Στατιστική για Ηλεκτρολόγους Μηχανικούς
ΣΥΣΧΕΤΙΣΗ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗ

Δημήτρης Κουγιουμτζής

20 Μαΐου 2010

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

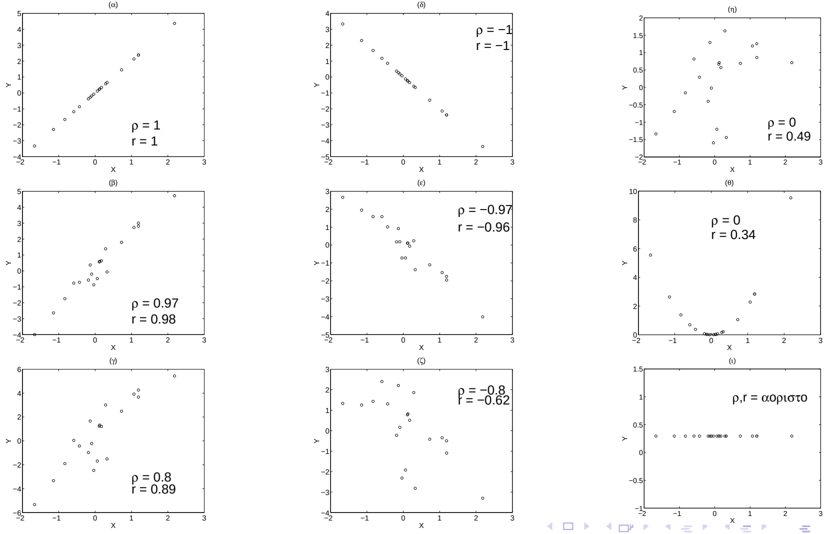
$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

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

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

Εκτίμηση του συντελεστή συσχέτισης

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

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

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 $100r^2\%$)

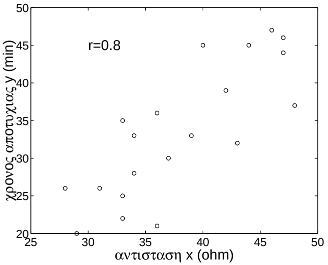
Δηλώνει το ποσοστό μεταβλητότητας που μπορούμε να ερμηνεύσουμε για τη μια τ.μ. όταν γνωρίζουμε την άλλη.

Παράδειγμα

Θέλουμε να εκτιμήσουμε τη συσχέτιση της αντίστασης και του χρόνου αποτυχίας κάποιου υπερφορτωμένου αντιστάτη

Δείγμα 20 δοκιμίων αντιστάσεων

Α/Α (i)	Αντίσταση x_i (ohm)	Χρόνος αποτυχίας y_i (min)
1	28	26
2	29	20
3	31	26
4	33	22
5	33	25
6	33	35
7	34	28
8	34	33
9	36	21
10	36	36
11	37	30
12	39	33
13	40	45
14	42	39
15	43	32
16	44	45
17	46	47
18	47	44
19	47	46
20	48	37



Υπολογίζουμε

$\bar{x} = 38 \qquad \bar{y} = 33.5$

$\sum_{i=1}^{20} x_i^2 = 29634$

$\sum_{i=1}^{20} y_i^2 = 23910$

$\sum_{i=1}^{20} x_i y_i = 26305$

$$r = \frac{26305 - 20 \cdot 38 \cdot 33.5}{\sqrt{(29634 - 20 \cdot 38^2) \cdot (23910 - 20 \cdot 33.5^2)}} = 0.804$$

Η αντίσταση και ο χρόνος αποτυχίας αντιστάτη έχουν γραμμική θετική συσχέτιση αλλά όχι ισχυρή.

Το ποσοστό μεταβλητότητας της αντίστασης που μπορούμε να εξηγήσουμε γνωρίζοντας τον χρόνο αποτυχίας αντιστάτη (και αντίστροφα) είναι 0.646.

Απλή Γραμμική Παλινδρόμηση

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

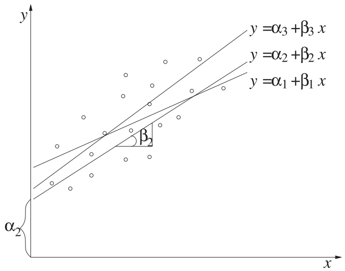
Y: εξαρτημένη μεταβλητή (τυχαία)
X: ανεξάρτητη μεταβλητή (καθορισμένη)

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 $\mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2$ για κάθε x_i

$$\mathrm{Var}(y_i | X = x_i) = \mathrm{Var}(\alpha + \beta x_i + \epsilon_i) = \mathrm{Var}(\epsilon_i)$$

$$\Downarrow$$

$$\mathrm{Var}(Y|X = x) \equiv \sigma_{Y|X}^2 = \sigma_\epsilon^2 \equiv \sigma^2$$

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:

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

Εκτιμήσεις των β και α είναι

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

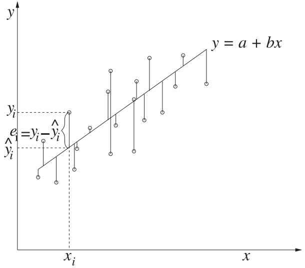
ευθεία ελαχίστων τετραγώνων: $\hat{y} = a + bx$

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 $(\bar{x}, \bar{y})$: $\quad a + b\bar{x} = \bar{y} - b\bar{x} + b\bar{x} = \bar{y}$
  Άρα η ευθεία ελαχίστων τετραγώνων μπορεί επίσης να οριστεί ως $\quad y_i - \bar{y} = b(x_i - \bar{x})$
- 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.

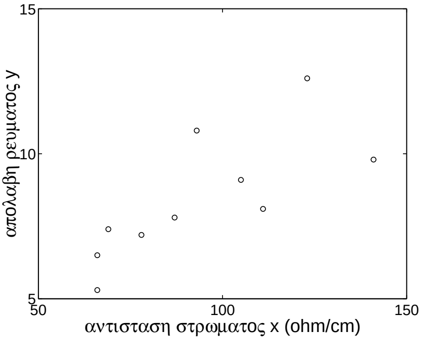
Παράδειγμα

Θέλουμε να μελετήσουμε σ' ένα ολοκληρωμένο κύκλωμα την εξάρτηση της απολαβής ρεύματος κρυσταλλολυχνίας (τρανζίστορ) από την αντίσταση του στρώματος της κρυσταλλολυχνίας.

Α/Α (i)	Αντίσταση στρώματος x_i (ohm/cm)	Απολαβή ρεύματος y_i
1	66	5.3
2	66	6.5
3	69	7.4
4	78	7.2
5	87	7.8
6	93	10.8
7	105	9.1
8	111	8.1
9	123	12.6
10	141	9.8

Εξαρτάται η απολαβή ρεύματος τρανζίστορ από την αντίσταση του στρώματος κρυσταλλολυχνίας;

Είναι η εξάρτηση γραμμική;



Παράδειγμα (συνέχεια)

Υπολογίζουμε $\bar{x} = 93.9$ $\bar{y} = 8.46$

$\sum_{i=1}^{10} x_i^2 = 94131$ $\sum_{i=1}^{10} y_i^2 = 757.64$ $\sum_{i=1}^{10} x_i y_i = 8320.2$

$s_{XY} = 41.81$ $s_X^2 = 662.1$ $s_Y^2 = 4.66$

Οι εκτιμήσεις b και a

$$\begin{aligned} b &= \frac{s_{XY}}{s_X^2} = \frac{41.81}{662.1} = 0.063 \\ a &= \bar{y} - b\bar{x} = 8.46 - 0.063 \cdot 93.9 = 2.53 \end{aligned}$$

Η εκτίμηση της διασποράς των σφαλμάτων παλινδρόμησης

$$s^2 = \frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2) = \frac{9}{8}(4.66 - 0.063^2 \cdot 41.81) = 5.056$$

Ευθεία ελαχίστων τετραγώνων: $y = 2.53 + 0.063x$
με διασπορά σφάλματος $s^2 = 5.056$

Παράδειγμα: Ερμηνεία των αποτελεσμάτων

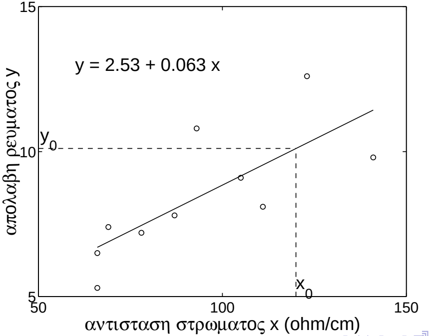
- b: Αύξηση αντίστασης στρώματος κατά μία μονάδα μέτρησης (1 ohm/cm)
  $\longrightarrow$ απολαβή του ρεύματος της κρυσταλλολυχνίας αυξάνεται κατά 0.063.
- a: Αντίσταση στρώματος 0
  $\longrightarrow$ απολαβή ρεύματος 2.53 [αδύνατον]
- s^2: Τυπικό σφάλμα εκτίμησης της παλινδρόμησης είναι $\sqrt{5.056} \rightarrow 2.249$ [σχετικά μεγάλο]

Παράδειγμα: Πρόβλεψη

Με βάση το μοντέλο παλινδρόμησης μπορούμε να προβλέψουμε την απολαβή ρεύματος για κάθε αντίσταση στρώματος κρυσταλλολυχνίας στο διάστημα $[66, 141]\,\mathrm{ohm/cm}$:

$$x_0 = 120: \quad y_0 = 2.53 + 0.063 \cdot 120 = 10.11$$

με ακρίβεια πρόβλεψης (προσεγγιστικά) 10.11 ± 2.249



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:

$$r = \frac{s_{XY}}{s_X \, s_Y} = \frac{41.81}{\sqrt{662.1 \cdot 4.66}} = 0.753$$

$r = 0.753$: ασθενής θετική συσχέτιση της απολαβής ρεύματος και αντίστασης στρώματος κρυσταλλολυχνίας