# ΣΥΣΧΕΤΙΣΗ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗ

**Παραδείγματα**

συνολική επιφάνεια κτιρίου ~
επιφάνεια που καλύπτεται από παράθυρα

παλαιότητα κτιρίου ~
απώλεια θερμικής ενέργειας

κατανάλωση ηλεκτρικής ενέργειας κατοικίας ~
κατανάλωση νερού ~
μέγεθος κατοικίας

Εξαρτάται η μια τ.μ. από την άλλη?

Εξαρτιούνται και οι δύο από κάποια άλλη?

Δύο τ.μ.: $X$ με διασπορά $\sigma_X^2$ , $Y$ με $\sigma_Y^2$

**συνδιασπορά** $\sigma_{XY} \equiv \mathrm{Cov}[X,Y]$

$$\sigma_{XY} = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[X\,Y] - \mu_X \mu_Y$$

- εκφράζει τη γραμμική συσχέτιση δύο τ.μ. δηλαδή την αναλογική μεταβολή (αύξηση ή μείωση) της μιας τ.μ. που αντιστοιχεί σε μεταβολή της άλλης μεταβλητής
- εξαρτάται από τις μονάδες μέτρησης των δύο τ.μ.

***συντελεστής συσχέτισης*** $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$

- $\rho$ δεν εξαρτάται από τη μονάδα μέτρησης των $X$ και $Y$
- $\rho$ είναι συμμετρικός ως προς τις $X$ και $Y$.
- $\rho \in$ [-1,1]

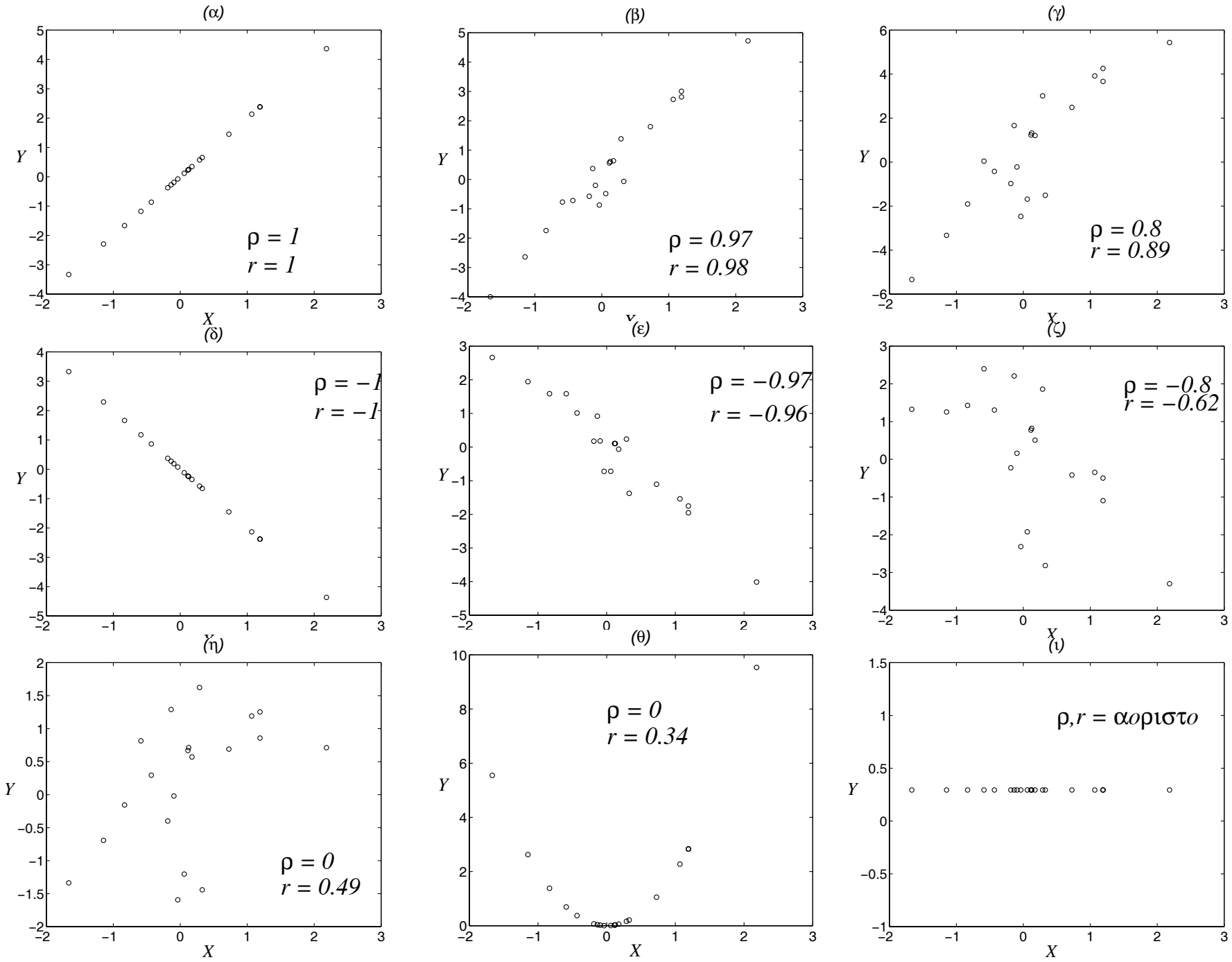
## Ερμηνεία της τιμής του συντελεστή συσχέτισης

- $\rho=1$: υπάρχει *τέλεια θετική* σχέση μεταξύ των *X* και *Y*.
- $\rho=0$: δεν υπάρχει καμιά (γραμμική) σχέση μεταξύ των *X* και *Y*.
- $\rho=-1$: υπάρχει *τέλεια αρνητική* σχέση μεταξύ των *X* και *Y*.
- "*ρ κοντά στο 1* "
  ➔ η γραμμική συσχέτιση των δύο τ.μ. είναι θετική και ισχυρή
- "*ρ κοντά στο -1* "
  ➔ η γραμμική συσχέτιση των δύο τ.μ. είναι αρνητική και ισχυρή
- "*ρ κοντά στο 0* " ➔ οι τ.μ. είναι πρακτικά ασυσχέτιστες

Παρατηρήσεις των δύο τ.μ. $X$ και $Y$ κατά ζεύγη

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

για ποιοτική εκτίμηση της συσχέτισης ➔ ***διάγραμμα διασποράς***


(α)
ρ = 1
r = 1
(β)
ρ = 0.97
r = 0.98
(γ)
ρ = 0.8
r = 0.89
(δ)
ρ = −1
r = −1
(ε)
ρ = −0.97
r = −0.96
(ζ)
ρ = −0.8
r = −0.62
(η)
ρ = 0
r = 0.49
(θ)
ρ = 0
r = 0.34
(ι)
ρ,r = αορισταο
X
Y

## Σημειακή εκτίμηση του συντελεστή συσχέτισης

Σημειακή εκτίμηση του $\rho$ από το δείγμα των $n$ ζευγαρωτών παρατηρήσεων των $X$ και $Y$

$$\hat{\rho} \equiv r = \frac{s_{XY}}{s_X s_Y}$$

αμερόληπτη εκτιμήτρια $s_{XY}$

$$\sigma_{XY} = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[XY] - \mu_X \mu_Y$$

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

αμερόληπτες εκτιμήτριες $s_X$ και $s_Y$
είναι οι τετραγωνικές ρίζες των δειγματικών διασπορών

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

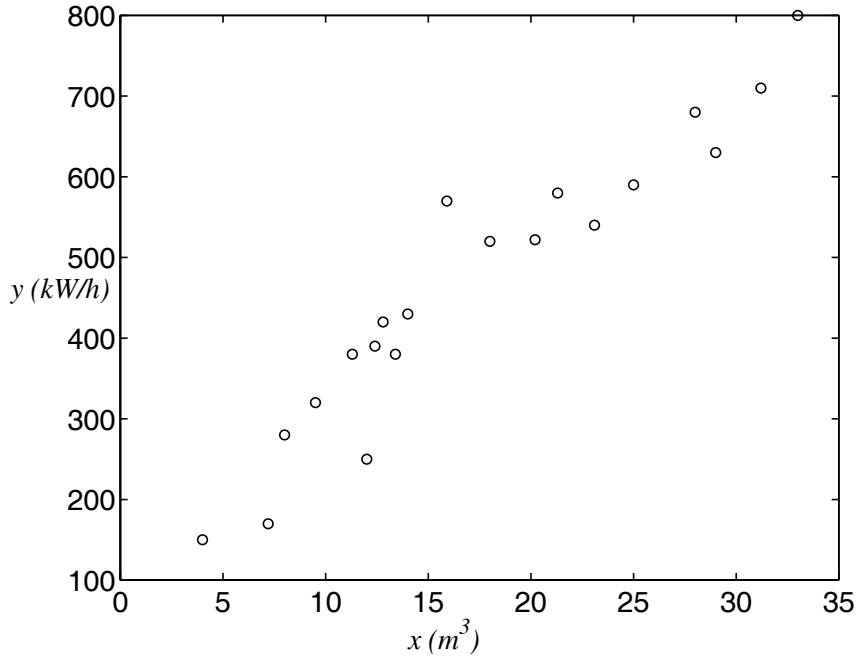
Άρα

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

***συντελεστής προσδιορισμού***
$r^2$ ή $100r^2$

## Παράδειγμα

Θέλουμε να διερευνήσουμε τη συσχέτιση της κατανάλωσης νερού και κατανάλωσης ρεύματος νοικοκυριού.




<u>Ποιοτική εκτίμηση</u>

συσχέτιση της κατανάλωσης νερού και ρεύματος είναι θετική και ισχυρή

| $A/A$ | $Νερό\ x_i$ ($m^3$) | $Ρεύμα\ y_i$ (kW/h) |
|---|---|---|
| 1 | 4.0 | 150 |
| 2 | 7.2 | 170 |
| 3 | 8.0 | 280 |
| 4 | 9.5 | 320 |
| 5 | 11.3 | 380 |
| 6 | 12.0 | 250 |
| 7 | 12.4 | 390 |
| 8 | 12.8 | 420 |
| 9 | 13.4 | 380 |
| 10 | 14.0 | 430 |
| 11 | 15.9 | 570 |
| 12 | 18.0 | 520 |
| 13 | 20.2 | 522 |
| 14 | 21.3 | 580 |
| 15 | 23.1 | 540 |
| 16 | 25.0 | 590 |
| 17 | 28.0 | 680 |
| 18 | 29.0 | 630 |
| 19 | 31.2 | 710 |
| 20 | 33.0 | 800 |

Υπολογισμός του $r$

$$\bar{x} = 17.465 \qquad \bar{y} = 465.6$$

$$\sum_{i=1}^{20} x_i^2 = 7471.53 \qquad \sum_{i=1}^{20} y_i^2 = 4944184 \qquad \sum_{i=1}^{20} x_i y_i = 190129.4$$

$$r = \frac{190129.4 - 20 \cdot 17.465 \cdot 465.6}{\sqrt{\left(7471.53 - 20 \cdot 17.465^2\right) \cdot \left(4944184 - 20 \cdot 465.6^2\right)}} = 0.952$$

Η μεταβλητότητα της μιας τ.μ. (κατανάλωση νερού ή ρεύματος) μπορεί να εξηγηθεί σε μεγάλο ποσοστό από τη συσχέτιση της με την άλλη

Συντελεστής προσδιορισμού $100r^2 = 100 \cdot 0.952^2 = 90.6\%$

Συμπέρασμα

Η γνώση της μιας τ.μ. μας επιτρέπει να προσδιορίσουμε την άλλη με μεγάλη ακρίβεια

# Απλή Γραμμική Παλινδρόμηση

*Επίδραση του μεγέθους της κατοικίας στην κατανάλωση νερού του νοικοκυριού?*

Ζητάμε να εκτιμήσουμε την (γραμμική) εξάρτηση της κατανάλωσης νερού από το μέγεθος της κατοικίας.

**εξαρτημένη μεταβλητή** $Y$: κατανάλωση νερού

**ανεξάρτητη μεταβλητή** $X$: μέγεθος κατοικίας

Μελέτη της μεταβλητότητα μιας τ.μ. *Y* χρησιμοποιώντας την πληροφορία από κάποια άλλη μεταβλητή *X*

➔ ***ανάλυση παλινδρόμησης***

***απλή γραμμική παλινδρόμηση***

απλή: σχέση εξάρτησης μόνο ως προς μια ανεξάρτητη μεταβλητή

γραμμική: η πιο απλή σχέση εξάρτησης

$F_Y(y \mid X = x)$ για κάθε τιμή $x$ της $X$?

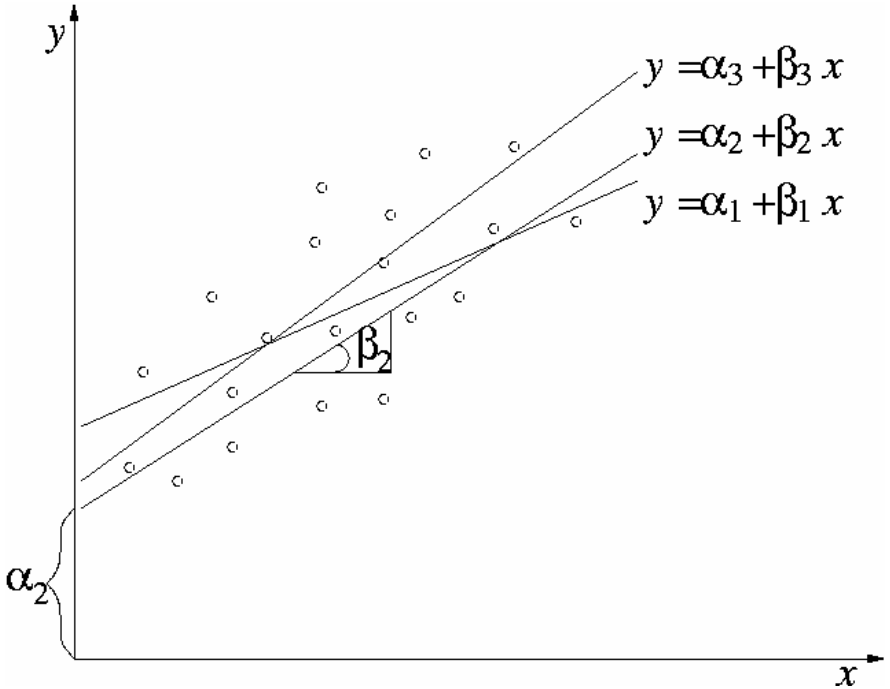
$\mathrm{E}[Y \mid X = x]$ για κάθε τιμή $x$ της $X$?

$$\mathrm{E}[Y \mid X = x] = \alpha + \beta\, x$$

***γραμμική παλινδρόμηση της Y στη X***

$\alpha$: ***διαφορά ύψους***, σταθερός όρος, η τιμή του $y$ για $x=0$

$\beta$: ***κλίση*** ή ***συντελεστής παλινδρόμησης***




για κάποια τιμή $x_i$ της $X$ μπορεί να αντιστοιχούν διαφορετικές τιμές $y_i$ της $Y$ ➔ $y_i$ είναι τ.μ. με $F_Y(y_i \mid X = x_i)$

$$y_i = \alpha + \beta\, x_i + \varepsilon_i$$

$\varepsilon_i$: ***σφάλμα παλινδρόμησης***

$$\varepsilon_i = y_i - \mathrm{E}[Y \mid X = x_i]$$

*Υποθέσεις για την ανάλυση της γραμμικής παλινδρόμησης*:

- Η μεταβλητή $X$ είναι *ελεγχόμενη*
- Η εξάρτηση της $Y$ από τη $X$ είναι γραμμική
- $\mathrm{E}[\varepsilon_i] = 0$ και $\mathrm{Var}[\varepsilon_i] = \sigma_\varepsilon^2$ ➔ $\mathrm{Var}[Y \mid X = x] \equiv \sigma_{Y|X}^2 = \sigma^2$

$\sigma_{X|Y}^2 = \sigma_\varepsilon^2 = \sigma^2$ (*ομοσκεδαστικότητα*)

**Σημειακή εκτίμηση παραμέτρων γραμμικής παλινδρόμησης**

Εκτίμηση των τριών παραμέτρων της παλινδρόμησης:

- της διαφοράς ύψους της ευθείας παλινδρόμησης $\alpha$
- του συντελεστή της ευθείας παλινδρόμησης $\beta$
- της διασποράς σφάλματος της παλινδρόμησης $\sigma^2$

## Εκτίμηση των $\alpha$ και $\beta$

Η εκτίμηση των παραμέτρων *α* και *β* γίνεται με τη μέθοδο των ***ελαχίστων τετραγώνων***
βρίσκει την ευθεία παλινδρόμησης με παραμέτρους *α* και *β* έτσι ώστε το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων από την ευθεία να είναι το ελάχιστο.

Οι εκτιμήσεις των $\alpha$ και $\beta$ δίνονται από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων

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

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

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

$s_{XY}$
δειγματική συνδιασπορά των $X$ και $Y$

$s_X^2$
δειγματική διασπορά της $X$

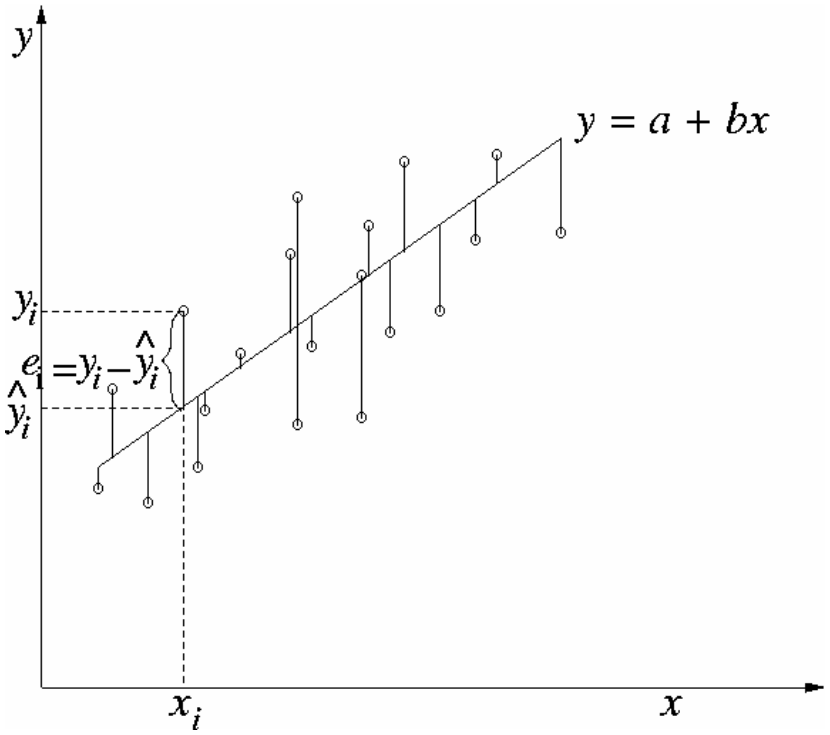
Τα $a$ και $b$ ορίζουν την ευθεία

$$\hat{y} = a + b\,x$$ *ευθεία ελαχίστων τετραγώνων*

Για κάθε $x_i$ ➔ $\hat{y}_i = a + b\,x_i$

***υπόλοιπο*** ή σφάλμα ελαχίστων τετραγώνων

$$e_i = y_i - \hat{y}_i = y_i - a - b\,x_i$$




$$\varepsilon_i = y_i - \alpha - \beta\,x_i$$

$e_i$ είναι η εκτίμηση του $\varepsilon_i$,

Εκτίμηση της διασποράς του σφάλματος $\sigma^2$

δειγματική διασπορά $s^2$ των υπολοίπων $e_i$

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

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

## Παρατηρήσεις

- Η ευθεία ελαχίστων τετραγώνων περνάει από το σημείο $(\bar{x}, \bar{y})$

  Άρα η ευθεία ελαχίστων τετραγώνων μπορεί επίσης να οριστεί ως

$$y_i - \bar{y} = b(x_i - \bar{x})$$

- Για κάθε τιμή $x_0$ της $X$ μπορούμε να ***προβλέψουμε*** την αντίστοιχη τιμή $y_0$ της $Y$ από την ευθεία ελαχίστων τετραγώνων ➔

$$\hat{y}_0 = a + b\,x_0$$

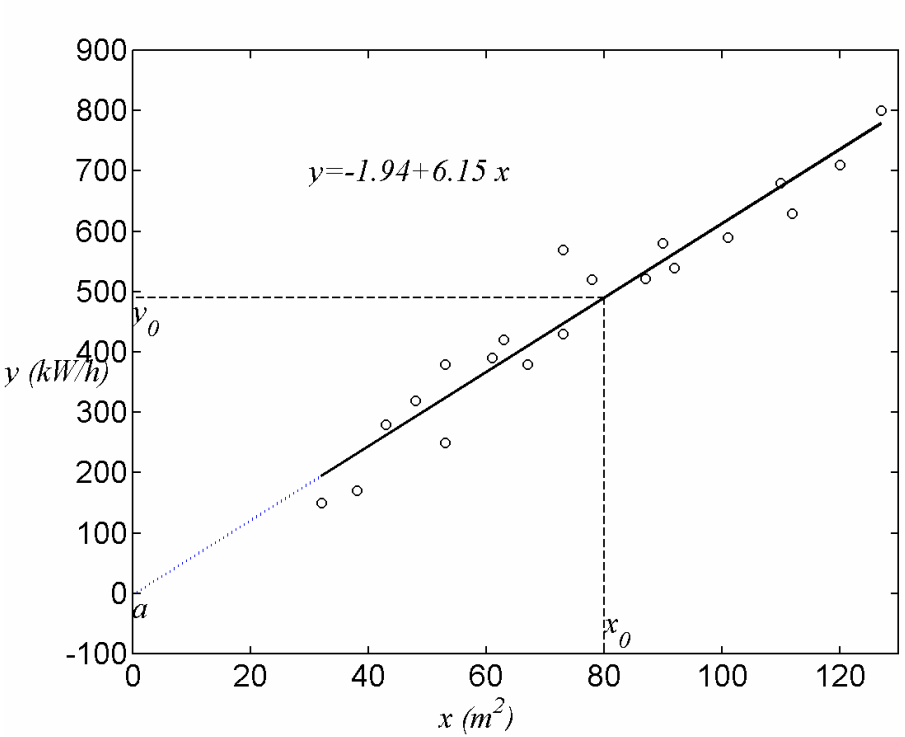
Προσοχή:
$x_0$ πρέπει να ανήκει στο εύρος των τιμών της $X$ που έχουμε από το δείγμα

## Παράδειγμα

Θέλουμε να μελετήσουμε τη μεταβολή της κατανάλωσης ρεύματος με το μέγεθος του διαμερίσματος για να προβλέψουμε, αν είναι δυνατόν, την κατανάλωση ρεύματος για κάποιο διαμέρισμα

*Υποθέτουμε* πως η κατανάλωση ηλεκτρικού ρεύματος για ένα διαμέρισμα εξαρτάται γραμμικά από το μέγεθος του διαμερίσματος.

Σωστή υπόθεση?




| Α/Α | Μέγεθος $x_i$ ($m^3$) | Ρεύμα $y_i$ (kW/h) |
|---|---|---|
| 1 | 32 | 150 |
| 2 | 38 | 170 |
| 3 | 43 | 280 |
| 4 | 48 | 320 |
| 5 | 53 | 380 |
| 6 | 53 | 250 |
| 7 | 61 | 390 |
| 8 | 63 | 420 |
| 9 | 67 | 380 |
| 10 | 73 | 430 |
| 11 | 73 | 570 |
| 12 | 78 | 520 |
| 13 | 87 | 522 |
| 14 | 90 | 580 |
| 15 | 92 | 540 |
| 16 | 101 | 590 |
| 17 | 110 | 680 |
| 18 | 112 | 630 |
| 19 | 120 | 710 |
| 20 | 127 | 800 |

Εκτίμηση των παραμέτρων $a$ και $b$ της ευθείας ελαχίστων τετραγώνων

$$\bar{x} = 76.05 \qquad \bar{y} = 465.6$$

$$\sum_{i=1}^{20} x_i^2 = 130667 \qquad \sum_{i=1}^{20} y_i^2 = 4944184 \qquad \sum_{i=1}^{20} x_i y_i = 800364$$

$$s_X^2 = 789.21 \qquad s_Y^2 = 32027.2 \qquad s_{XY} = 4851.92$$

$$b = \frac{4851.92}{789.21} = 6.148 \qquad a = 465.6 - 6.148 \cdot 76.05 = -1.94$$

Εκτίμηση διασποράς των σφαλμάτων παλινδρόμησης

$$s^2 = \frac{19}{18} \cdot \left(32027.2 - 6.148^2 \cdot 789.21\right) = 2318.8$$

## Ερμηνεία των αποτελεσμάτων:

$b$: Για αύξηση του μεγέθους του διαμερίσματος κατά $1m^2$ η κατανάλωση ηλεκτρικού ρεύματος αυξάνει κατά περίπου 6kW/h (για ακρίβεια 6.148kW/h)

$a$: Για μηδενικό μέγεθος διαμερίσματος η κατανάλωση ρεύματος είναι -1.94kW/h (???)

$s^2$: Η εκτίμηση της διασποράς γύρω από την ευθεία παλινδρόμησης για κάθε τιμή του $X$ (που ανήκει στο διάστημα τιμών του πειράματος) είναι $2318.8(kW/h)^2$
➔ Το τυπικό σφάλμα της εκτίμησης της παλινδρόμησης είναι 48.15kW/h ➔

η τυπική κατανάλωση ρεύματος για κάποιο διαμέρισμα μεγέθους $x_i$ βρίσκεται στο διάστημα

$$\hat{y}_i \pm 48.15$$

όπου $\hat{y}_i$ είναι η εκτίμηση από την ευθεία ελαχίστων τετραγώνων της μέσης κατανάλωσης ρεύματος για διαμέρισμα μεγέθους $x_i$

Συμπέρασμα:

Δε μπορούμε να βγάλουμε συμπεράσματα για την κατανάλωση του ρεύματος σε μεγέθη διαμερισμάτων εκτός του εύρους των τιμών του δείγματος

Με βάση το μοντέλο παλινδρόμησης που εκτιμήσαμε μπορούμε να προβλέψουμε κατανάλωση ρεύματος για κάθε διαμέρισμα μεγέθους από 32m$^2$ ως 127m$^2$.

Για διαμέρισμα 80m$^2$ ( $x_0 = 80$ ) ➔

$$y_0 = -1.94 + 6.148 \cdot 80 = 489.9$$

η κατανάλωση ρεύματος είναι 489.9 kW/h