23-2-2023

Μοντέλα Παλινδρόμησης

Y : εξαρτημένη μεταβλητή / απόκριση

dependent variable / response variable

$X_1, X_2, \dots, X_p$ ανεξάρτητες μεταβλητές / παράγοντες

independent variables / predictors / factors

Y: ποσοτική (scale)

$X_1, X_2, \dots, X_p$: (αρχικά) όλες ποσοτικές

(θα το γενικεύσουμε)

Στατιστικό Μοντέλο

συνάρτηση παλινδρόμησης

$$E(Y \mid X_1 = x_1, X_2 = x_2, \dots, X_p = x_p) = f(x_1, x_2, \dots, x_p)$$

regression function

(αγνώστη . . . --)

Παράδειγμα

Y = score

X_1 = age

X_2 = χρόνια άσκησης

$$f(52, 120) = E(Y \mid X_1 = 52, X_2 = 120) =$$

= μέσο ύψος ατόμων ηλικίας 52 ετών που ασχολούνται 120 min/εβδ.

Παραμετρικά

1) Αν $f(x_1, \dots, x_p) = C$ = σταθερό

$\nexists$ συσχέτιση της Y με αυτές τις ανεξ. μεταβλητές

2) f : άγνωστη συνάρτηση, πρέπει να "εκτιμηθεί" από δεδομένα

π.χ. 1) $f(x_1, x_2) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$

2) $f(x_1, x_2) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1^2 + \alpha_4 x_2^2 + \alpha_5 x_1 x_2$

$\alpha_0, \alpha_1, \dots$ (άγνωστες παράμετροι που πρέπει να εκτιμηθούν)

3) $f(x_1, x_2, \dots, x_p) = C$

$E(Y) = C$ (άγνωστο)

↳ εκτίμηση του $E(Y)$ για ό,τι πληθυσμό

(εκτίμηση ου μέσης τιμής)

4) $f(x) = a_0 + a_1 \sin(2\pi f x)$

a_0, a_1, f : άγνωστες παράμετροι

ή μόνο α_0, α_1 : άγνωστα ε' αν f γνωστή.

6) $f(x) = \alpha_0 e^{\alpha_1 x_1 + \alpha_2 x_2}$

Γραμμικά Μοντέλα

Γενική μορφή:

$$E(Y) = b_0 + b_1 x_1 + \cdots + b_p x_p$$

γραμμική συνάρτηση των $b_0, b_1, \ldots, b_p$!!

$x_1, x_2, \ldots, x_p$ μπορεί να είναι οποιεσδήποτε συναρτήσεις άλλων ανεξαρτήτων μεταβλητών ή μη.

Παραβολικά

$$E(Y) = b_0 + b_1 x + b_2 x^2 \quad (x = \text{age})$$

ηλικία

γραμμικό μοντέλο

$x_1 = x, \quad x_2 = x^2$

$$E(Y) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_1 x_2$$

x_1 x_2 x_3 x_4

γραμμικό

$$E(Y) = b_0 + b_1 e^{x_1} + b_2 \sin(2\pi f x_2)$$

x_1 x_2

f: γνωστή

$E(Y) = b_0 + e^{b_1 x_1} + e^{b_2 x_2}$ b_0, b_1, b_2 : άγνωστοι

μη γραμμικό.

$E(Y) = b_0 + b_1 x_1 + b_1^2 x_2$ μη γραμμικό μοντέλο

Τύποι πολυωνύμων

$E(Y) = b_0 + b_1 x$: γραμμική πολυωνύμιση

(γραμμική σχέση της $E(Y)$ ως προς X

$E(Y) = b_0 + b_1 x + b_2 x^2 + b_3 x^3$ πολυωνυμική πολυωνύμιση

πολυωνυμική σχέση EY ως προς X

γραμμικά μοντέλα

Απλό Μονοπαραγοντικό Μοντέλο

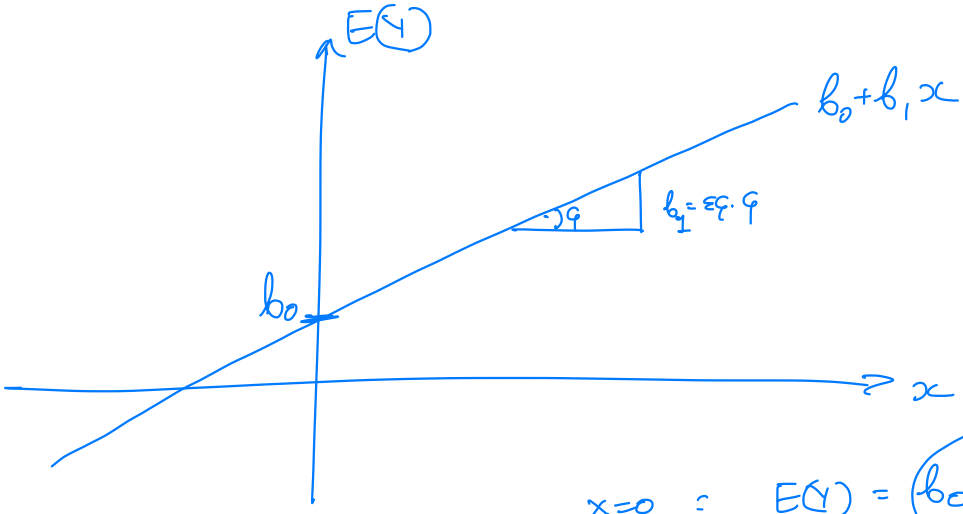
X : ανεξ. μεταβλητή

Y : εξαρτ. "

μαθηματική σχέση μεταξύ x κ' εξ υπόθεση

$$E(Y|X=x) = b_0 + b_1 x$$

b_0, b_1 } άγνωστες παράμετροι



$x=0$: $E(Y) = b_0$ = σταθ. όρος (intercept)

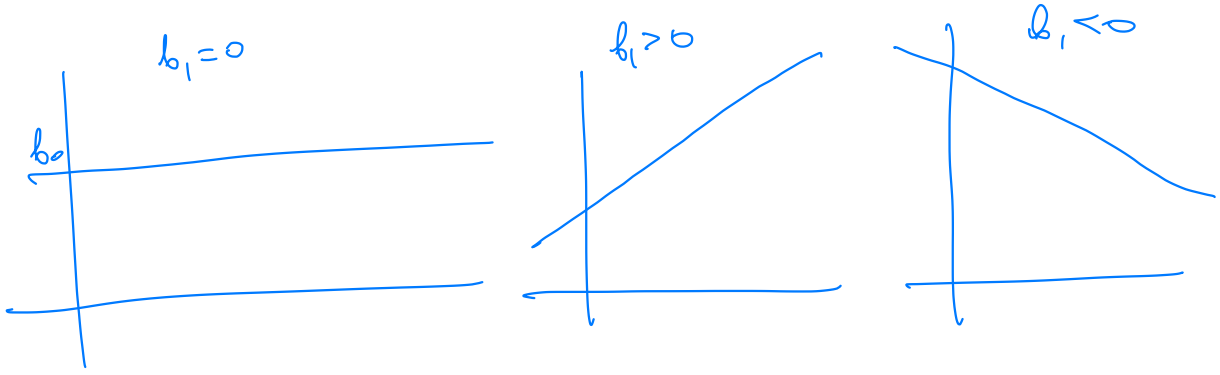
b_1: κλίση (ρυθμός μεταβολής)

Αν $x = x_0$: $E(Y) = b_0 + b_1 x_0$

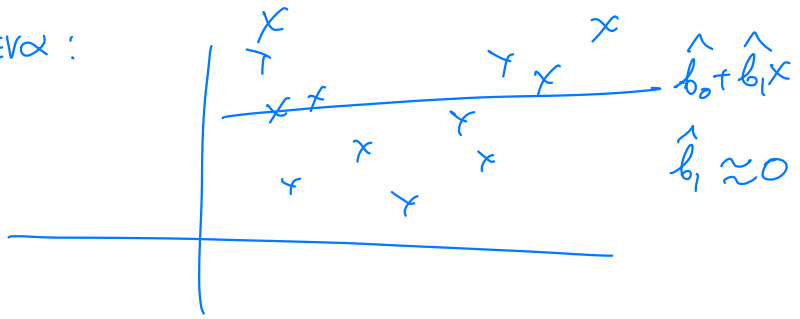
$x = x_0+1$: $E(Y) = b_0 + b_1(x_0+1) =$

$= b_0 + b_1 x_0 + b_1$

b_1: μεταβολή της $E(Y)$ για κάθε μοναδιαία μεταβολή της X.



Αν δεδομένα:



Αναμενόμενη μορφή μοντέλου παλινδρόμησης

$$E(Y|X=x) = b_0 + b_1 x$$

σήμαινε εξ. παλινδρόμησης

ισοδύναμα: $Y = b_0 + b_1 x + \varepsilon$

ε: τυχαία διακύμανση

$E(\varepsilon) = 0$

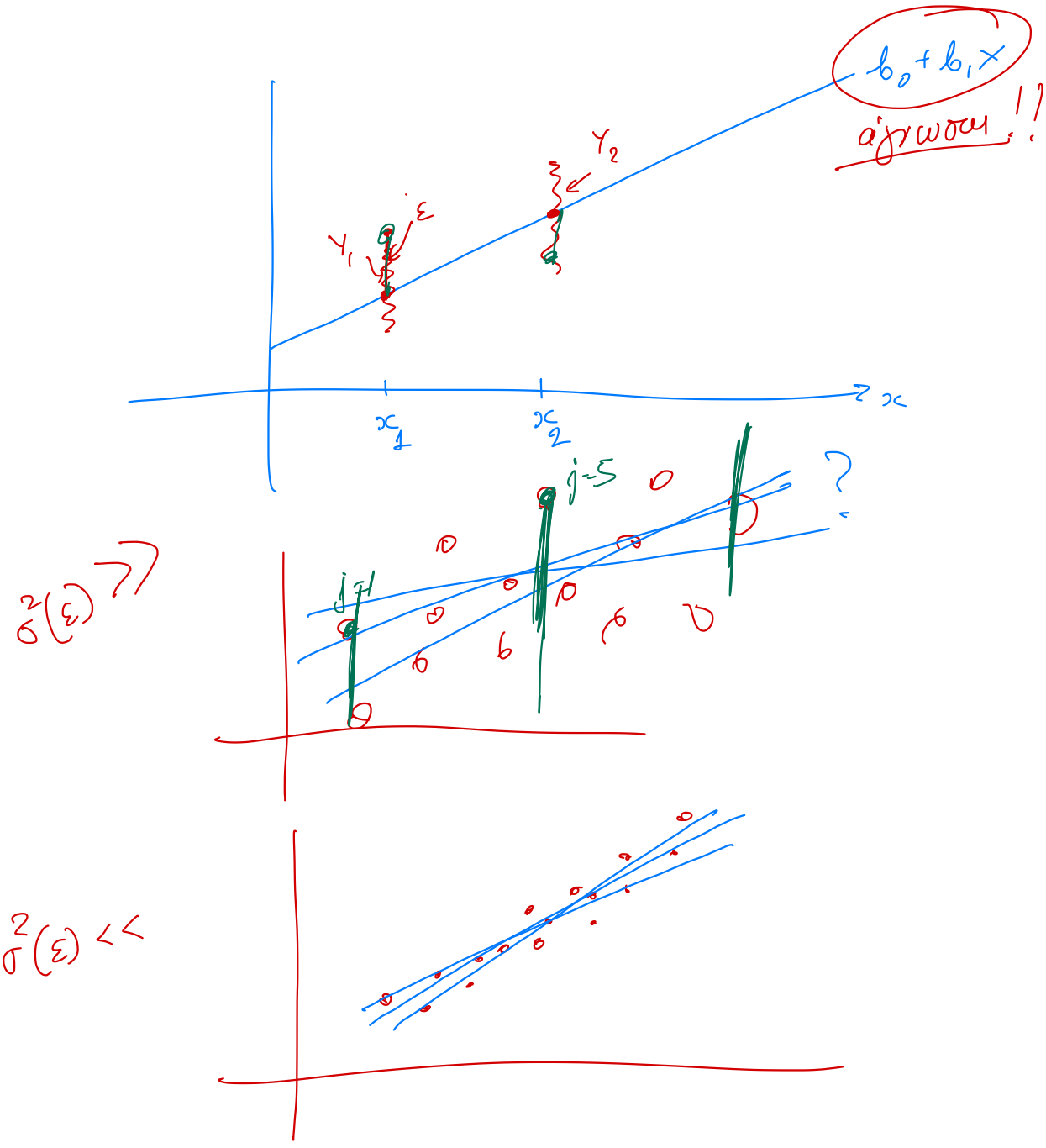
Σε ένα δείγμα η παρατήρηση $\sigma^2(\varepsilon)$?

X	Y
x_1	y_1
⋮	⋮
x_n	y_n

$Y_j = b_0 + b_1 x_j + \varepsilon_j$

η j παρατήρηση έχει $x = x_j$ κ' η απόκριση της είναι τυχαία μεταβλητή Y_j με μέση τιμή $b_0 + b_1 x_j$ κ' μια τυχαία απόκλιση ε_j

Ο πληθυσμός των ατόμων με $X = x_j$ έχει ως απόκριση Y_j τυχαία με μέση τιμή $b_0 + b_1 x_j$



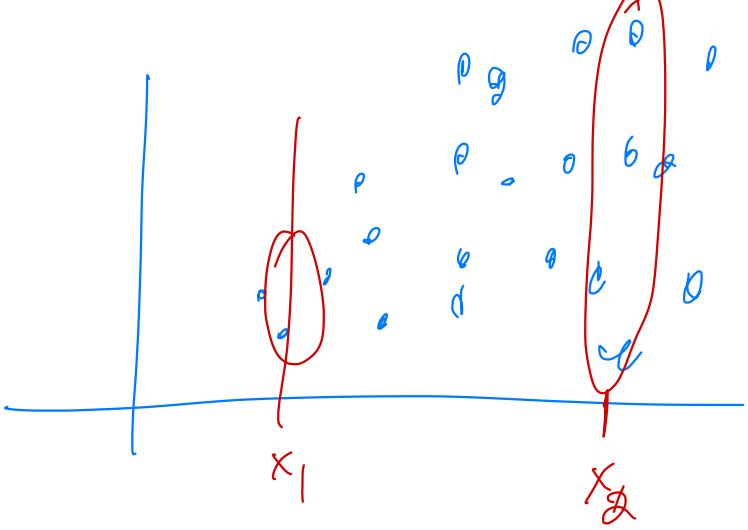
Υποθέσεις Παλινδρόμησης

(1) $Y_j = b_0 + b_1 x_j + \varepsilon_j \qquad j = 1, 2, \ldots, n$

(2) $E(\varepsilon_j) = 0 \qquad (E(Y_j) = b_0 + b_1 x_j)$

(3) $Var(\varepsilon_j) = \sigma^2 \quad \forall j$ (σταθερή ως προς j) ← ομοσκεδαστικότητα

(4) $\varepsilon_1, \varepsilon_2, \ldots$ ασυσχέτιστα (οι παρατηρούμενες δείγμα ασυσχέτιστες !!)



Ετεροσκεδαστικότητα

Παραβιάζει τις υποθέσεις

(4) Παραβιάζεται π.χ. όταν

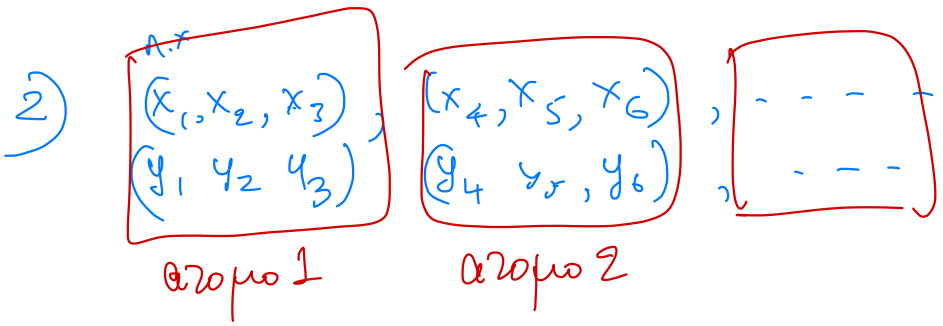
1) $x_1, x_2 \ldots$ διαδοχικές χρονικές σειρές.

π.χ. $x_1, x_2 \ldots$ ώρες του μήνα

$y_1, y_2 \ldots$ θερμοκρασίες.

χρονοσειρές (αυτοσυσχέτιση)

μοντέλα χρονοσειρών

2) 

Μοντέλα Επαναλαμβανόμενων Μετρήσεων

(Longitudinal / Repeated measurements)

Με βάση τις υποθέσεις 1-4 ⇒ αποδεκτές αμεροληψία εκτιμήσεις

ΔΕ, ελέγχους ??

5) $\varepsilon_i \sim N(0, \sigma^2)$ 4: ε_i : ανεξάρτητα

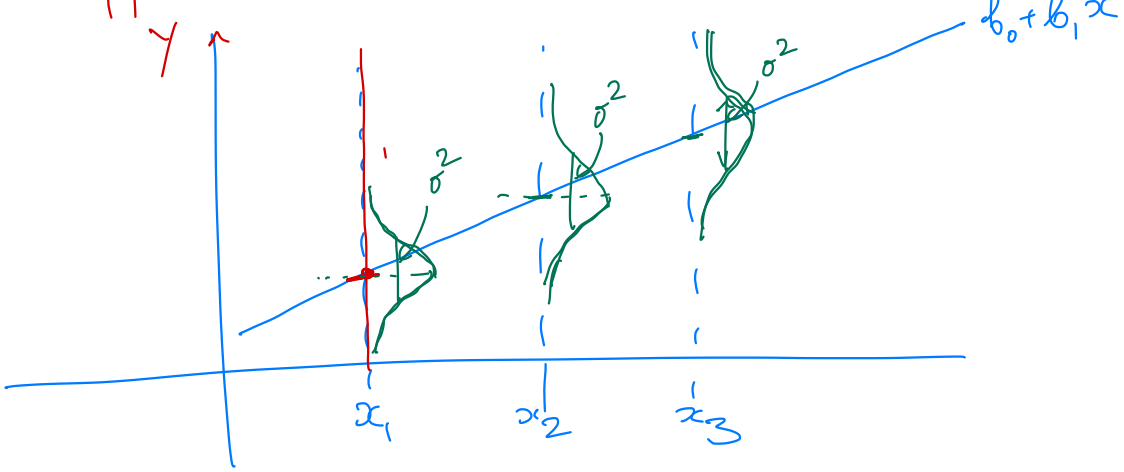
Γενικευμένο Γραμμικό Μοντέλο: άλλες υποθέσεις κατανομής

Οι υποθέσεις 1-5 γίνονται για το γενικό γραμμικό μοντέλο

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + \varepsilon$$

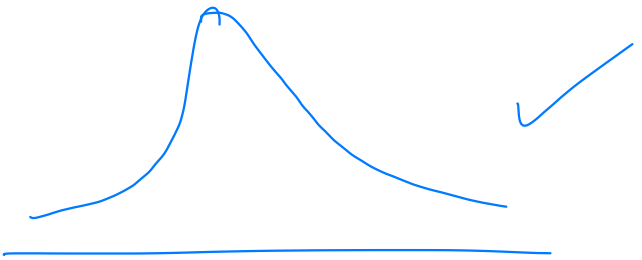
Άγνωστες παράμετροι $\{b_0, b_1, \ldots, b_p, \sigma^2\}$

Γεωμετρική ερμηνεία

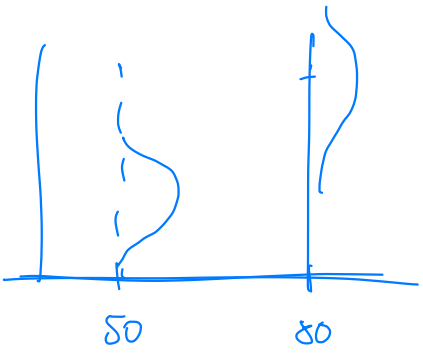


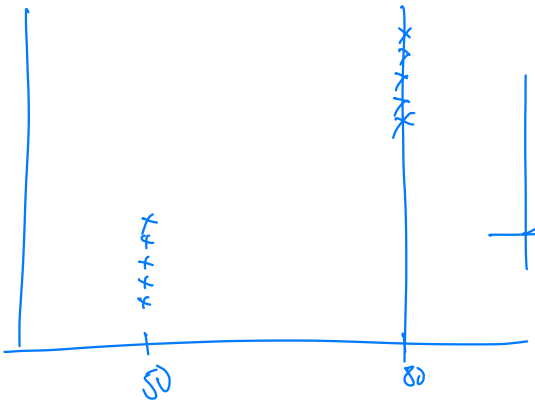
$$Y_j \sim N(b_0 + b_1 x_j, \sigma^2)$$

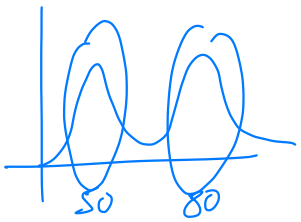
Y_j από την ίδια κανονική κατανομή



Y_j από διαφορετικές κανονικές πχ. μαλλιές $x_1 = 50$, $x_2 = 80$



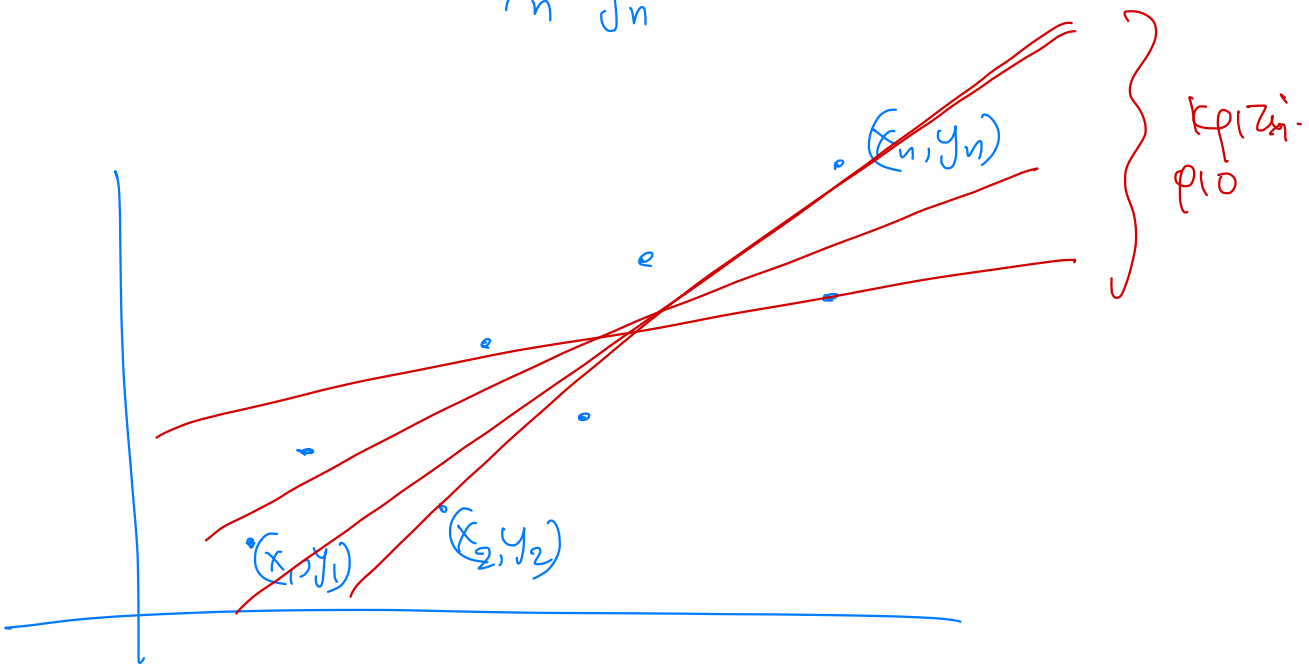




Εκτίμηση Παραμέτρων

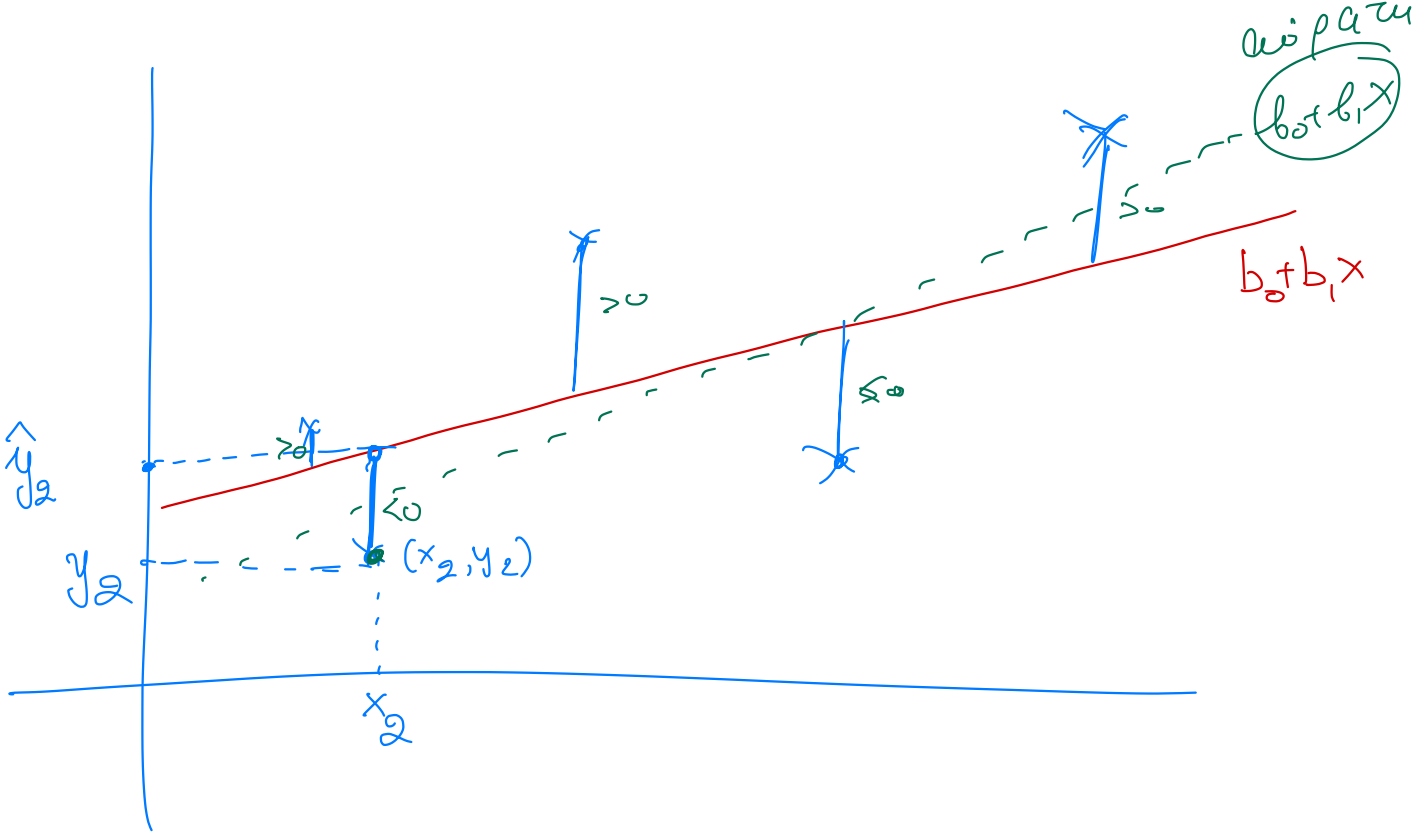
Δείγμα

$$\begin{matrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{matrix}$$



Κριτήριο Ελαχίστων Τετραγώνων

($\approx$ μέγιστη πιθανοφάνεια λόγω $N(0, \sigma^2)$)



y_2 : πραγματική τιμή

$\hat{y}_2 = b_0 + b_1 \cdot x_2$: προβλεπόμενη τιμή του Y από το μοντέλο $y = b_0 + b_1 x$ για $x = x_2$

Απόκλιση του j : $(y_j - \hat{y}_j) = e_j$: residual (κατάλοιπο)

Κριτήριο τετραγωνικής απόκλισης

$$SSE(b_0, b_1) = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 = \sum_{j=1}^{n} (y_j - (b_0 + b_1 x_j))^2$$

(sum of square errors)

Θέλουμε εκείνα τα b_0, b_1 :

$SSE(b_0, b_1)$: ελάχιστο

$\frac{\partial SSE}{\partial b_1} = 0$, $\frac{\partial SSE}{\partial b_0} = 0 \Rightarrow \dots$

$\Rightarrow \cdots \Rightarrow$ $\hat{b}_0, \hat{b}_1 \ldots$ Εκτιμώσες ελαχ. τετραγώνων

$$SSE(\hat{b}_0, \hat{b}_1) \leq SSE(b_0, b_1) \quad \forall\, b_0, b_1$$

$\hat{b}_0, \hat{b}_1$: μοναδικοί

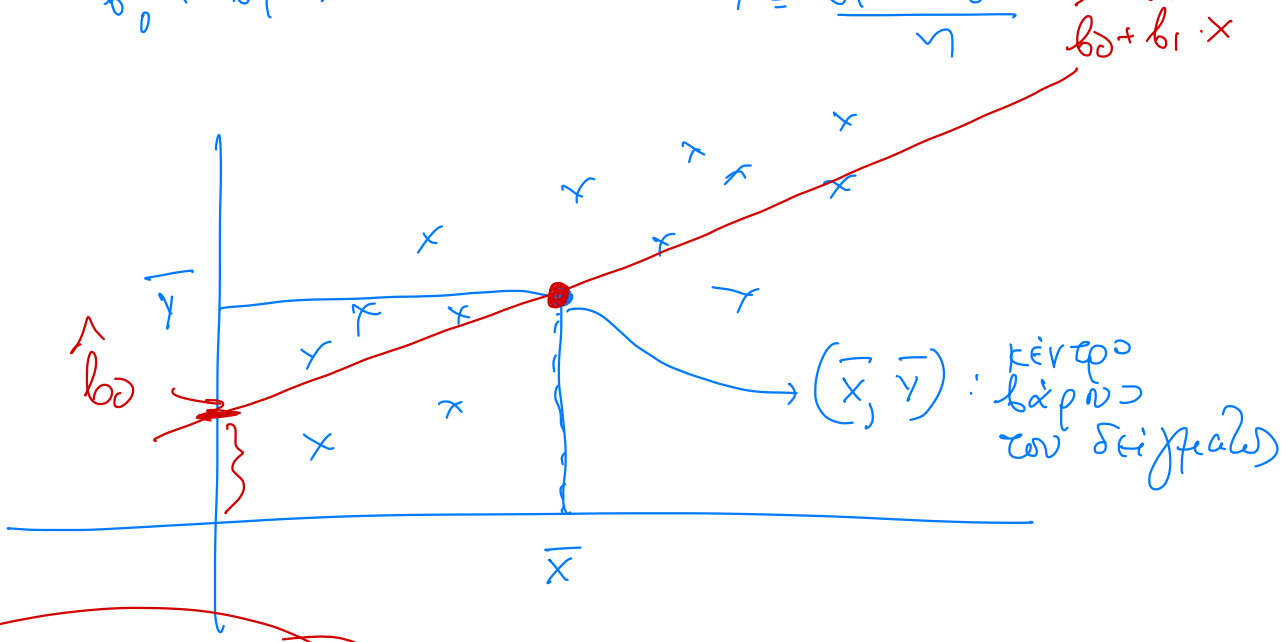
$$\hat{b}_1 = \frac{\sum x_j y_j - \frac{1}{n}\left(\sum x_j\right)\left(\sum y_j\right)}{\sum x_j^2 - \frac{1}{n}\left(\sum x_j\right)^2}$$

Παρατηρήσεις

① $\hat{b}_0 + \hat{b}_1 \cdot \bar{X} = \bar{Y}$

$$\bar{X} = \frac{x_1 + \cdots + x_n}{n}$$

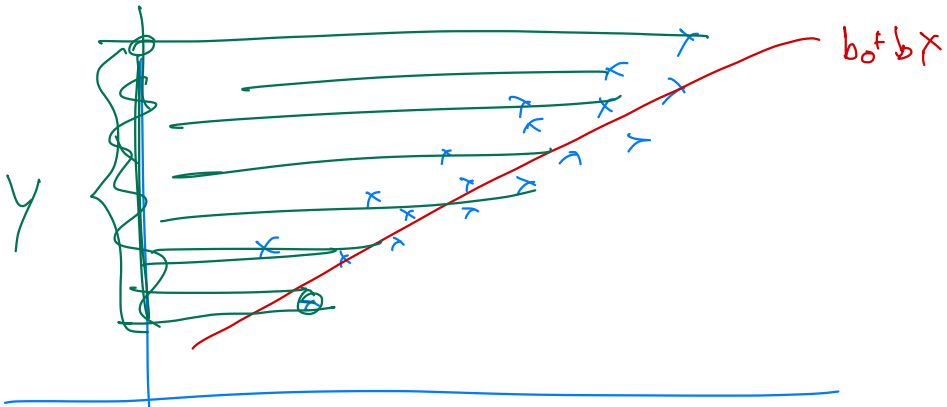
$$\bar{Y} = \frac{y_1 + \cdots + y_n}{n}$$



$\hat{Y} = \hat{b}_0 + \hat{b}_1 \cdot X$: εξίσωση παλινδρόμησης με βάση δεδομένα

Παράδειγμα

① Δείγμα 1



② Δείγμα 2