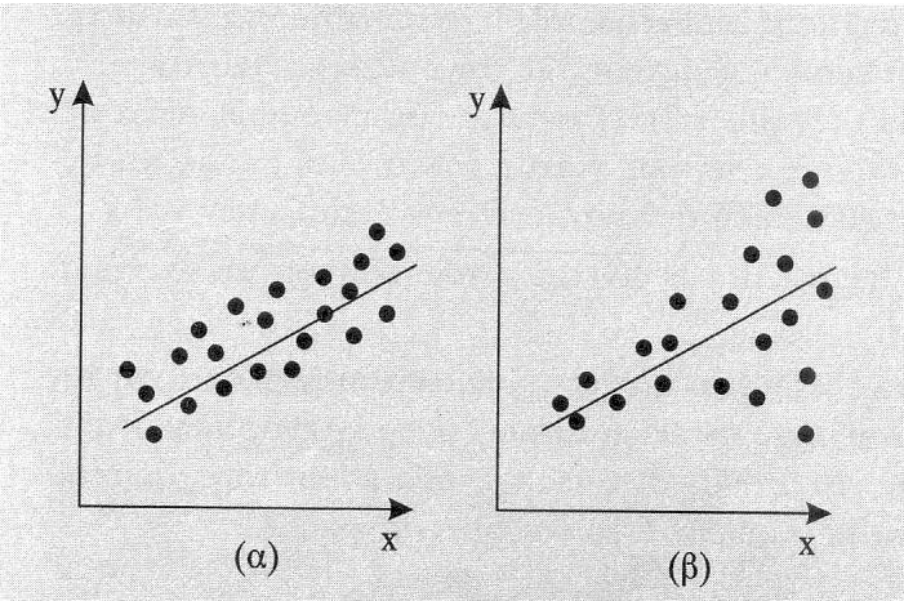
X_i: ύψος πατέρα → εκτίμηση $\widehat{Y}_i$ (αναμενόμενο ύψος παιδιού για δεδομένο ύψος πατέρα)

$$Y_i = \hat{Y}_i + e_i$$

- Y_i: Πραγματικό ύψος παιδιού
- $\hat{Y}_i$: Εκτιμώμενο ύψος παιδιού για δεδομένο ύψος πατέρα
- e_i: Υπόλοιπο

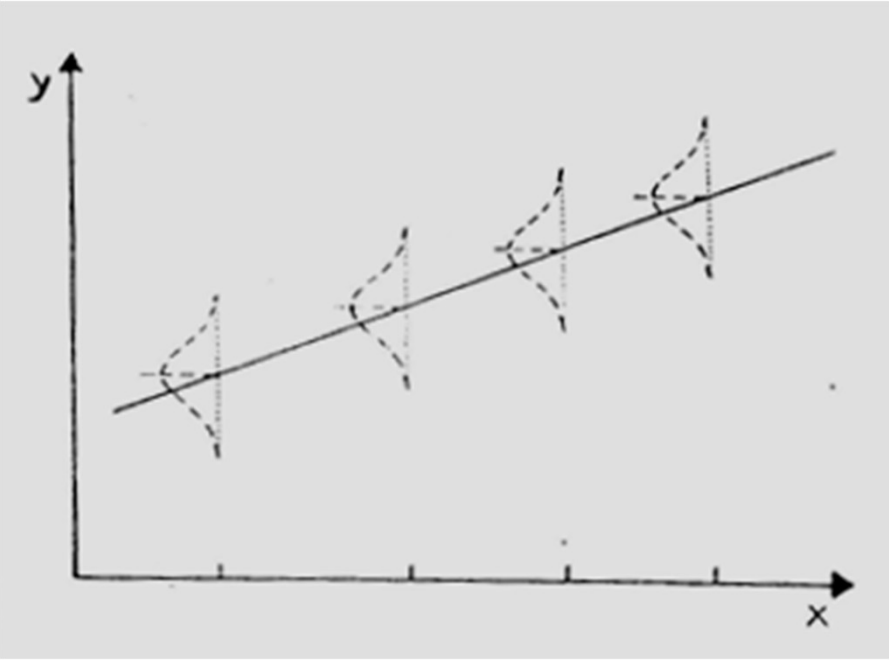
Προϋποθέσεις για την εφαρμογή της γραμμικής εξάρτησης

1. Η ευθεία γραμμή να είναι ικανοποιητική προσέγγιση της σχέσης y και x.
2. Η διασπορά των σημείων πάνω και κάτω από την ευθεία να είναι περίπου ομοιόμορφη σε όλο το μήκος της γραμμής.



Προϋποθέσεις

3. Η κατανομή συχνοτήτων της εξαρτημένης Υ, οι οποίες αντιστοιχούν σε ορισμένη τιμή της Χ, πρέπει να είναι κατά προσέγγιση κανονική



Συντελεστής εξάρτησης b (regression coefficient)

$$\hat{Y} = a + bX$$ ή $$\hat{Y} = \bar{Y} + b(X - \bar{X})$$

Ο συντελεστής εξάρτησης b εκφράζει πόσο μεταβάλλεται κατά μέσο όρο η Υ, όταν η Χ μεταβάλλεται κατά 1 μονάδα.

Π.χ.

έστω δύο άτομα με τιμές Χ (π.χ. ηλικία) ίσες με X_1 και X_1+1. Με βάση τη γραμμή εξάρτησης, πόσο αναμένεται να διαφοροποιείται κατά μέσο όρο η τιμή του Υ τους (π.χ. τριγλυκερίδια);

Συντελεστής εξάρτησης b (regression coefficient)

$$\hat{Y}_1 = \bar{Y} + b(X_1 - \bar{X})$$

$$\hat{Y}_2 = \bar{Y} + b[(X_1 + 1) - \bar{X})$$

$$\longrightarrow \quad \hat{Y}_2 - \hat{Y}_1 = \bar{Y} + bX_1 + b - b\bar{X} - (\bar{Y} + b(X_1 - \bar{X}))$$

$$\hat{Y}_2 - \hat{Y}_1 = \cancel{\bar{Y}} + \cancel{bX_1} + b - \cancel{b\bar{X}} - \cancel{\bar{Y}} - \cancel{bX_1} + \cancel{b\bar{X}}$$

$$\hat{Y}_2 - \hat{Y}_1 = b$$

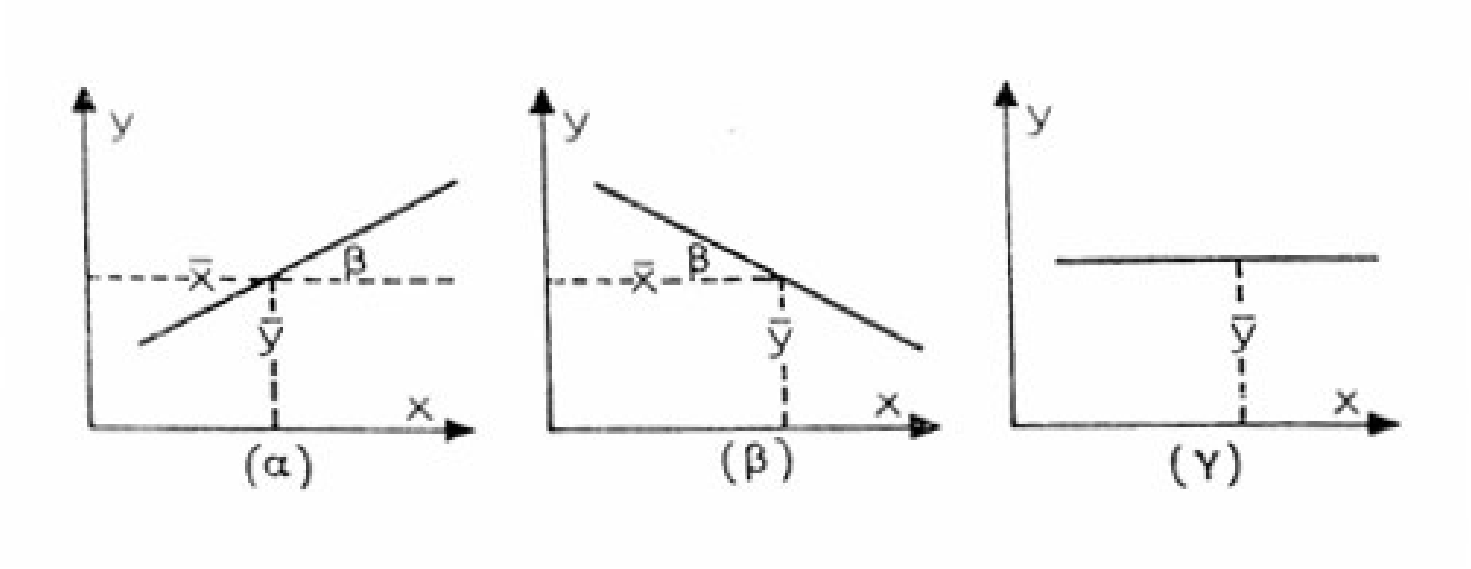
Ερμηνεία

- Ο b εκφράζει την **κατά μέσο όρο** μεταβολή (αύξηση ή μείωση ανάλογα με το πρόσημο) της εξαρτημένης μεταβλητής (Y) όταν η ανεξάρτητη (X) μεταβληθεί (αυξηθεί) κατά μία μονάδα

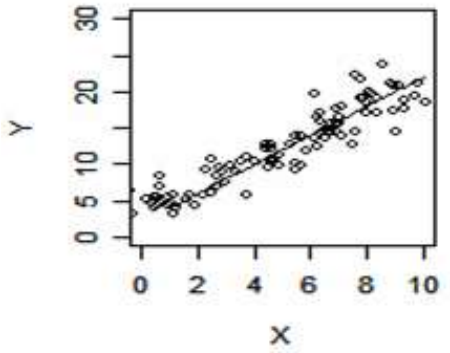
- Ο συντελεστής εξάρτησης b μπορεί να είναι αρνητικός (αρνητική εξάρτηση) ή θετικός αριθμός (θετική εξάρτηση) ή να ισούται προς 0 (απουσία εξάρτησης).

- Έχει σαν μονάδες το λόγο των μονάδων της εξαρτημένης προς τις μονάδες της ανεξάρτητης μεταβλητής (μονάδες Y ανά μονάδες X)

Θετική εξάρτηση (α), αρνητική εξάρτηση (β) και απουσία εξάρτησης (γ) της y από την x

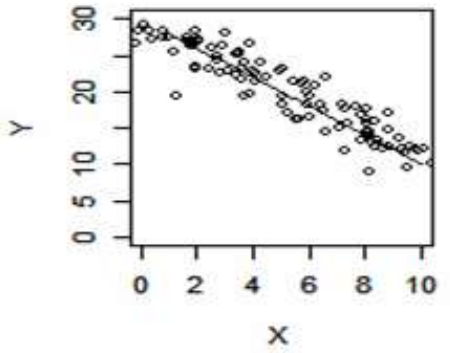


Στα διαγράμματα συσχέτισης, ο b εκφράζει τη γωνία που σχηματίζει η ευθεία εξάρτησης με τον οριζόντιο άξονα (συντελεστής κλίσης, γωνία κλίσης)



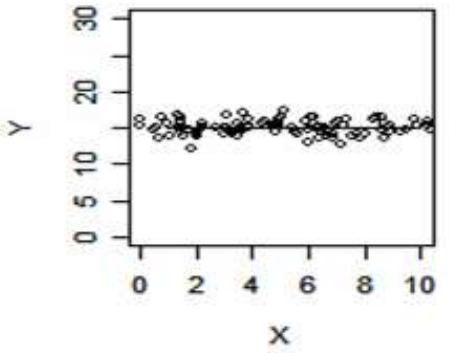
b>0

θετική σχέση



b<0

αρνητική σχέση



b=0

απουσία σχέσης

Εκτίμηση του συντελεστή εξάρτησης (μέσω της μεθόδου ελαχίστων τετραγώνων)

$$b = \frac{\sum_{i=1}^{n} \{(Y_i - \bar{Y})(X_i - \bar{X})\}}{\sum_{i=1}^{n} (X_i - \bar{X})^2} =$$

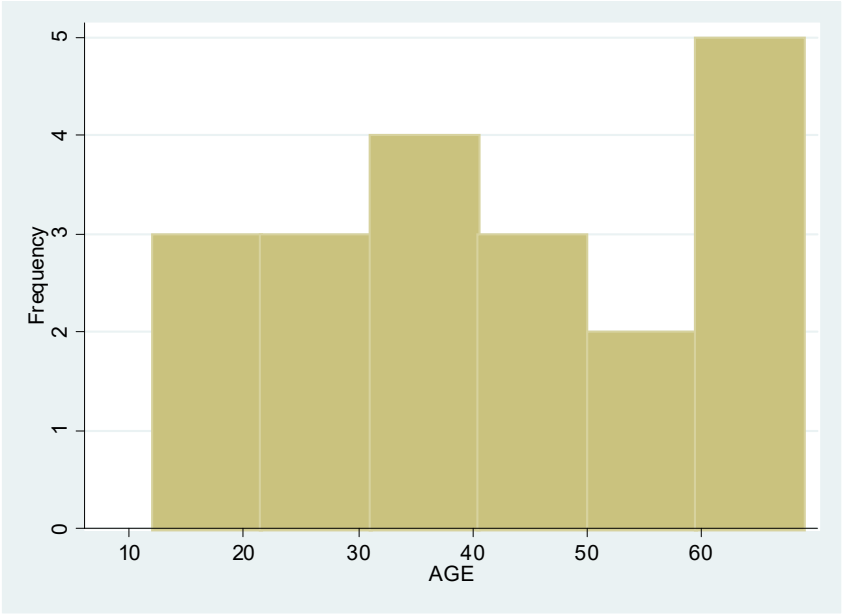
$$= r \frac{SD_Y}{SD_X} \quad (\mu o \nu \acute{a} \delta \varepsilon \varsigma \ Y \ / \ \mu o \nu \acute{a} \delta \varepsilon \varsigma \ X)$$

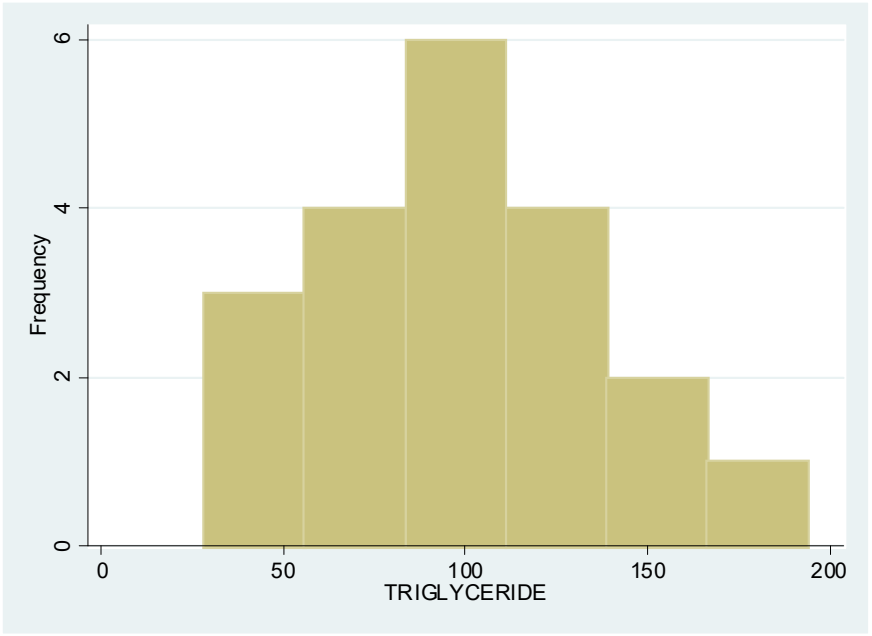
$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

Παράδειγμα

Ηλικία	Τριγλυκερίδια ορού (mg/100ml)
12	28
12	52
18	106
24	87
26	90
27	67
33	99
35	80
38	130
40	50
44	83
46	95
48	111
51	124
57	83
62	119
63	194
67	165
68	152
69	91

- Ηλικία και τριγλυκερίδια σε δείγμα 20 υγιών ανδρών
- Ποια η σχέση μεταξύ ηλικίας και τριγλυκεριδίων;


Frequency
0
1
2
3
4
5
10
20
30
40
50
60
AGE


Frequency
0
2
4
6
0
50
100
150
200
TRIGLYCERIDE

Ποια θα θεωρηθεί εξαρτημένη μεταβλητή και ποια ανεξάρτητη;

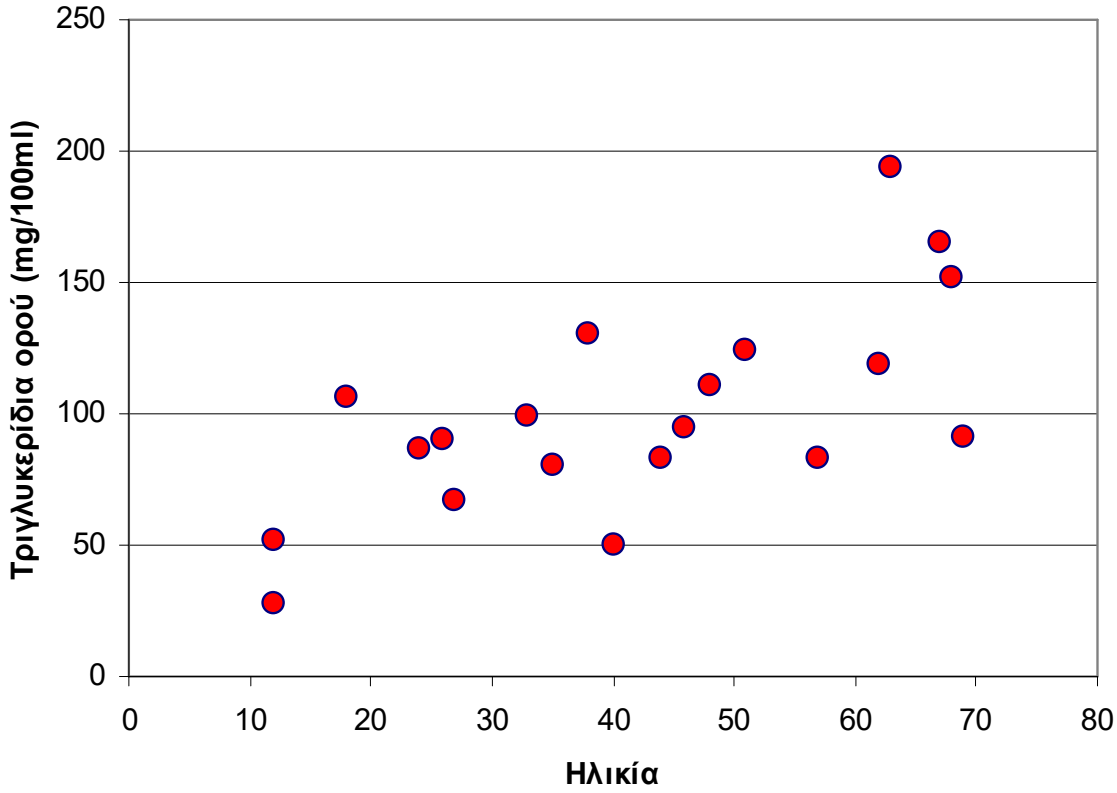
- Εννοιολογικά: τριγλυκερίδια εξαρτώνται από την ηλικία και όχι η ηλικία από τα τριγλυκερίδια

- Με βάση την κατανομή τους: η ηλικία δεν έχει κανονική κατανομή ενώ τα τριγλυκερίδια έχουν κατά προσέγγιση κανονική κατανομή

→Εξαρτημένη (Υ) : ΤΡΙΓΛΥΚΕΡΙΔΙΑ

→Ανεξάρτητη (Χ): ΗΛΙΚΙΑ

Scatter plot (στικτόγραμμα) ηλικίας (X) και τριγλυκεριδίων (Y)



- Τι τιμή αναμένετε να έχει ο b;
  - Περίπου 0, >0 ή <0;

Ηλικία (χ) (σε έτη) και τιμή των τριγλυκεριδίων του ορού (y) (σε χιλιοστόγραμμα ανά 100 χιλιοστόλιτρα) είκοσι υγιών ανδρών

x	ψ	$(x-\bar{x})$	$(\psi-\bar{\psi})$	$(x-\bar{x})^2$	$(\psi-\bar{\psi})^2$	$(x-\bar{x})*(\psi-\bar{\psi})$
12	28	-30	-72	900	5184	+2160
12	52	-30	-48	900	2304	+1440
18	106	-24	+6	576	36	-144
24	87	-18	-13	324	169	+234
26	90	-16	-10	256	100	+160
27	61	-15	-39	225	1521	+585
33	99	-9	-1	81	1	+9
35	80	-7	-20	49	400	+140
38	130	-4	+30	16	900	-120
40	50	-2	-50	4	2500	+100
44	83	+2	-17	4	289	-34
46	95	+4	-5	16	25	-20
48	111	+6	+11	36	121	+66
51	124	+9	+24	81	576	+216
57	83	+15	-17	225	289	-255
62	119	+20	+19	400	361	+380
63	194	+21	+94	441	8836	+1974
67	165	+25	+65	625	4225	+1625
68	152	+26	+52	676	2704	+1352
69	91	+27	-9	729	81	-243
840	**2000**	**Πάντοτε =0**	**Πάντοτε =0**	**6564**	**30622**	**+9625**

$$\bar{X} = \frac{\sum_i X_i}{n} = \frac{840}{20} = 42 \ \text{έτη}$$

$$\bar{Y} = 100 \ \text{mgr/100 ml}$$

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

$$= + \frac{9625}{6564} = +1.466 \ \text{mgr/100 ml/year}$$

Εξίσωση της ευθείας

- a=100, b=1,466

- Η εξίσωση που περιγράφει την εξάρτηση της τιμής των τριγλυκεριδίων του ορού από την ηλικία είναι:

$$\hat{Y}_i = a + b\ (X_i - \bar{X}) = 100 + 1{,}466(X_i - 42)$$

Εξίσωση της ευθείας

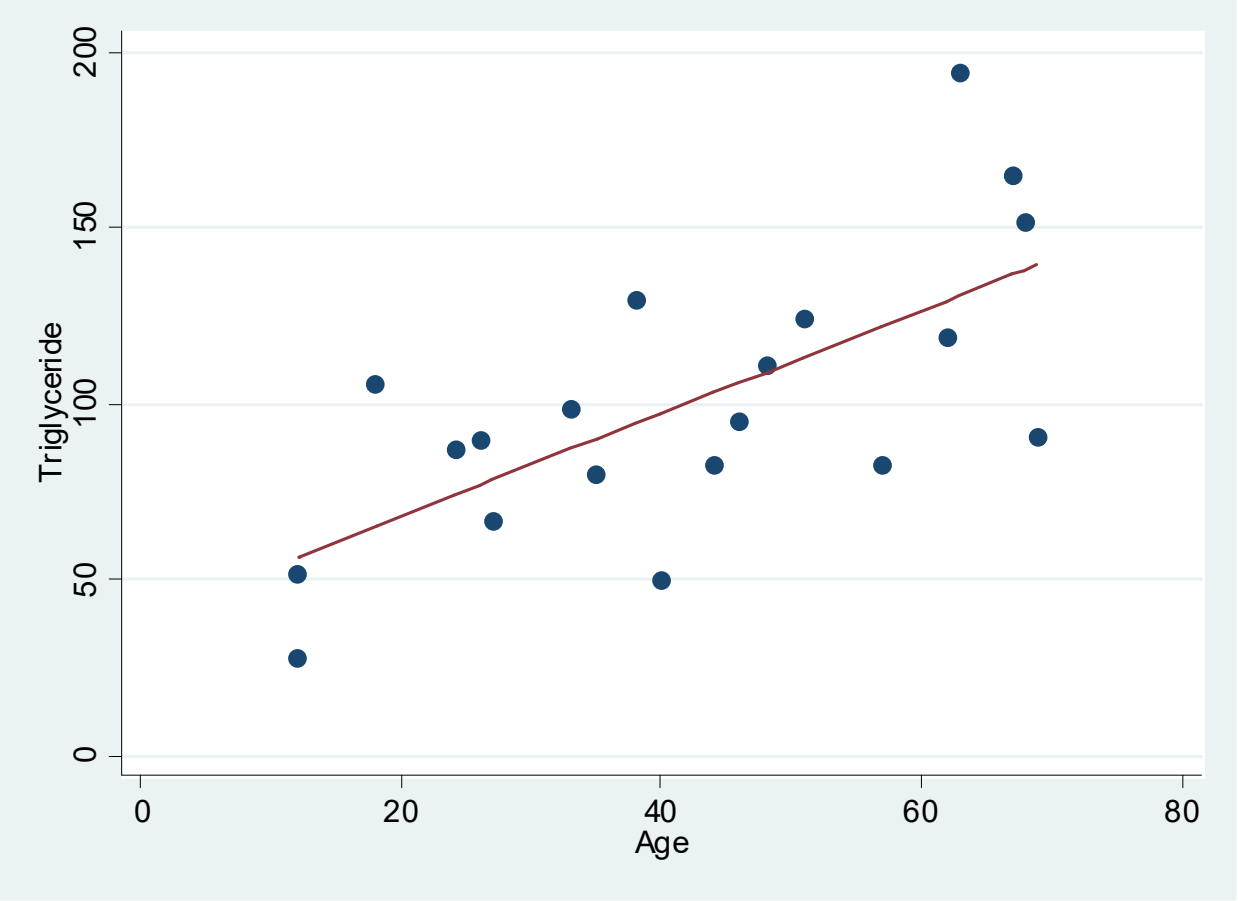
$$\hat{Y}_i = a + b\ (X_i - \bar{X}) = 100 + 1{,}466(X_i - 42)$$

Για να σχεδιάσω αυτή την ευθεία, αρκεί να βρω 2 σημεία της για αυθαίρετα επιλεγμένα Χ και στη συνέχεια να τα ενώσω, π.χ.

Υπολογισμός προβλεπόμενης τιμής:

- Για Χ=22, $\hat{Y} = 100 + 1{,}466(22 - 42) = 70{,}68\ mgr/100\,mlt$
- Για Χ=62, $\hat{Y} = 100 + 1{,}466(62 - 42) = 129{,}32\ mgr/100\,mlt$

Εξίσωση της ευθείας



Ερμηνεία

$$\hat{Y}_i = a + b\ (X_i - \bar{X}) = 100 + 1{,}466(X_i - 42)$$

- Όταν η ηλικία αυξηθεί κατά ένα έτος η τιμή των τριγλυκεριδίων του ορού αναμένεται να αυξηθεί **κατά μέσο όρο** κατά 1,466 mgr/100mlt

- Εναλλακτικά, σε άτομα που η ηλικία τους διαφέρει κατά 1 έτος, η τιμή των τριγλυκεριδίων του ορού αναμένεται να διαφέρει κατά μέσο όρο κατά 1,466 mgr/mlt

- **Είναι όμως σημαντική η σχέση αυτή;**

Στατιστική αξιολόγηση του b

- Η τιμή του b εκφράζει την απουσία (b=0) ή την παρουσία θετικής (b>0) ή αρνητικής συσχέτισης (b<0) μεταξύ 2 μεταβλητών

- Αφού υπολογίσουμε το συντελεστή εξάρτησης b μας ενδιαφέρει στη συνέχεια να διερευνήσουμε αν η σχέση μεταξύ των 2 μεταβλητών είναι στατιστική σημαντική

  → Αν δηλαδή η τιμή του πραγματικού β (από τον πληθυσμό) διαφέρει στατιστικά σημαντικά από το 0

H_0: $\beta=0$
H_1: $\beta\neq0$

Στατιστική αξιολόγηση του b

Μέθοδος Α: Όταν το πιθανό σφάλμα (SE_b) του b είναι άγνωστο

Αξιολόγηση μέσω του αντίστοιχου **συντελεστή συσχέτισης (r)** δεδομένου ότι:

$$b = \frac{\sum_{i=1}^{n}\{(Y_i - \overline{Y})(X_i - \overline{X})\}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = r\frac{SD_Y}{SD_X}$$

b και r ίδιο πρόσημο

Στατιστική αξιολόγηση του b

Μέθοδος Β: Όταν το πιθανό σφάλμα (SE_b) του b γνωστό

Τι θα παίξει ρόλο στο αν τελικά θα βρούμε μια διαφορά ή όχι;

1. Το μέγεθος του **b**
   π.χ. Όσο πιο κοντά στο 0 → τόσο πιο πιθανό να μην υπάρχει κάποια σχέση

2. Το πόσο σίγουροι είμαστε για την εκτίμησή μας → **SE_b**

Υπολογίζουμε το πηλίκο: $\frac{|b|}{SE_b}$

και με βάση αυτό θα πραγματοποιήσουμε τον έλεγχο της υπόθεσης

Όσο πιο μεγάλη η τιμή του → τόσο πιο πιθανό να υπάρχει όντως σχέση

Πώς κρίνουμε αν η τιμή είναι μεγάλη;;

Στατιστική αξιολόγηση του b (όταν το πιθανό σφάλμα (SEb) του b γνωστό

Η ποσότητα $\frac{b}{SE_b}$ ακολουθεί την **t κατανομή με n-2 ΒΕ**

Ελέγχουμε την τιμή $\frac{b}{SE_b}$ στον πίνακα με τις οριακές τιμές (συνήθως με την τιμή που αντιστοιχεί στο 5% επίπεδο σημαντικότητας σε n-2 βαθμούς ελευθερίας)

Αν $\frac{|b|}{SE_b}$ ≥οριακή τιμή → απορρίπτω H_0 και συμπεραίνω ότι η σχέση είναι στατιστικά σημαντική

Αν $\frac{|b|}{SE_b}$ <οριακή τιμή → η σχέση δεν είναι στατιστικά σημαντική

Θυμόμαστε από τα μαθηματικά:

Αν $\theta > 0$ τότε:

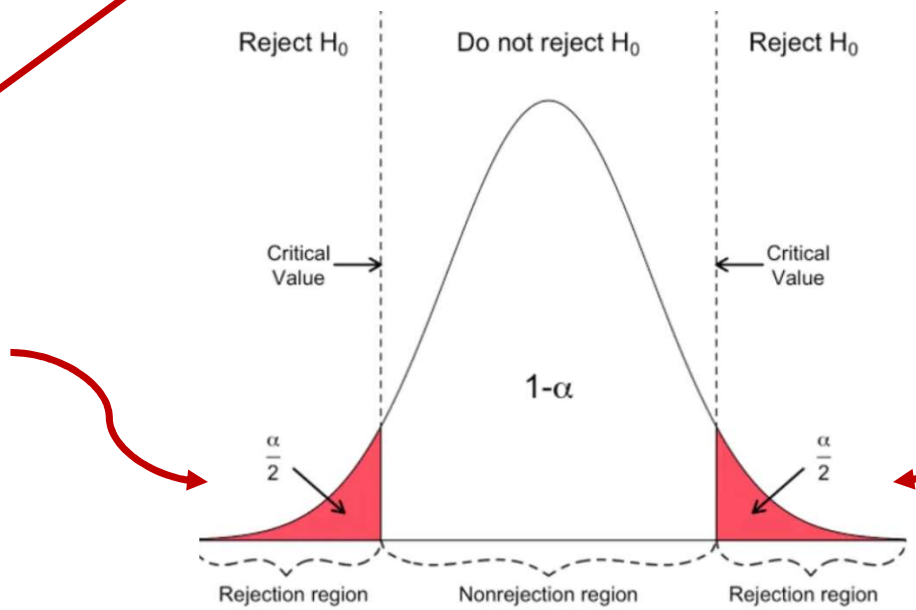
1. $|x| \geq \theta \Leftrightarrow x \leq -\theta$ ή $x \geq \theta$

2. $|x| \leq \theta \Leftrightarrow -\theta \leq x \leq \theta$

$$\frac{|b|}{SE_b} > t_{n-2,\,0.05}$$

$$\frac{b}{SE_b} < -t_{n-2,\,0.05}$$

$$\frac{b}{SE_b} > t_{n-2,\,0.05}$$



95% όρια αξιοπιστίας (CI) του b

- Τα **95% CI** του συντελεστή εξάρτησης b υπολογίζονται από τον τύπο:

$$\mathbf{b \pm t_{0.05,(n-2)} \cdot SE(b)}$$

- Τα 95% CI μας επιτρέπουν να αξιολογήσουμε αν ο b διαφέρει στατιστικά σημαντικά από το 0

  - Αν τα όρια δεν περιλαμβάνουν το 0

  → διαφέρει στατιστικά σημαντικά π.χ. (-3, -1) ή (0.25, 0.81)

  - Αν τα όρια περιλαμβάνουν το 0

  → μη στατιστικά σημαντική σχέση π.χ. (-0.92, 1,12)

Αξιολόγηση του b στο παράδειγμα ηλικίας-τριγλυκεριδίων (μέθοδος Α)

Ηλικία (χ) (σε έτη) και τιμή των τριγλυκεριδίων του ορού (y) (σε χιλιοστόγραμμα ανά 100 χιλιοστόλιτρα) είκοσι υγιών ανδρών

x	ψ	$(x-\bar{x})$	$(\psi-\bar{\psi})$	$(x-\bar{x})^2$	$(\psi-\bar{\psi})^2$	$(x-\bar{x})*(\psi-\bar{\psi})$
12	28	-30	-72	900	5184	+2160
12	52	-30	-48	900	2304	+1440
18	106	-24	+6	576	36	-144
24	87	-18	-13	324	169	+234
26	90	-16	-10	256	100	+160
27	61	-15	-39	225	1521	+585
33	99	-9	-1	81	1	+9
35	80	-7	-20	49	400	+140
38	130	-4	+30	16	900	-120
40	50	-2	-50	4	2500	+100
44	83	+2	-17	4	289	-34
46	95	+4	-5	16	25	-20
48	111	+6	+11	36	121	+66
51	124	+9	+24	81	576	+216
57	83	+15	-17	225	289	-255
62	119	+20	+19	400	361	+380
63	194	+21	+94	441	8836	+1974
67	165	+25	+65	625	4225	+1625
68	152	+26	+52	676	2704	+1352
69	91	+27	-9	729	81	-243
840	**2000**	**Πάντοτε =0**	**Πάντοτε =0**	**6564**	**30622**	**+9625**

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$= +\frac{9625}{\sqrt{6564 * 30622}}$$

$$= +0.679$$

BE: n-2=20-2=18

B.E.	10%	5%	1%
18	**0,378**	**0,444**	**0,561**

r=0.679

|r|=0.679 >0.561

→ Στατιστικά σημαντικό στο 1% επίπεδο

Επίπεδο σημαντικότητας

DF	.10	.05	.02	.01
1	.9877	.9969	.9995	.9999
2	.9000	.9500	.9800	.9900
3	.8054	.8783	.9343	.9587
4	.7293	.8114	.8822	.9172
5	.6694	.7545	.8329	.8745
6	.6215	.7067	.7887	.8343
7	.5822	.6664	.7498	.7977
8	.5494	.6319	.7155	.7646
9	.5214	.6021	.6851	.7348
10	.4973	.5760	.6581	.7079
11	.4762	.5529	.6339	.6835
12	.4575	.5324	.6120	.6614
13	.4409	.5140	.5923	.6411
14	.4259	.4973	.5742	.6226
15	.4124	.4821	.5577	.6055
16	.4000	.4683	.5425	.5897
17	.3887	.4555	.5285	.5751
18	.3783	.4438	.5155	.5614
19	.3687	.4329	.5034	.5487
20	.3598	.4227	.4921	.5368
21	.3515	.4132	.4815	.5256
22	.3438	.4044	.4716	.5151
23	.3365	.3961	.4622	.5052
24	.3297	.3882	.4534	.4958
25	.3233	.3809	.4451	.4869
26	.3172	.3739	.4372	.4785
27	.3115	.3673	.4297	.4705
28	.3061	.3610	.4226	.4629
29	.3009	.3550	.4158	.4556
30	.2960	.3494	.4093	.4487
31	.2913	.3440	.4032	.4421

Αξιολόγηση του b στο παράδειγμα ηλικίας-τριγλυκεριδίων (μέθοδος Β)

- b=1,466 και SE(b)=0,374

  Επομένως: $b/SE_b = 1,466/0,374 = 3,92$

- Ελέγχουμε με πίνακα t κατανομής στους 20-2=18 ΒΕ

Β.Ε.	10%	5%	1%
18	**1,73**	**2,10**	**2,88**

b/SE_b =3,92>2.88

Στατιστικά πολύ σημαντική συσχέτιση (p<1%)

df	0.10	0.05	0.025	0.01
2	2.9200	4.3027	6.2054	9.9250
3	2.3534	3.1824	4.1765	5.8408
4	2.1318	2.7765	3.4954	4.6041
5	2.0150	2.5706	3.1634	4.0321
6	1.9432	2.4469	2.9687	3.7074
7	1.8946	2.3646	2.8412	3.4995
8	1.8595	2.3060	2.7515	3.3554
9	1.8331	2.2622	2.6850	3.2498
10	1.8125	2.2281	2.6338	3.1693
11	1.7959	2.2010	2.5931	3.1058
12	1.7823	2.1788	2.5600	3.0545
13	1.7709	2.1604	2.5326	3.0123
14	1.7613	2.1448	2.5096	2.9768
15	1.7531	2.1315	2.4899	2.9467
16	1.7459	2.1199	2.4729	2.9208
17	1.7396	2.1098	2.4581	2.8982
18	1.7341	2.1009	2.4450	2.8784
19	1.7291	2.0930	2.4334	2.8609
20	1.7247	2.0860	2.4231	2.8453
21	1.7207	2.0796	2.4138	2.8314
22	1.7171	2.0739	2.4055	2.8188
23	1.7139	2.0687	2.3979	2.8073
24	1.7109	2.0639	2.3910	2.7970
25	1.7081	2.0595	2.3846	2.7874
26	1.7056	2.0555	2.3788	2.7787
27	1.7033	2.0518	2.3734	2.7707
28	1.7011	2.0484	2.3685	2.7633
29	1.6991	2.0452	2.3638	2.7564
30	1.6973	2.0423	2.3596	2.7500
31	1.6955	2.0395	2.3556	2.7440
32	1.6939	2.0369	2.3518	2.7385
33	1.6924	2.0345	2.3483	2.7333
34	1.6909	2.0322	2.3451	2.7284
35	1.6896	2.0301	2.3420	2.7238
36	1.6883	2.0281	2.3391	2.7195
37	1.6871	2.0262	2.3363	2.7154
38	1.6860	2.0244	2.3337	2.7116
39	1.6849	2.0227	2.3313	2.7079
40	1.6839	2.0211	2.3289	2.7045
41	1.6829	2.0195	2.3267	2.7012
42	1.6820	2.0181	2.3246	2.6981
43	1.6811	2.0167	2.3226	2.6951

95% CI του b

- 95% CI :

$$1{,}466 \pm 2{,}10 * 0{,}374 \rightarrow (0{,}681 - 2{,}251)$$

- Τα όρια δεν περιλαμβάνουν το 0 → στατιστικά σημαντική σχέση ηλικίας-τριγλυκεριδίων στο 5% επίπεδο σημαντικότητας

Ερμηνεία

- Όταν η ηλικία αυξηθεί κατά ένα έτος η τιμή των τριγλυκεριδίων του ορού αναμένεται να αυξηθεί **κατά μέσο όρο** κατά 1,466 mgr/100 ml (ενώ με 95% πιθανότητα η αύξηση αυτή μπορεί να κυμαίνεται από 0,68 μέχρι 2,25 mgr/100 ml)

- Εναλλακτικά, σε άτομα που η ηλικία τους διαφέρει κατά 1 έτος, η τιμή των τριγλυκεριδίων του ορού αναμένεται να διαφέρει κατά μέσο όρο κατά 1,466 mgr/100 ml

- **Ανά δεκαετία:** b=1,466*10=14,66 mgr/100 ml/έτος

  95% CI: 6,81-22,51 → Σε άτομα που η ηλικία τους διαφέρει κατά 10 έτη, η τιμή των τριγλυκεριδίων του ορού αναμένεται να διαφέρει κατά μέσο όρο κατά 14,66 χιλιοστόγραμμα ανά 100 χιλιοστόλιτρα

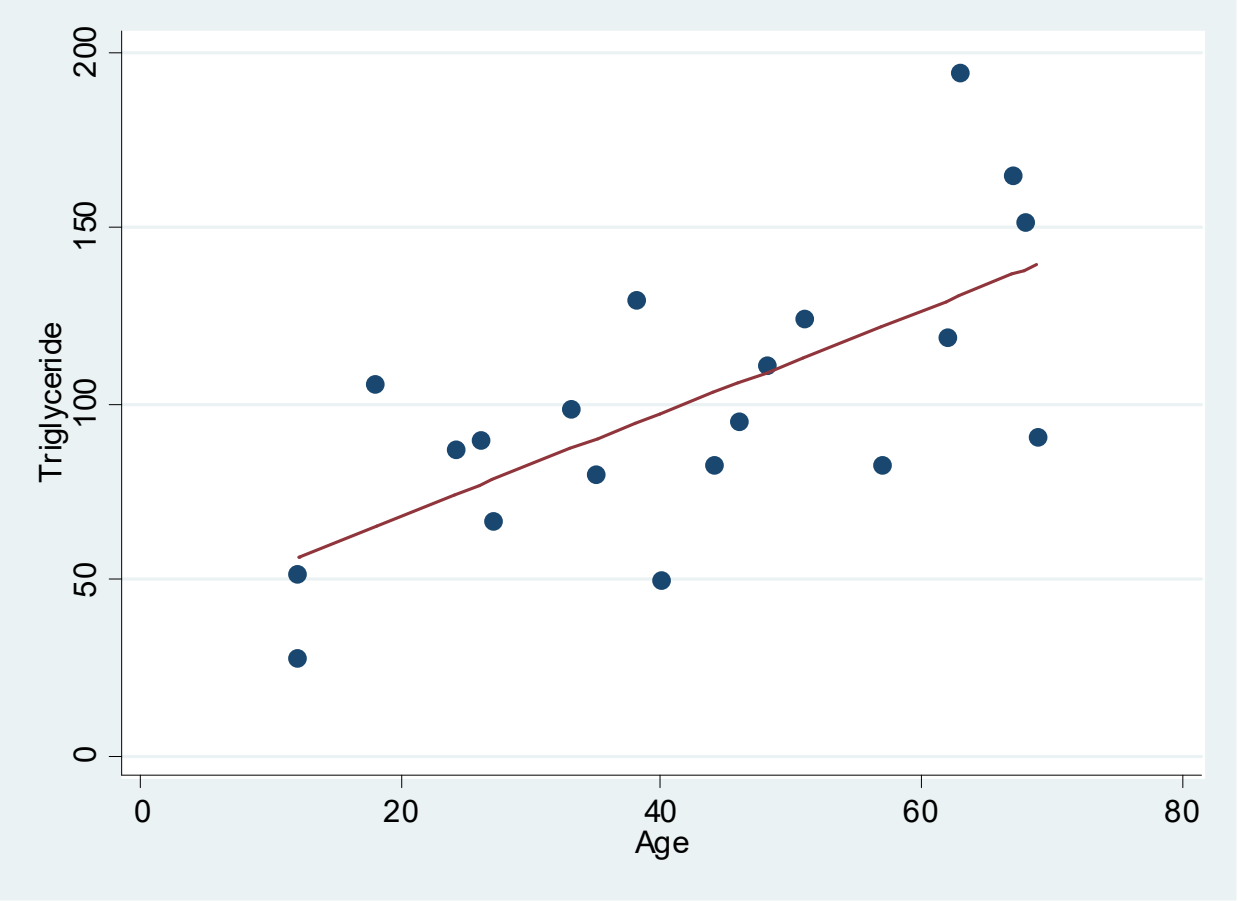
Πληροφορίες που μας δίνει η ευθεία παλινδρόμησης

1. Η τιμή και το πρόσημο του συντελεστή εξάρτησης υποδηλώνει την ύπαρξη γραμμικής σχέσης μεταξύ των 2 μεταβλητών και το είδος της (θετική ή αρνητική σχέση) → όπως και ο συντελεστής συσχέτισης r

Επιπλέον

2. Η τιμή του b μας δίνει την πληροφορία κατά πόσο αναμένεται να μεταβληθεί κατά μέσο όρο η Υ για κάθε μία μονάδα μεταβολής της Χ
3. Επιτρέπει να εκτιμήσουμε το αναμενόμενο Υ για ένα άτομο με δεδομένη τιμή Χ

Πρόβλεψη των τιμών του Y για δοθείσες τιμές του X



Πρόβλεψη των τιμών του Υ για δοθείσες τιμές του Χ

- Πρέπει να αποφεύγονται οι προβλέψεις για τιμές της Χ εκτός του εύρους των τιμών της στο δείγμα (**extrapolation**)

  - Π.χ. στο παράδειγμά μας, μπορεί η σχέση ηλικίας-τριγλυκεριδιών να μην είναι γραμμική για άτομα ηλικίας >70 ετών ή <10 ετών

Στατιστικές δοκιμασίες για τη διερεύνηση σχέσης μεταξύ 2 παραγόντων

Παράγοντας 1	Παράγοντας 2	
	Ποσοτική	Ποιοτική
Ποιοτική Με 2 επίπεδα	t-test για ανεξάρτητα δείγματα	χ^2-test
Ποσοτική	• Συντελεστής συσχέτισης με στατιστική αξιολόγηση • Απλή γραμμικη εξάρτηση	

Σημείωση: Οι περισσότερες δοκιμασίες για **ποσοτικά** χαρακτηριστικά υποθέτουν την **κανονική** κατανομή των χαρακτηριστικών

Παράδειγμα από δημοσιευμένη μελέτη

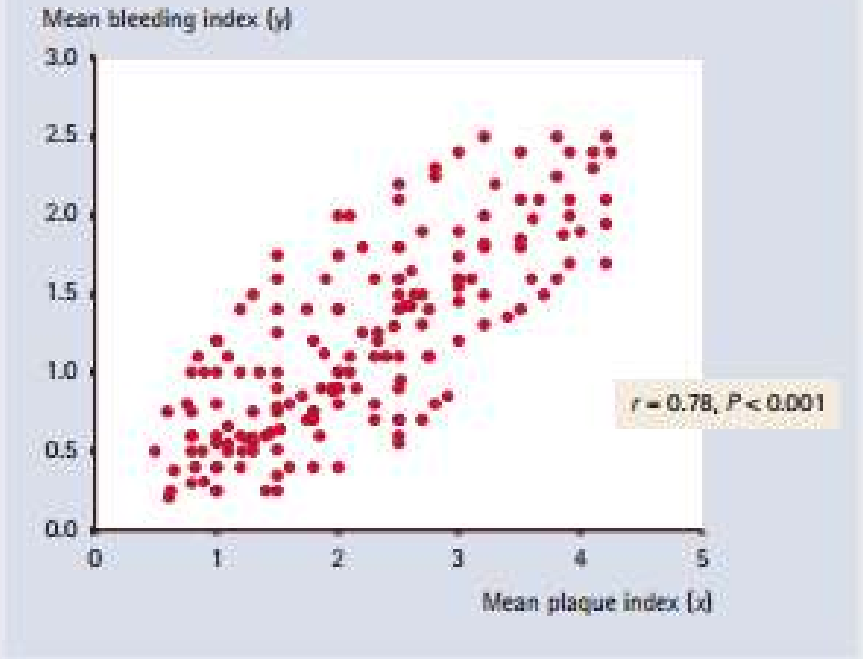


Fig. 1a Scatter diagram showing the relationship between the mean bleeding index per child and the mean plaque index per child in a sample of 170 12-year-old schoolchildren (derived from data kindly provided by Dr Gareth Griffiths of the Eastman Dental Institute for Oral Health Care Sciences, University College London)

Petrie et al. Further statistics in dentistry Part 6: Multiple linear regression. Bristish Dental Journal, 2002

Παράδειγμα

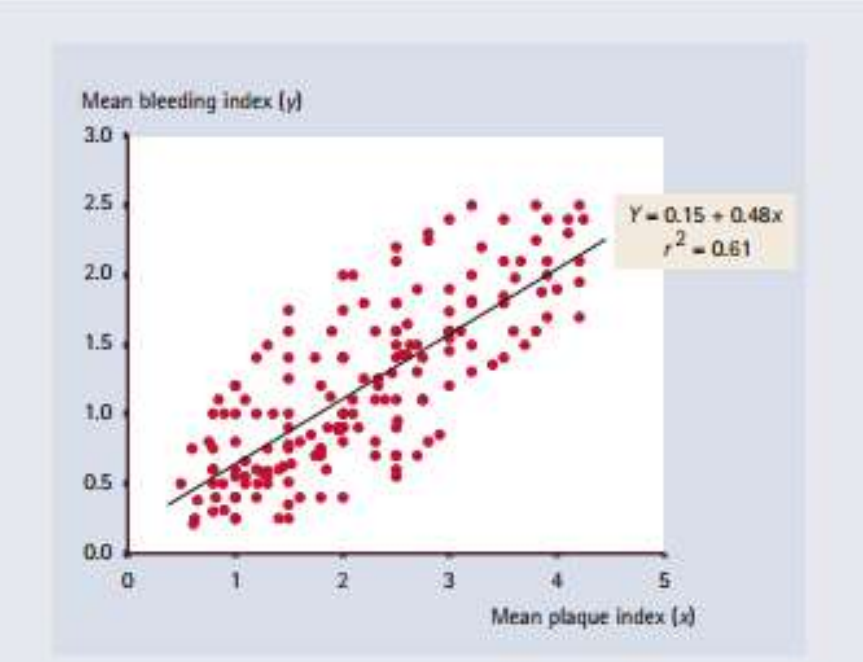
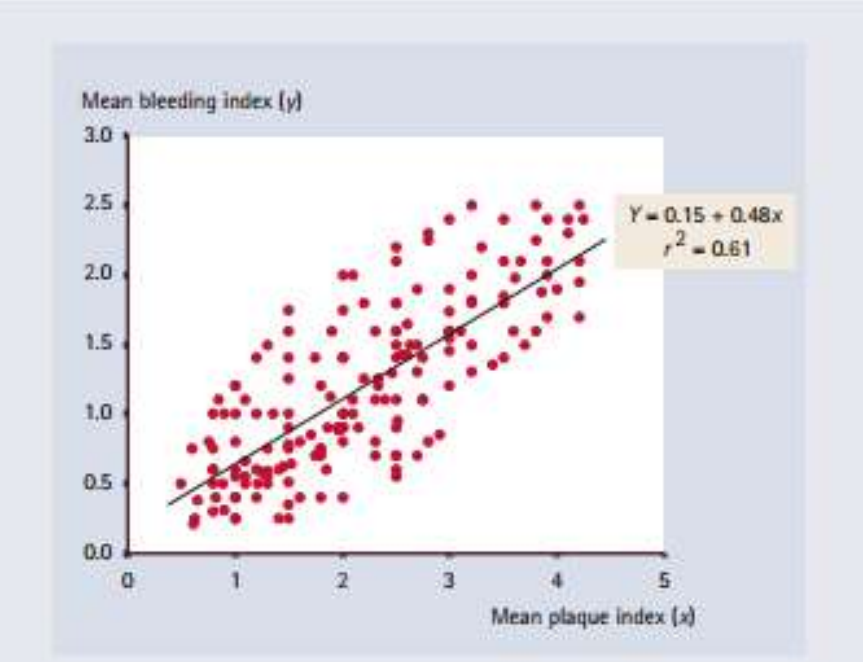


Fig. 1b Estimated linear regression line of the mean bleeding index against the mean plaque index using the data of Fig. 1a. Intercept, $a = 0.15$; slope, $b = 0.48$ (95% CI = 0.42 to 0.54, $P < 0.001$), indicating that the mean bleeding index increases on average by 0.48 as the mean plaque index increases by one

- Τι πληροφορίες μας δίνει αυτό;

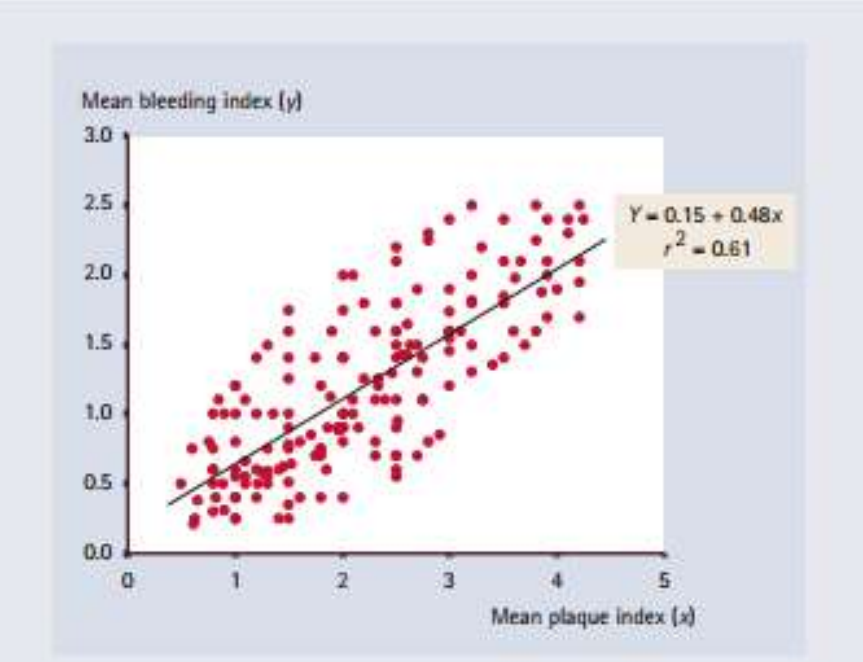
1. Αν υπάρχει στατιστικά σημαντική σχέση και το είδος της σχέσης (θετική-αρνητική)



- Στατιστικά σημαντική σχέση
(γιατί b=0.48, p<0.001)

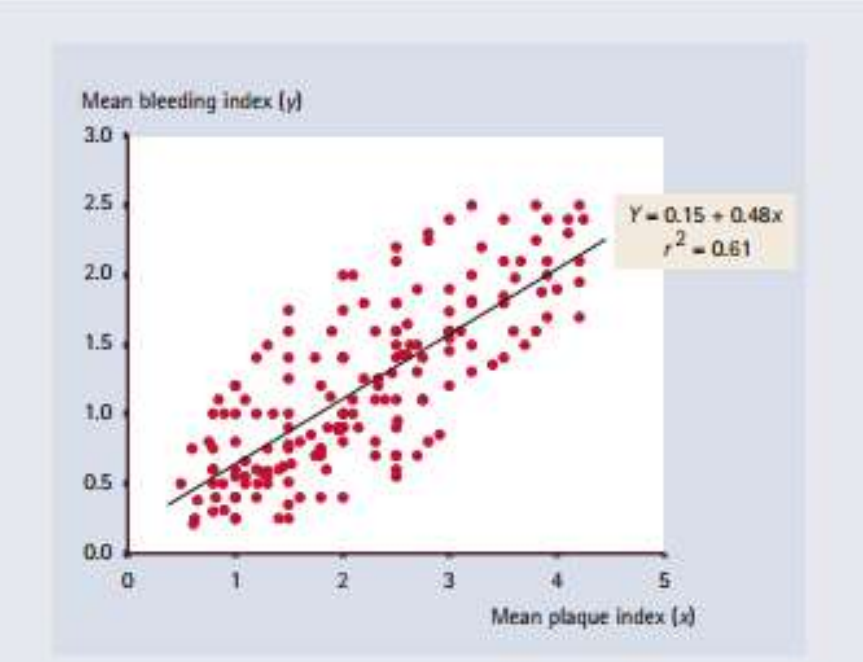
- Θετική συσχέτιση
(γιατί b>0)

2. Πόσο θα διαφέρει κατά μέσο όρο bleeding index σε ένα παιδί που έχει 1 μονάδα υψηλότερο plaque index από ένα άλλο παιδί?



→ Κατά 0.48 μονάδες

3. Πόσο αναμένεται να είναι κατά μέσο όρο το bleeding index σε παιδί με plaque index=3?



Bleeding index
= 0.15 + 0.48 * 3
= 1.59