Στατιστική για Πολιτικούς Μηχανικούς
ΣΥΣΧΕΤΙΣΗ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗ

Δημήτρης Κουγιουμτζής

18 Δεκεμβρίου 2012

Συσχέτιση

Δύο τ.μ. X και Y συσχετίζονται:

- Η μία επηρεάζει την άλλη
- Επηρεάζονται και οι δύο από κάποια άλλη

σ_X^2, σ_Y^2: διασπορά

συνδιασπορά των X και Y:
$\sigma_{XY} = \mathrm{Cov}(X, Y) = \mathrm{E}(X, Y) - \mathrm{E}(X)\mathrm{E}(Y)$,

συντελεστής συσχέτισης ρ

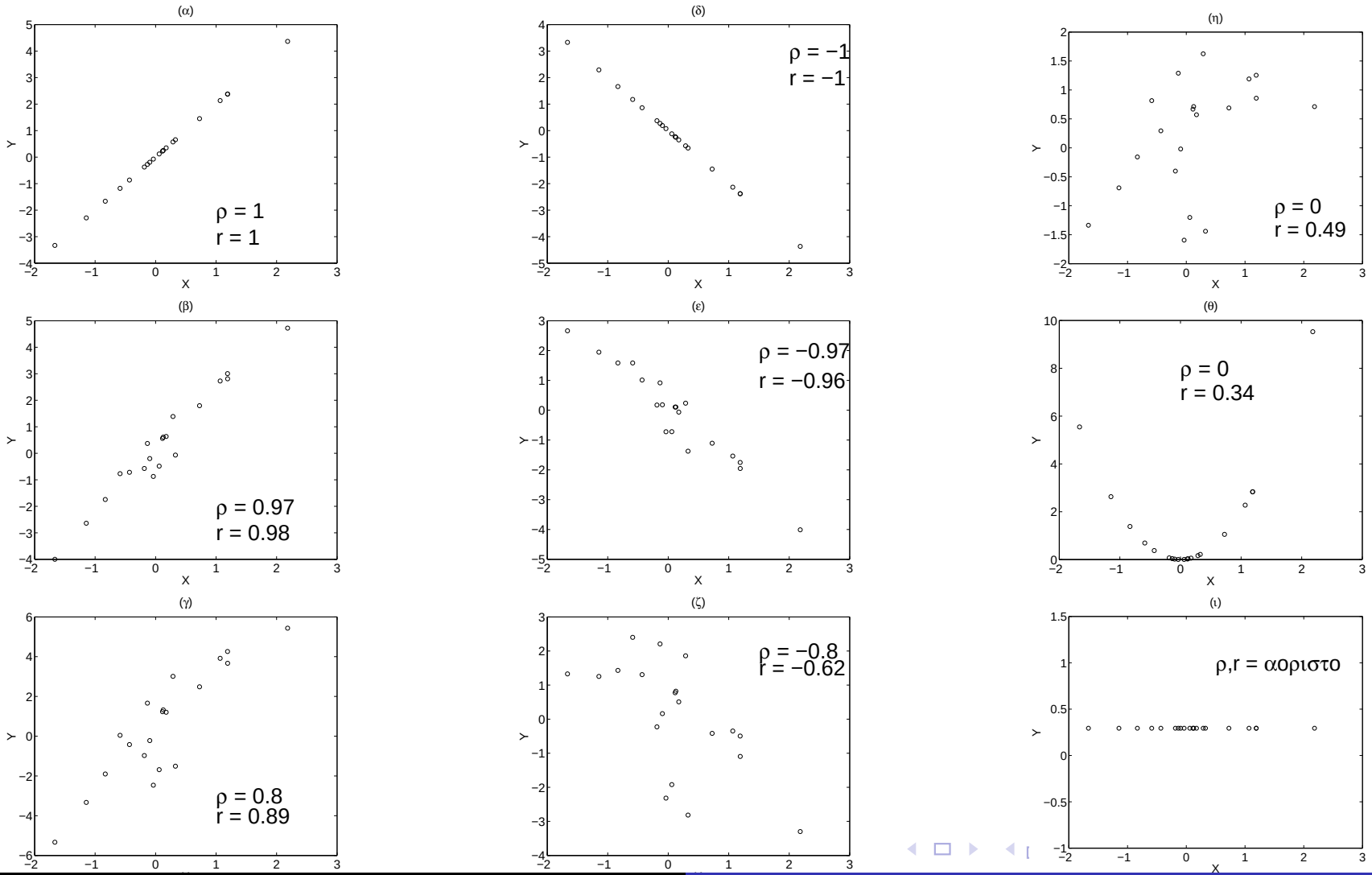
$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Ιδιότητες του ρ

- $\rho \in [-1, 1]$
- $\rho = 1$: *τέλεια θετική* συσχέτιση
- $\rho = -1$: *τέλεια αρνητική* συσχέτιση
- ρ 'κοντά' στο -1 ή 1 $\rightarrow$ ισχυρή συσχέτιση
- ρ 'κοντά' στο 0 $\rightarrow$ οι τ.μ. είναι πρακτικά ασυσχέτιστες
- ρ δεν εξαρτάται από τη μονάδα μέτρησης των X και Y
- ρ είναι συμμετρικός ως προς τις X και Y

Διάγραμμα διασποράς

Δείγμα των X και Y κατά ζεύγη: $(x_1, y_1), \ldots, (x_n, y_n)$



Σημειακή εκτίμηση του ρ

Εκτίμηση διασποράς

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

Εκτίμηση συνδιασποράς

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

Εκτίμηση του συντελεστή συσχέτισης

$$\rho = \frac{\sigma_{XY}}{\sigma_X\,\sigma_Y} \quad \rightarrow \quad \hat{\rho} \equiv r = \frac{s_{XY}}{s_X\,s_Y}$$

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

Σημειακή εκτίμηση του ρ (συνέχεια)

- Το r είναι η σημειακή εκτίμηση του ρ από το δείγμα και λέγεται συντελεστής συσχέτισης Pearson
- Μπορούν να υπολογιστούν παραμετρικά διαστήματα εμπιστοσύνης για το ρ.
- Μπορούν να γίνουν παραμετρικοί έλεγχοι υπόθεσης για κάποια τιμή του ρ.
  Η πιο σημαντική υπόθεση είναι H_0: $\rho = 0$.

Συντελεστής προσδιορισμού r^2 (ή σε ποσοστά $100r^2\%$)

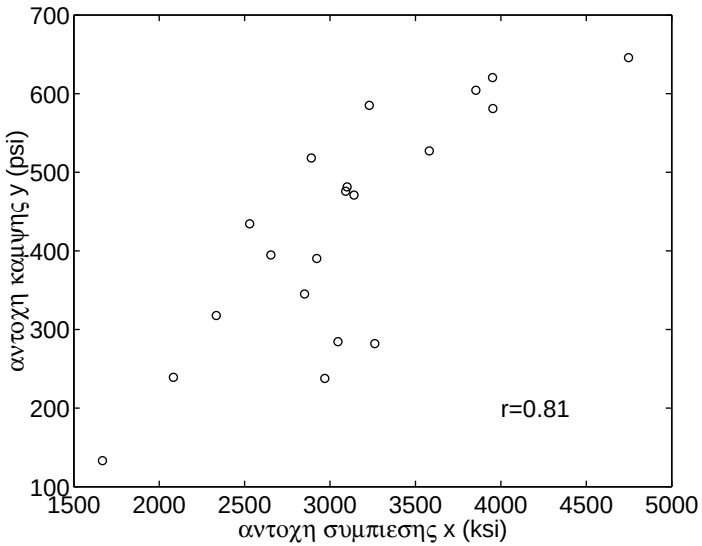
Δηλώνει το ποσοστό μεταβλητότητας που μπορούμε να ερμηνεύσουμε για τη μια τ.μ. όταν γνωρίζουμε την άλλη.

Παράδειγμα

Θέλουμε να εκτιμήσουμε τη συσχέτιση της αντοχής συμπίεσης και αντοχής κάμψης σκυροδέματος κάποιας παρασκευής.

Δείγμα 20 δοκιμίων σκυροδέματος

Α/Α	Αντοχή συμπίεσης x_i (ksi)	Αντοχή κάμψης y_i (psi)
1	1668	133
2	2083	239
3	2334	318
4	2529	434
5	2654	395
6	2851	345
7	2891	518
8	2923	390
9	2970	238
10	3047	284
11	3091	476
12	3100	481
13	3140	471
14	3230	585
15	3262	282
16	3581	527
17	3853	604
18	3951	621
19	3953	581
20	4747	646



Υπολογίζουμε

$\bar{x} = 3092.9 \quad \bar{y} = 428.5$

$\sum_{i=1}^{20} x_i^2 = 200680560$

$\sum_{i=1}^{20} y_i^2 = 4083287$

$\sum_{i=1}^{20} x_i y_i = 28088503$

$$r = \frac{28088503 - 20 \cdot 3092.9 \cdot 428.5}{\sqrt{(200680560 - 20 \cdot 3092.9^2) \cdot (4083287 - 20 \cdot 428.5^2)}} = 0.807$$

Η αντοχή συμπίεσης και η αντοχή κάμψης έχουν γραμμική θετική συσχέτιση αλλά όχι ισχυρή.

Το ποσοστό μεταβλητότητας της αντοχής κάμψης που μπορούμε να εξηγήσουμε γνωρίζοντας την αντοχή συμπίεσης (και αντίστροφα) είναι 0.65.

Θέμα 13

Έλεγχος σημαντικότητας (significance test) για το συντελεστή συσχέτισης Pearson. Παρουσίαση και παράδειγμα.

Θέμα 14

Συντελεστής συσχέτισης Spearman δύο τυχαίων μεταβλητών. Παρουσίαση, ιδιότητες, παραδείγματα.

Θέμα 15

Συντελεστής συσχέτισης Kendall δύο τυχαίων μεταβλητών. Παρουσίαση, ιδιότητες, παραδείγματα.

Θέμα 16

Η αμοιβαία πληροφορία (mutual information) ως μη-γραμμικό μέτρο συσχέτισης δύο μεταβλητών. Παρουσίαση, ιδιότητες, παραδείγματα.

Απλή Γραμμική Παλινδρόμηση

συσχέτιση: γραμμική σχέση δύο τ.μ. X και Y
παλινδρόμηση: εξάρτηση μιας τ.μ. Y από μια άλλη μεταβλητή X

Y: εξαρτημένη μεταβλητή (τυχαία)
X: ανεξάρτητη μεταβλητή (καθορισμένη)

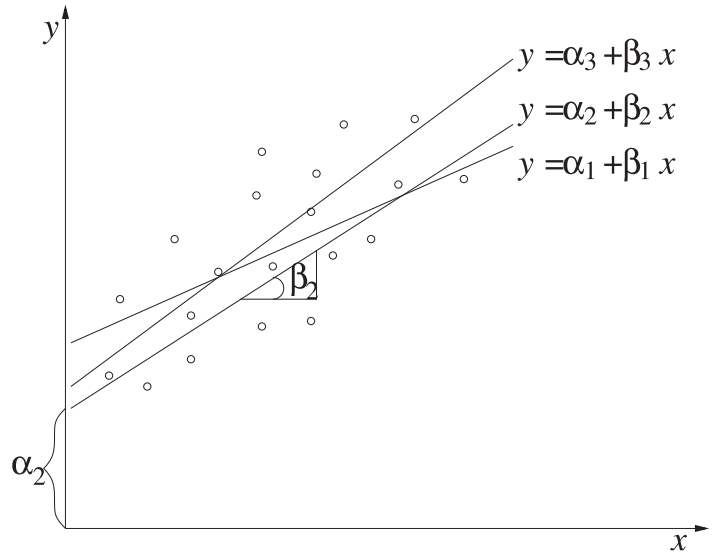
Παράδειγμα: διατμητική αντοχή αργίλου σε διάφορα βάθη
X ? Y ? ;

Γενικά θέλουμε να βρούμε $F_Y(y|X = x)$
Περιοριζόμαστε στη μέση τιμή
και υποθέτουμε γραμμική εξάρτηση

$$\mathrm{E}(Y|X = x) = \alpha + \beta x$$

γραμμική παλινδρόμηση της Y στη X

Παρατηρήσεις
$(x_1, y_1), \ldots, (x_n, y_n)$



α: σταθερός όρος
β: συντελεστής του x (κλίση ευθείας)

Η τ.μ. y_i για κάποια τιμή x_i της X είναι
$y_i = \alpha + \beta x_i + \epsilon_i$

$\epsilon_i = y_i - \mathrm{E}(Y|X = x_i)$ *σφάλμα παλινδρόμησης*

Πρόβλημα παλινδρόμησης:
Ποια είναι η 'καλύτερη' ευθεία;
Ποιες είναι οι 'καλύτερες' εκτιμήσεις των α, β;

Συνθήκες απλής γραμμικής παλινδρόμησης

- Η X είναι *ελεγχόμενη* (καθορισμένη)
- Η εξάρτηση της Y από τη X είναι γραμμική
- $\mathrm{E}(\epsilon_i) = 0$ και $\mathrm{Var}(\epsilon_i) = \sigma^2_\epsilon$ για κάθε x_i

$$\mathrm{Var}(y_i|X = x_i) = \mathrm{Var}(\alpha + \beta x_i + \epsilon_i) = \mathrm{Var}(\epsilon_i)$$
$$\Downarrow$$
$$\mathrm{Var}(Y|X = x) \equiv \sigma^2_{Y|X} = \sigma^2_\epsilon \equiv \sigma^2$$

ομοσκεδαστικότητα: η διασπορά της Y δε μεταβάλλεται με τη X
ετεροσκεδαστικότητα: η διασπορά της Y μεταβάλλεται με τη X.

Άγνωστοι (παράμετροι) παλινδρόμησης: α, β, σ^2

[Συνήθως υποθέτουμε $\quad Y|X = x \sim \mathrm{N}(\alpha + \beta x, \sigma^2)$]

Εκτίμηση των παραμέτρων της ευθείας παλινδρόμησης

Μέθοδος ελαχίστων τετραγώνων:

Το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων από την ευθεία είναι το ελάχιστο

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \epsilon_i^2 \qquad \text{ή} \qquad \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Λύση:

$$\left.\begin{array}{l} \frac{\partial \sum (y_i-\alpha-\beta x_i)^2}{\partial \alpha} = 0 \\ \frac{\partial \sum (y_i-\alpha-\beta x_i)^2}{\partial \beta} = 0 \end{array}\right\} \begin{array}{l} \sum_{i=1}^{n} y_i = n\alpha + \beta \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 \end{array}$$

Εκτιμήσεις των β και α είναι

$$b = \frac{s_{XY}}{s_X^2} \qquad a = \bar{y} - b\bar{x}$$

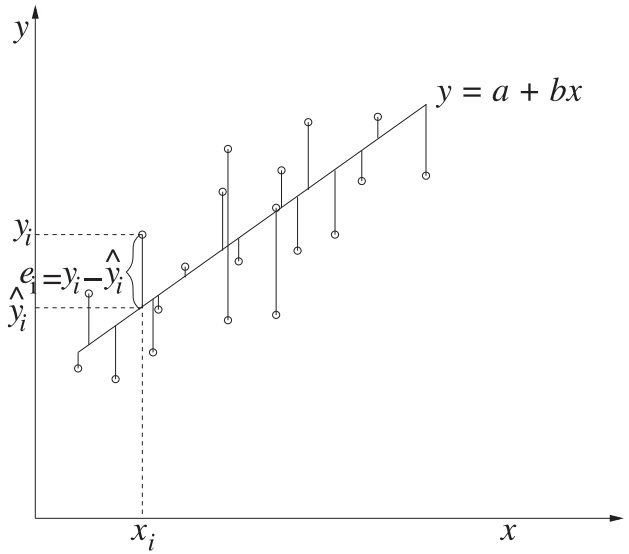
ευθεία ελαχίστων τετραγώνων: $\hat{y} = a + bx$

Εκτίμηση της διασποράς των σφαλμάτων

Για κάθε x_i: $\hat{y}_i = a + bx_i$

$e_i = y_i - \hat{y}_i$: σφάλμα ελαχίστων τετραγώνων ή **υπόλοιπο**

e_i: εκτίμηση του σφάλματος παλινδρόμησης ϵ_i



Η εκτίμηση της διασποράς σ^2 του σφάλματος

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

θέτοντας $\hat{y}_i = a + bx_i$

$$s^2 = \frac{n-1}{n-2}\left(s_Y^2 - \frac{s_{XY}^2}{s_X^2}\right) = \frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2)$$

Παρατηρήσεις

- Η ευθεία ελαχίστων τετραγώνων περνάει από το σημείο $(\bar{x}, \bar{y})$: $a + b\bar{x} = \bar{y} - b\bar{x} + b\bar{x} = \bar{y}$
  Άρα η ευθεία ελαχίστων τετραγώνων μπορεί επίσης να οριστεί ως $y_i - \bar{y} = b(x_i - \bar{x})$
- Η εκτίμηση των α και β με τη μέθοδο των ελαχίστων τετραγώνων **δεν** προϋποθέτει
  (i) σταθερή διασπορά της Y για κάθε x και
  (ii) κανονική κατανομή της Y για κάθε x
- Για κάθε τιμή x_0 της X, η *πρόβλεψη* της y_0 από την ευθεία ελαχίστων τετραγώνων είναι

$$y_0 = a + bx_0$$

Προσοχή: Η τιμή x_0 πρέπει να ανήκει στο εύρος των γνωστών τιμών της X.

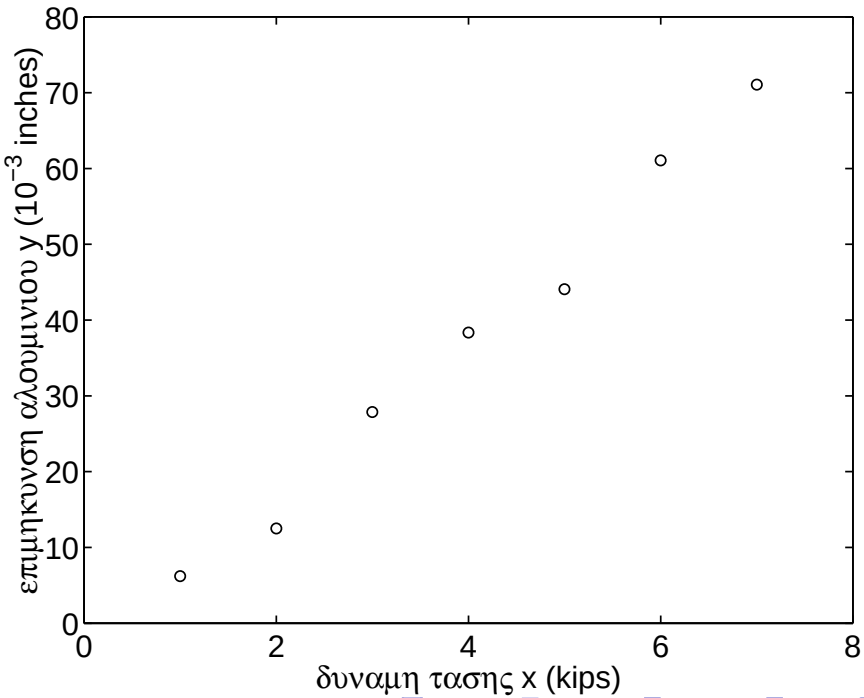
Παράδειγμα

Θέλουμε να μελετήσουμε την αντοχή αλουμινίου και για αυτό κάναμε ένα πείραμα και μετρήσαμε την επιμήκυνση δοκιμίου αλουμινίου για διάφορες τάσεις.

Δύναμη τάσης x_i (kips)	1	2	3	4	5	6	7
Επιμήκυνση y_i (10^{-3} inches)	6	13	28	38	44	61	71

Εξαρτάται η παραμόρφωση αλουμινίου από την τάση;

Είναι η εξάρτηση γραμμική;



Παράδειγμα (συνέχεια)

Υπολογίζουμε $\bar{x} = 4$ $\bar{y} = 37.31$

$\sum_{i=1}^{7} x_i^2 = 140$ $\sum_{i=1}^{7} y_i^2 = 13165$ $\sum_{i=1}^{7} x_i y_i = 1352.5$

$s_{XY} = 51.32$ $s_X^2 = 4.67$ $s_Y^2 = 570.50$

Οι εκτιμήσεις b και a

$$b = \frac{s_{XY}}{s_X^2} = \frac{51.32}{4.67} = 10.99$$

$$a = \bar{y} - b\bar{x} = 37.31 - 10.99 \cdot 4 = -6.65$$

Η εκτίμηση της διασποράς των σφαλμάτων παλινδρόμησης

$$s^2 = \frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2) = \frac{6}{5}(570.50 - 10.99^2 \cdot 4.67) = 7.75$$

Ευθεία ελαχίστων τετραγώνων: $y = -6.65 + 10.99x$
με διασπορά σφάλματος $s^2 = 7.75$

Παράδειγμα: Ερμηνεία των αποτελεσμάτων

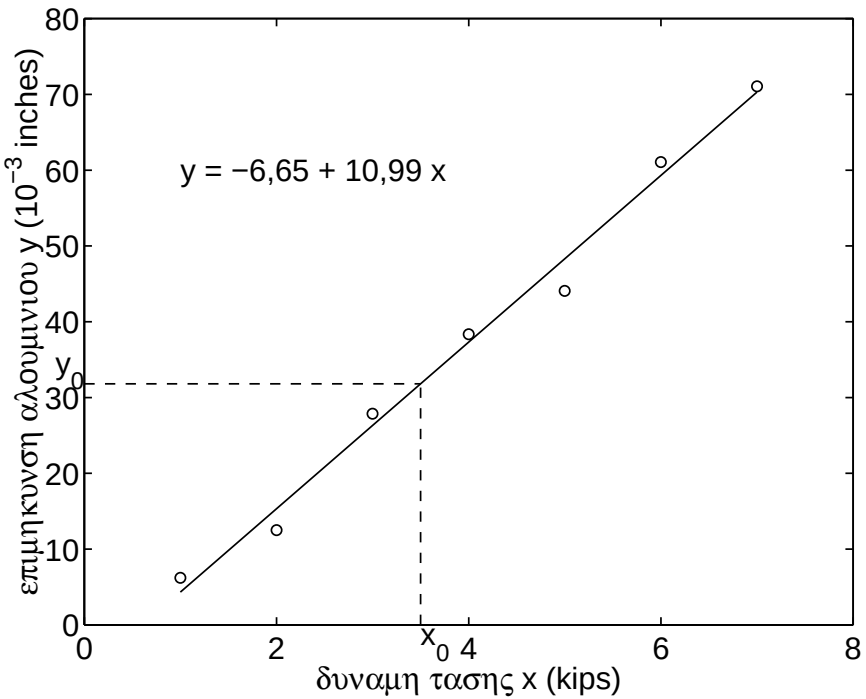
- b: Αύξηση δύναμης τάσης κατά 1 kips
  $\rightarrow$ επιμήκυνση κατά ~ 0.011 ίντσες $[10.99 \cdot 10^{-3}$ ίντσες$]$.
- a: Δύναμη τάσης 0
  $\rightarrow$ επιμήκυνση $-6.65 \cdot 10^{-3}$ ίντσες [αδύνατον]
- s^2: Τυπικό σφάλμα εκτίμησης παλινδρόμησης είναι
  $\sqrt{7.75} \rightarrow 2.78 \cdot 10^{-3}$ ίντσες [σχετικά μικρό]

Παράδειγμα: Πρόβλεψη

Με βάση το μοντέλο παλινδρόμησης μπορούμε να προβλέψουμε την επιμήκυνση του αλουμινίου για κάθε δύναμη τάσης στο διάστημα $[1, 7]$ kips:

$$x_0 = 3.5: \quad y_0 = -6.65 + 10.99 \cdot 3.5 = 31.82$$

με ακρίβεια πρόβλεψης (προσεγγιστικά) 31.82 ± 2.78



Σχέση r και b

Για το πρόβλημα της παλινδρόμησης, 'αγνοούμε' ότι η X δεν είναι τ.μ. και ορίζουμε το συντελεστή συσχέτισης ρ.

Σχέση μεταξύ του r και του b ($r = \frac{s_{XY}}{s_X s_Y}$ και $b = \frac{s_{XY}}{s_X^2}$)

$$r = b\frac{s_X}{s_Y} \quad \text{ή} \quad b = r\frac{s_Y}{s_X}$$

- r και b εκφράζουν ποιοτικά τη γραμμική συσχέτιση των X και Y
- b εξαρτάται από τη μονάδα μέτρησης των X και Y
- r παίρνει τιμές στο διάστημα $[-1, 1]$
- $r > 0 \Rightarrow b > 0 \qquad (r < 0 \Rightarrow b < 0)$
- $r = 0 \Rightarrow b = 0$

Σχέση r και s^2

Σχέση του r^2 και της διασποράς του σφάλματος s^2

$$s^2 = \frac{n-1}{n-2} s_Y^2 (1 - r^2) \quad \text{ή} \quad r^2 = 1 - \frac{n-2}{n-1} \frac{s^2}{s_Y^2}$$

Όσο μεγαλύτερο είναι το r^2 τόσο μικρότερο είναι το s^2 και καλύτερη η πρόβλεψη.

Συνέχεια παραδείγματος:

Συντελεστής συσχέτισης:

$$r = \frac{s_{XY}}{s_X \, s_Y} = \frac{41.32}{\sqrt{4.67 \cdot 570.50}} = 0.995$$

$r = 0.995$: πολύ ισχυρή θετική συσχέτιση της επιμήκυνσης του αλουμινίου και της τάσης

Θέμα 17

Διάστημα εμπιστοσύνης για τους συντελεστές της απλής γραμμικής παλινδρόμησης. Παρουσίαση και παράδειγμα.

Θέμα 18

Έλεγχος σημαντικότητας (significance test) για το συντελεστή της ανεξάρτητης μεταβλητής της απλής γραμμικής παλινδρόμησης. Παρουσίαση και παράδειγμα.

Θέμα 19

Διάστημα εμπιστοσύνης για τη μέση πρόβλεψη της απλής γραμμικής παλινδρόμησης. Παρουσίαση και παράδειγμα.

Θέμα 20

Πολλαπλή γραμμική παλινδρόμηση (multiple linear regression). Παρουσίαση και παράδειγμα.