Στατιστική για Χημικούς Μηχανικούς ΣΥΣΧΕΤΙΣΗ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗ

Δημήτρης Κουγιουμτζής

14 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

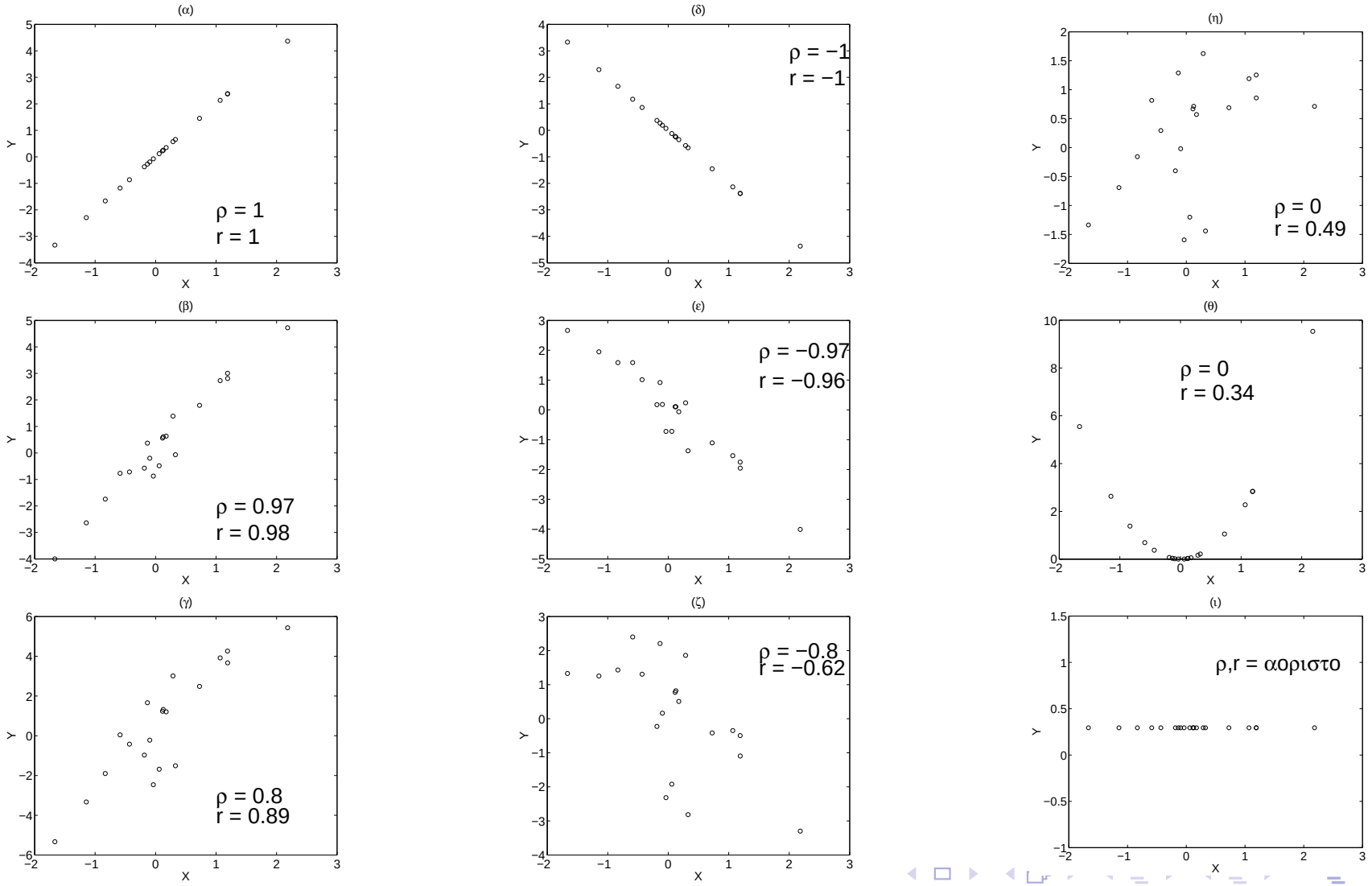
$$\rho = \frac{\sigma_{XY}}{\sigma_X \, \sigma_Y}$$

Ιδιότητες του ρ

- $\rho \in [-1, 1]$
- $\rho = 1$: *τέλεια θετική* συσχέτιση
- $\rho = -1$: *τέλεια αρνητική* συσχέτιση
- ρ 'κοντά' στο -1 ή 1 $\longrightarrow$ ισχυρή συσχέτιση
- ρ 'κοντά' στο 0 $\longrightarrow$ 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 $100r^2\%$)

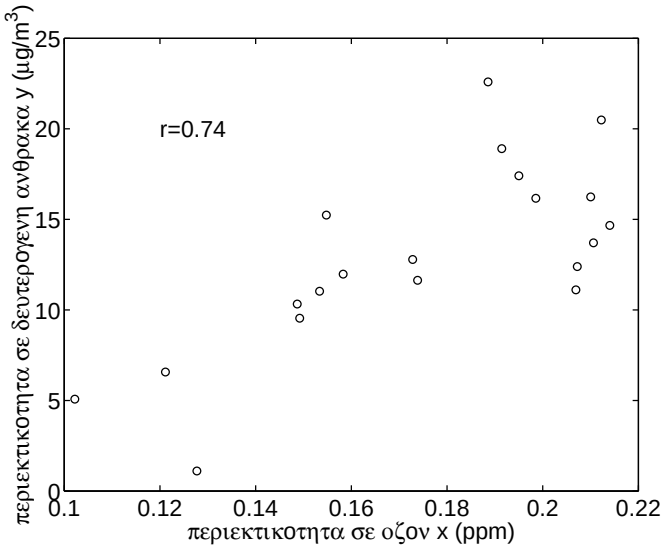
Δηλώνει το ποσοστό μεταβλητότητας που μπορούμε να ερμηνεύσουμε για τη μια τ.μ. όταν γνωρίζουμε την άλλη.

Παράδειγμα

Θέλουμε να εκτιμήσουμε τη συσχέτιση της περιεκτικότητας σε όζον και σε δευτερογενή άνθρακα στον αέρα κάποιας περιοχής.

Δείγμα από 20 μετρήσεις

A/A	Περιεκτικότητα σε όζον x_i (ppm)	Περιεκτικότητα σε δευτερογενή άνθρακα y_i ($\mu g/m^3$)
1	0.102	5.07
2	0.121	6.57
3	0.128	1.11
4	0.149	10.32
5	0.149	9.54
6	0.153	11.03
7	0.155	15.23
8	0.158	11.98
9	0.173	12.78
10	0.174	11.64
11	0.189	22.59
12	0.191	18.91
13	0.195	17.41
14	0.199	16.17
15	0.207	11.11
16	0.207	12.39
17	0.210	16.24
18	0.211	13.71
19	0.212	20.49
20	0.214	14.67



Υπολογίζουμε

$\bar{x} = 0.175 \quad \bar{y} = 12.95$

$\sum_{i=1}^{20} x_i^2 = 0.633$

$\sum_{i=1}^{20} y_i^2 = 3860.17$

$\sum_{i=1}^{20} x_i y_i = 247.74$

$$r = \frac{47.74 - 20 \cdot 0.175 \cdot 12.95}{\sqrt{(0.633 - 20 \cdot 0.175^2)\,(3860.17 - 20 \cdot 12.95^2)}} = 0.74$$

Περιεκτικότητα σε όζον και περιεκτικότητα σε δευτερογενή άνθρακα έχουν γραμμική θετική συσχέτιση αλλά όχι ισχυρή.

Το ποσοστό μεταβλητότητας της περιεκτικότητας σε δευτερογενή άνθρακα που μπορούμε να εξηγήσουμε γνωρίζοντας την περιεκτικότητα σε όζον (και αντίστροφα) είναι 0.55.

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)

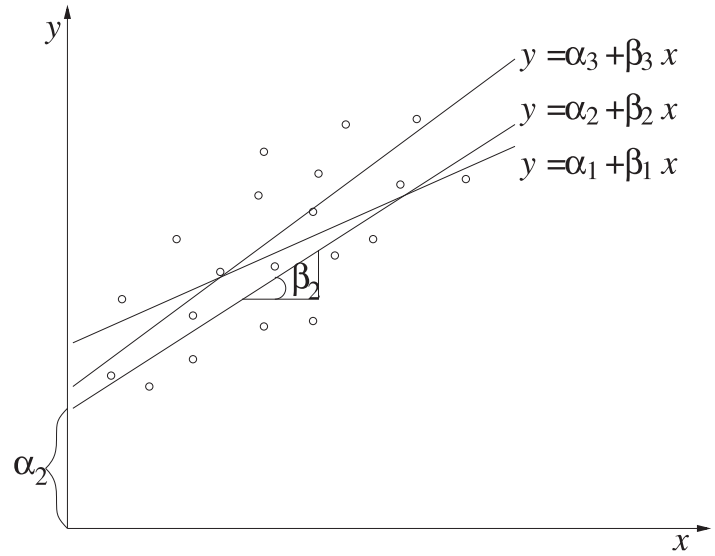
Παράδειγμα: διατμητική αντοχή αργίλου σε διάφορα βάθη
X ? Y ? ;

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

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

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

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

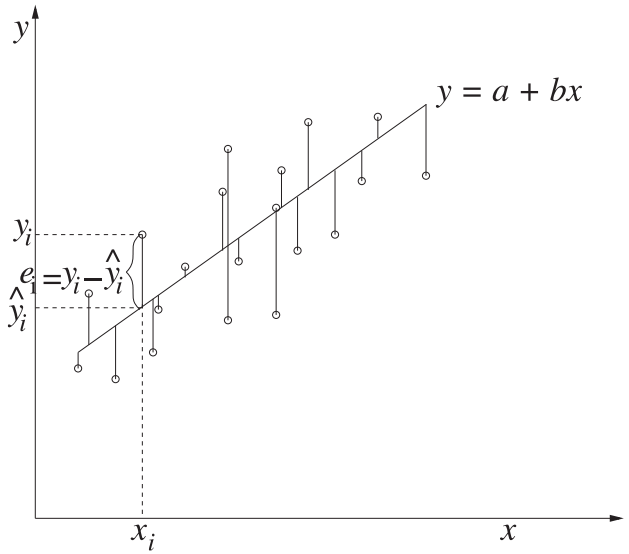
ευθεία ελαχίστων τετραγώνων: $\hat{y} = a + bx$

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

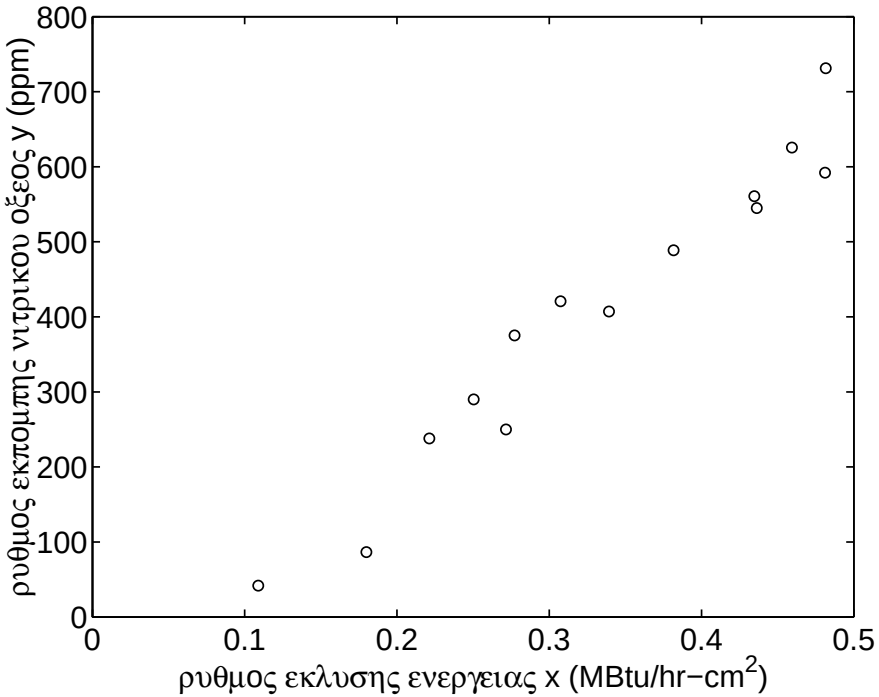
Έγινε ένα πείραμα σε λέβητες παραγωγής θερμότητας για τη μελέτη της εξάρτησης της παραγωγής νιτρικού οξέος από το 'ρυθμό έκλυσης σε επιφάνεια καύσης' (burner area liberation rate) (είναι ένα μέτρο της ενέργειας που παράγεται από τη μονάδα ανά τετραγωνικό εκατοστό της επιφάνειας του καυστήρα)

Α/Α	Ρυθμός έκλυσης από καυστήρα x_i (MBtu/hr-cm^2)	Ρυθμός εκπομπής νιτρικού οξέος y_i (ppm)
1	0.109	41.7
2	0.180	86.5
3	0.221	238.1
4	0.250	289.9
5	0.272	250.1
6	0.277	375.2
7	0.307	420.7
8	0.339	407.0
9	0.382	488.8
10	0.435	560.8
11	0.436	545.1
12	0.459	625.4
13	0.481	592.3
14	0.482	731.5

Παράδειγμα (συνέχεια)

Εξαρτάται η παραγωγή νιτρικού οξέος από το ρυθμό έκλυσης σε επιφάνεια καύσης;

Είναι η εξάρτηση γραμμική;



Παράδειγμα (συνέχεια)

Υπολογίζουμε $\bar{x} = 0.331$ $\bar{y} = 403.8$

$\sum_{i=1}^{14} x_i^2 = 1.716$ $\sum_{i=1}^{14} y_i^2 = 2823556.9$ $\sum_{i=1}^{14} x_i y_i = 2177.42$

$s_X^2 = 0.014$ $s_Y^2 = 41603.9$ $s_{XY} = 23.66$

Οι εκτιμήσεις b και a

$$b = \frac{s_{XY}}{s_X^2} = \frac{23.66}{0.014} = 1672.85$$

$$a = \bar{y} - b\bar{x} = 403.8 - 1672.85 \cdot 0.331 = -149.50$$

Η εκτίμηση της διασποράς των σφαλμάτων παλινδρόμησης

$$s^2 = \frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2) = \frac{13}{12}(41603.9 - 1672.85^2 \cdot 0.014) = 2186.0$$

Ευθεία ελαχίστων τετραγώνων: $y = -149.5 + 1672.85x$
με διασπορά σφάλματος $s^2 = 2186$

Παράδειγμα: Ερμηνεία των αποτελεσμάτων

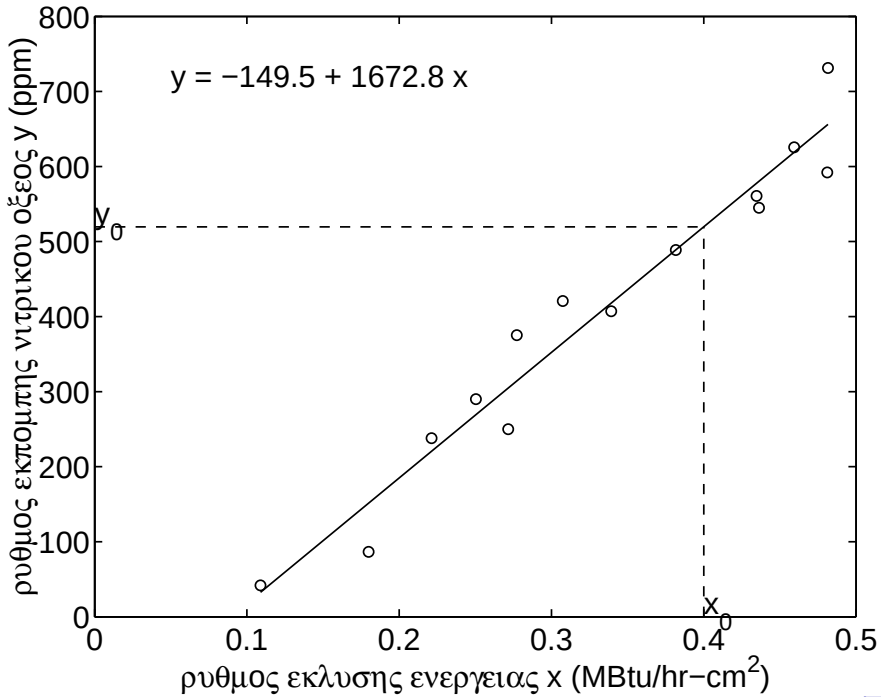
- b: Αύξηση του ρυθμού έκλυσης ενέργειας κατά μία μονάδα μέτρησης ($1\,\text{MBtu/hr-cm}^2$)
  $\longrightarrow$ αύξηση ρυθμού εκπομπής νιτρικού οξέος κατά $1672.85\,\text{ppm}$
- a: Μηδενική έκλυση ενέργειας από την επιφάνεια του καυστήρα ($x = 0$)
  $\longrightarrow$ εκπομπή νιτρικού οξέος $-149.5\,\text{ppm}$ [αδύνατον]
- s^2: Τυπικό σφάλμα εκτίμησης της παλινδρόμησης είναι $\sqrt{2186.0} \longrightarrow 46.75\text{ppm}$ [σχετικά μικρό]

Παράδειγμα: Πρόβλεψη

Με βάση το μοντέλο παλινδρόμησης μπορούμε να προβλέψουμε το ρυθμό εκπομπής νιτρικού οξέος για κάθε τιμή του ρυθμού έκλυσης ενέργειας στο διάστημα $[0.1, 0.5]$ ppm (προσεγγιστικά):

$$x_0 = 0.4: \quad y_0 = -149.5 + 1672.85 \cdot 0.4 = 519.6$$

με ακρίβεια πρόβλεψης (προσεγγιστικά) 519.6 ± 46.75



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:

$$r = \frac{s_{XY}}{s_X \, s_Y} = \frac{23.66}{\sqrt{0.014 \cdot 41603.9}} = 0.975$$

$r = 0.975$: ισχυρή θετική συσχέτιση του ρυθμού της εκπομπής νιτρικού οξέος και και του ρυθμού έκλυσης ενέργειας