2023-12-06

Μοντέλο Παλινδρόμησης

Y = εξαρτημένη μεταβλητή / απόκριση
dependent variable / response variable

$X_1, X_2, X_3, \dots, X_p$: ανεξάρτητες μεταβλητές / παράγοντες
Independent variables / predictors / factors

Y : ποσοτική μεταβλητή (scale)

$X_1, \dots, X_p$: ποσοτικές αρχικά (θα το γενικεύσουμε)

Στατιστικό Μοντέλο Παλινδρόμησης

$$E(Y | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p) = \boxed{f(x_1, x_2, \dots, x_p)}$$

συνάρτηση παλινδρόμησης (regression function)

(άγνωστη συνάρτηση)

π.χ. Y = score

X_1 = age

X_2 = exercise time

$$E(Y|X_1=x_1, X_2=x_2) = f(x_1,x_2)$$

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

= μέση τιμή των ατόμων ηλικίας 52 ετών που αθλούνται 120 min ανα εβδομάδα

Παρατηρήσεις

① Αν $f(x_1,\dots,x_p) = C$ = σταθερή

δεν υπάρχει συσχέτιση των X με την Y

δεν υπάρχει συσχέτιση της τιμής/ες απο αυτές τις μεταβλητές

② f: άγνωστη συνάρτηση που πρέπει να εκτιμηθεί από τα δεδομένα.

[στα πλαίσια του machine learning θέλουμε να μάθουμε την f από δεδομένα]

π.χ. ① $f(x_1,x_2) = a_0 + a_1x_1 + a_2x_2$

a_0, a_1, a_2 : άγνωστες παράμετροι

γραμμική παλινδρόμηση

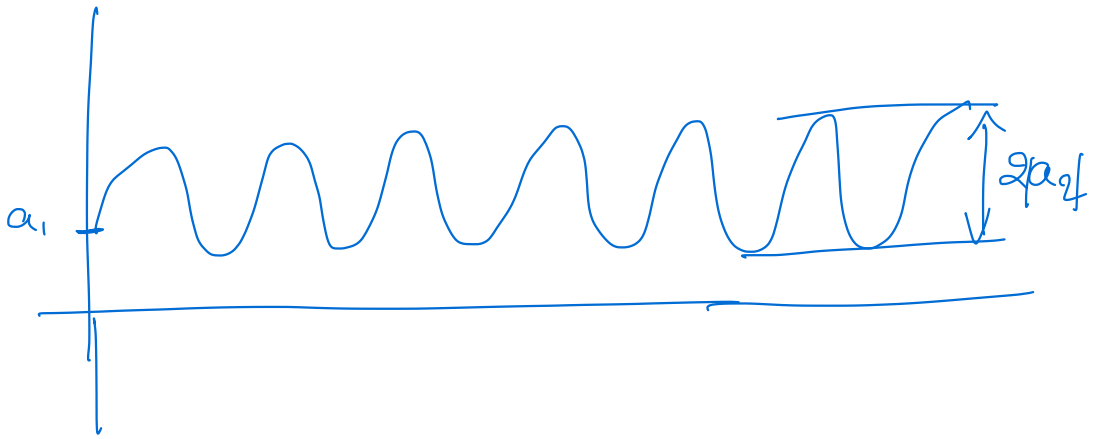
f: γραμμική ως προς x_1, x_2

(2) $f(x_1, \ldots, x_p) = C$ = σταθερά

C: άγνωστο.

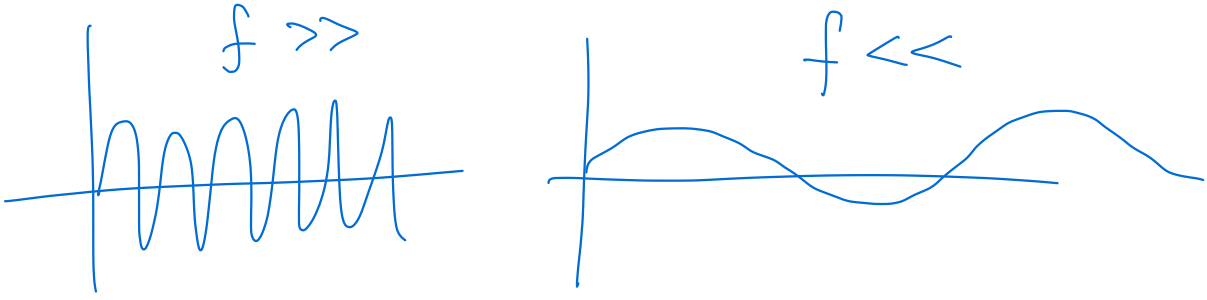
(3) $C = E(Y) \Rightarrow$ Εκτίμηση του C ⇕ Εκτίμηση της μέσης τιμής του Y σε όρο του πληθυσμού

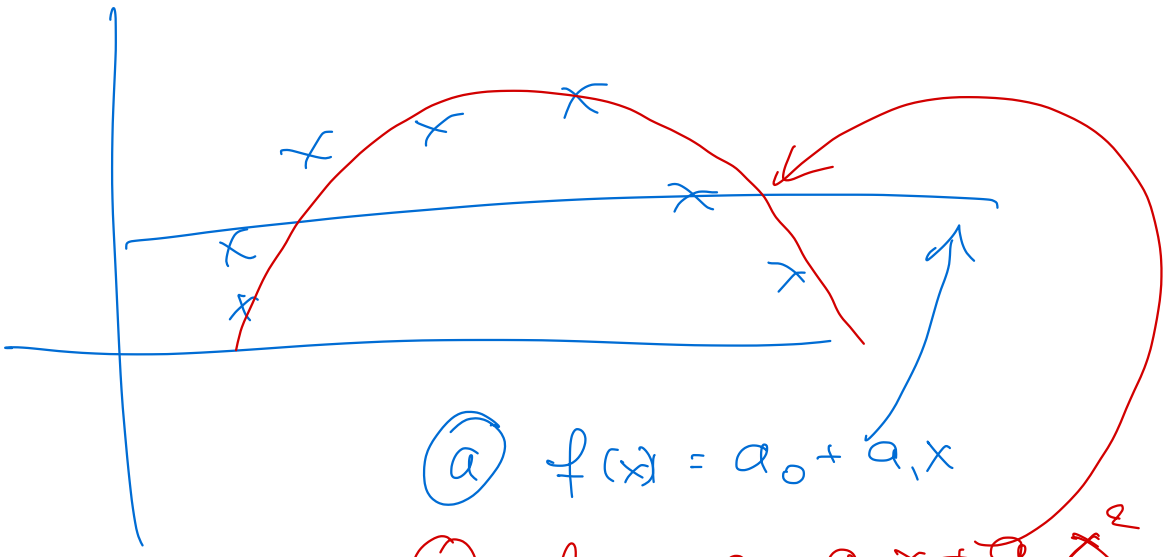
(4) $f(x) = a_1 + a_2 \sin(2x)$



(5) $f(x) = a_1 + a_2 \cdot \sin(f \cdot x)$

a_1, a_2, f : άγνωστα.





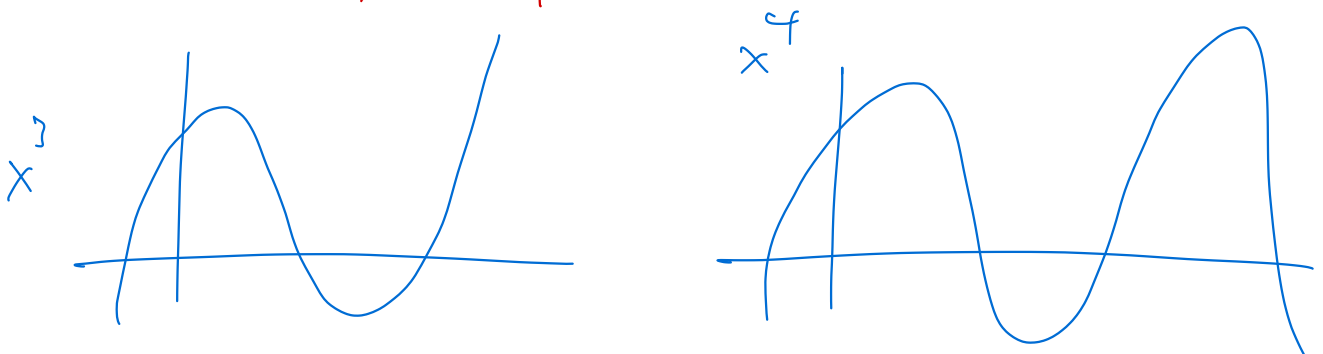
(a) $f(x) = a_0 + a_1 x$

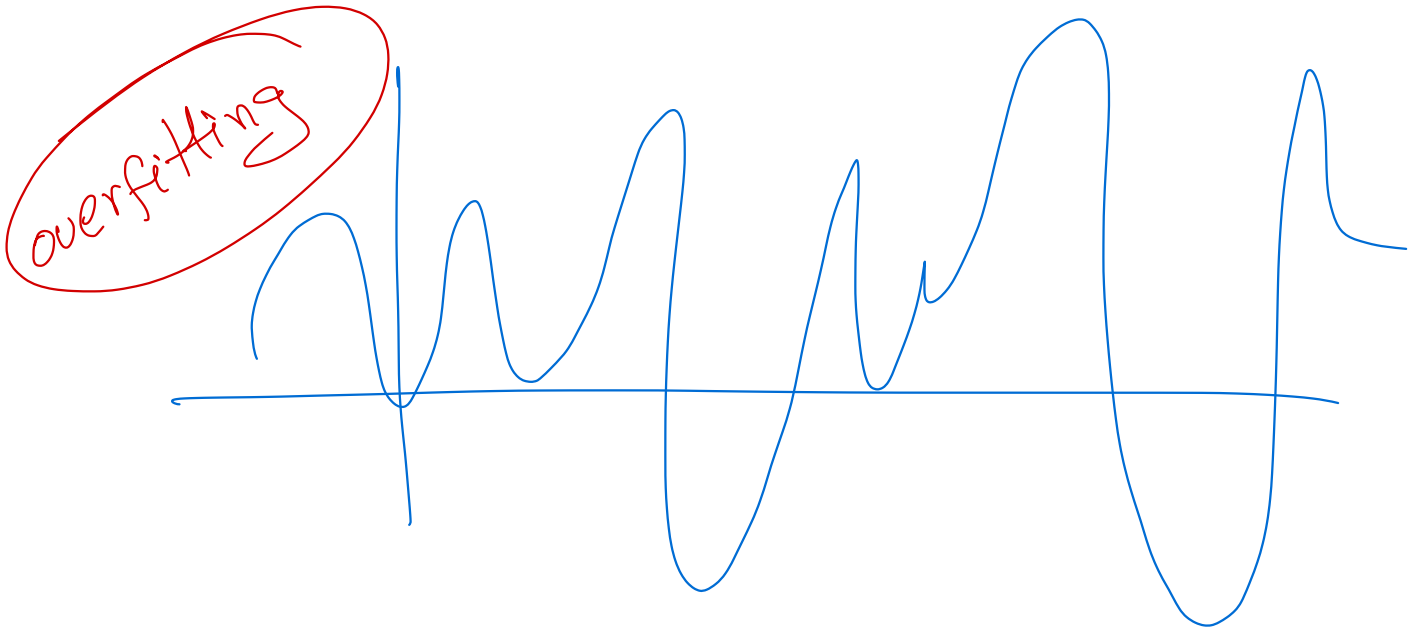
(β) $f(x) = a_1 + a_1 x + a_2 x^2$

μαι να μη θεωρώ εξ αρχής

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \dots + a_{100} x^{100}$$

Πολυωνυμική παρεμβολή





$$f(x_1, x_2, x_3) = a_0 + a_1 x_1 + a_2 x_1 x_2 + a_3 x_1^2 x_2$$
$$+ a_4 x_1^3 x_2^2 - - -$$

πολυωνυμική παλινδρόμηση

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

$f(x) = a_1 + a_2 e^{b x_1}$ εκθετική $- - -$

Γραμμικό Μοντέλο

$$E(Y | x_1, \dots, x_p) = f(x_1, \dots, x_p) =$$
$$= b_0 + b_1 g_1(x_1, \dots, x_p) + b_2 g_2(x_1, \dots, x_p) + \dots$$
$$+ \dots + b_k g_k(x_1, \dots, x_p)$$

$b_0, b_1 \dots, b_k$: άγνωστες παράμετροι

$g_1, g_2 \dots, g_k$: γνωστές συναρτήσεις

Η f είναι γραμμική ως προς τις άγνωστες παραμέτρους !!

Γραμμικό μοντέλο

π.χ. $f(x) = a_0 + a \underbrace{\sin 2x}_{g_1(x)}$

γραμμικό μοντέλο
όχι γραμμική παρεμβολή

$f(x) = a_0 + a_1 \sin(a_2 x)$

όχι γραμμικό μοντέλο

a_0, a_1, a_2 : άγνωστα

$f(x) = b_0 + b_1 e^{\gamma_1 x_1} + b_2 e^{\gamma_2 x_2}$ μη γραμμικό

$\gamma_1, \gamma_2, b_0, b_1, b_2$ άγνωστα

$f(x_1, x_2) = a_1 + a_2 x_1 + a_3 x_2 =$ γραμ. μοντέλο γραμ. παρεμβολή

$f(x_1, x_2) = a_0 + a_1 x_1 + a_1^2 x_2$ όχι γραμ. μοντ. γραμμική παρεμβ.

$f(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2$

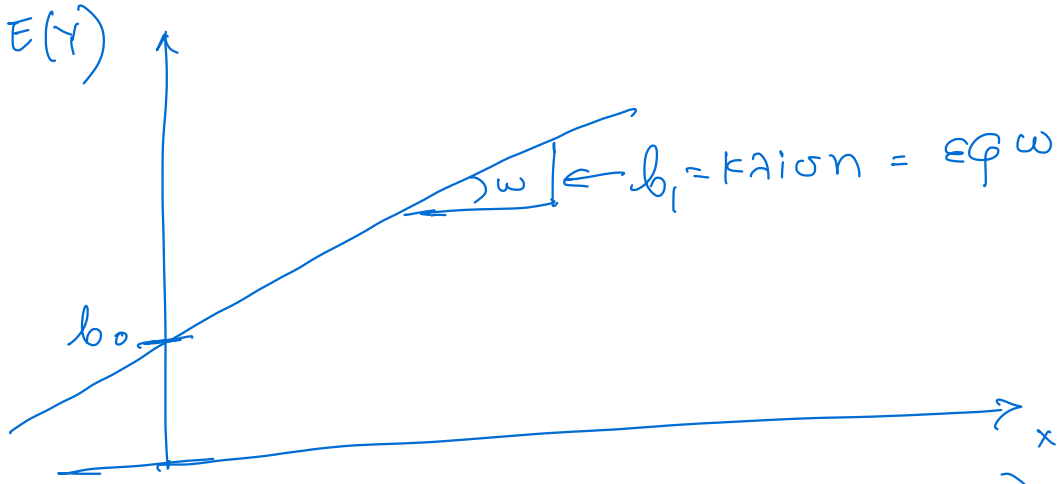
με περιορισμό $\boxed{a_2 = a_1^2}$

Απλό Μονοπαραγοντικό Μοντέλο

X : ανεξάρτητη μεταβλητή

Y : εξαρτημένη μεταβλητή

Μοντέλο $E(Y|X=x) = b_0 + b_1 x$ $\left.\begin{matrix} b_0 \\ b_1 \end{matrix}\right\}$ άγνωστες παράμετροι



b_0 : σταθερός όρος (intercept)

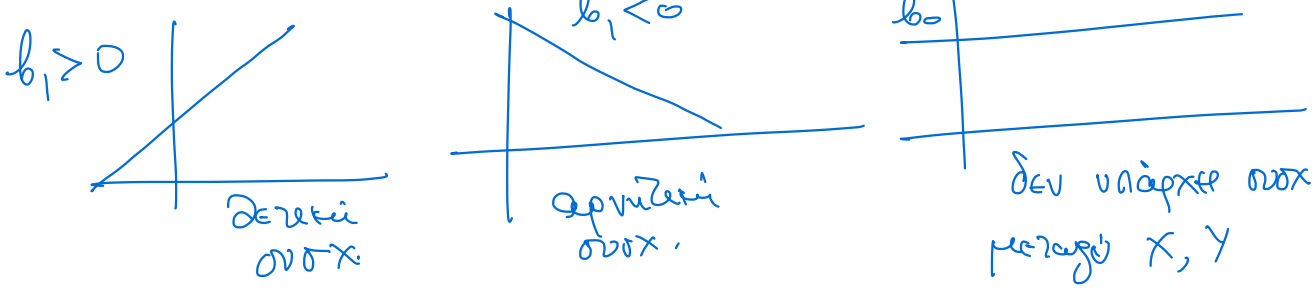
b_1 : κλίση (slope).

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

$E(Y|X=x+1) = b_0 + b_1 x + b_1$

b_1 = μεταβολή της μέσης τιμής του Y ανά μονάδα αύξησης της X.

(**Επίπτωση** της X)



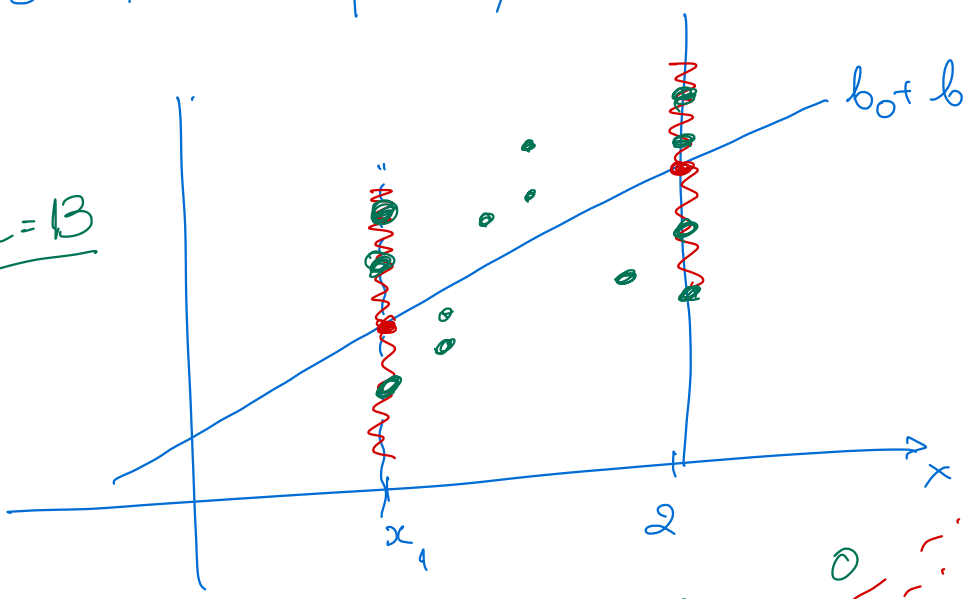
Μοντέλο $E(Y|X=x) = b_0 + b_1 x$

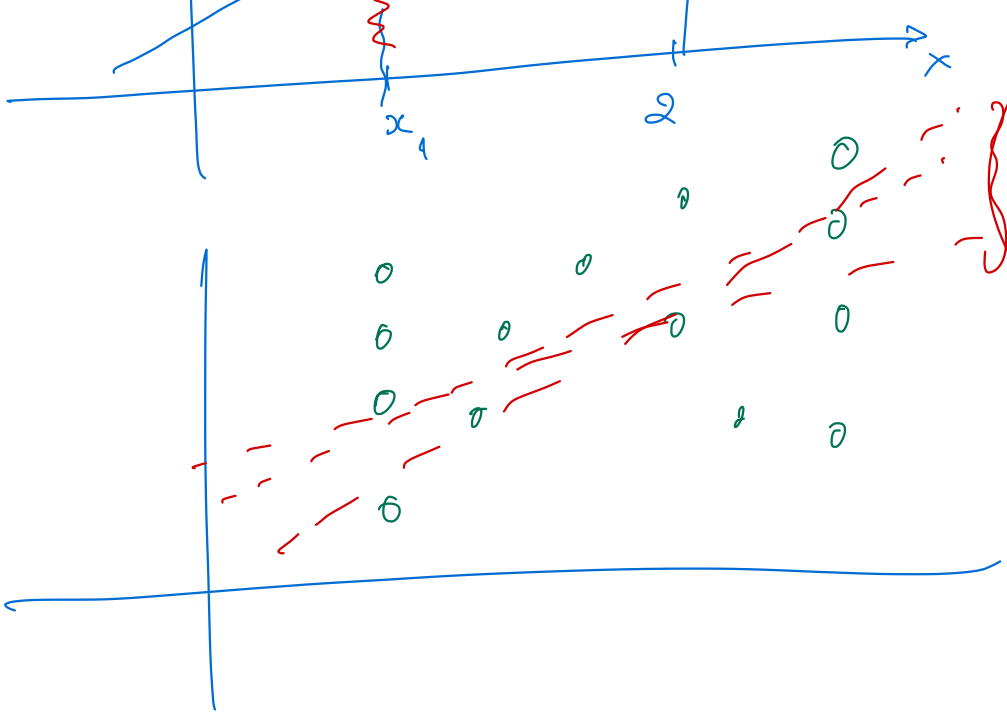
Ισοδύναμα $Y = b_0 + b_1 \cdot x + \varepsilon$,

ε τυχαία μεταβλητή
$E(\varepsilon) = 0$.

ε: τυχαία διαταραχή / απόκλιση

Δείγμα μεγέθους $n = 13$





Δείγμα

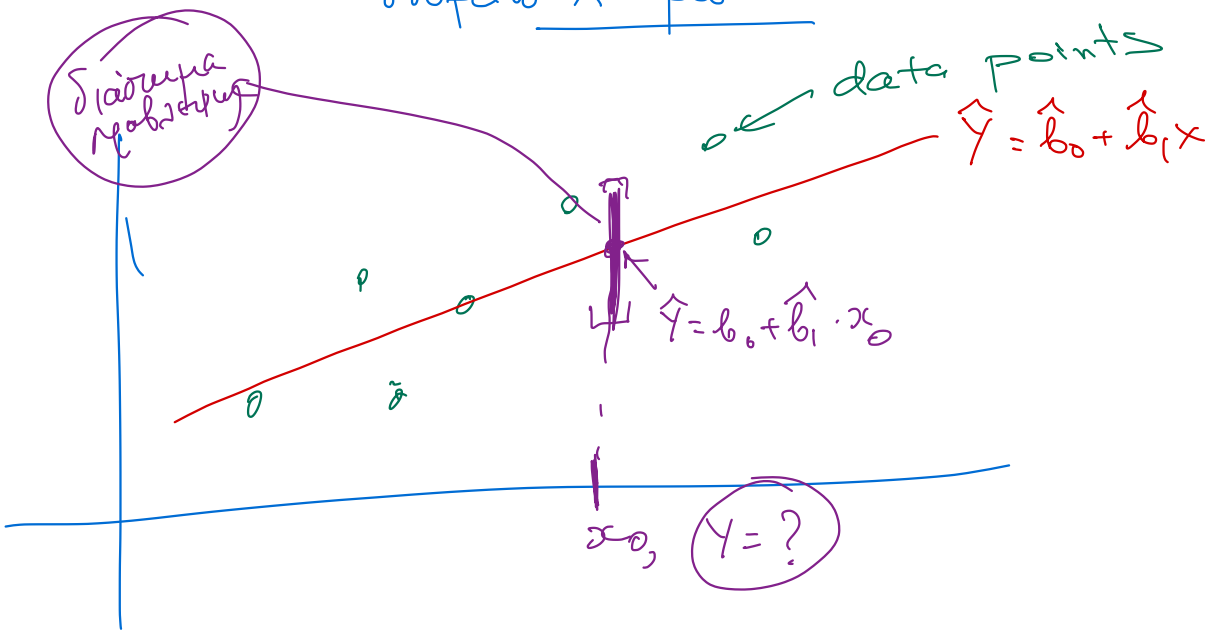
X	Y
x_1	y_1
x_2	y_2
⋮	
x_n	y_n

$j = 1, \dots, n$ παρατηρήσεις

① Από το δείγμα υπολογίζουμε $\hat{b}_0, \hat{b}_1$ εκτιμήσεις των b_0, b_1

$\hat{Y} = \hat{b}_0 + \hat{b}_1 x$: εκτίμηση της συνάρτησης παλινδρόμησης

② Μπορούμε να χρησιμοπ. τη $\hat{Y} = \hat{b}_0 + \hat{b}_1 x$ για να προβλέψουμε τιμές του Y για νέες παρατηρήσεις με δοσμένο X μόνο



Στο πλαίσιο της μηχανικής μάθησης

μεθοδολογία supervised learning (Επιβλεπόμενη μάθηση)

↓

το δείγμα χωρίζω σε X και y

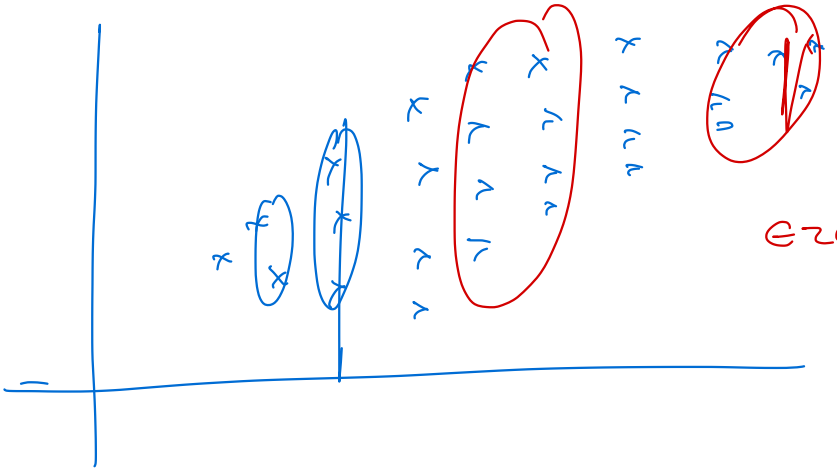
Υποθέσεις Παλινδρόμησης

(1) $Y_j = b_0 + b_1 x_j + \varepsilon_j \quad j = 1, \dots, n$

(2) $E(\varepsilon_j) = 0 \quad j = 1, \dots, n$

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

(4) $\varepsilon_1, \varepsilon_2 \dots$ αυτοσυσχετισμένα [οι παρατηρήσεις είναι αυτοσυσχετισμένες μεταξύ τους]



ετεροσκεδαστικό

Η (4) παραβιάζεται π.χ. όταν οι παρατηρήσεις προκύπτουν από διαδοχικές χρονικές στιγμές (χρονοσειρά)

μοντέλα χρονοσειρών ←

ή από επαναλαμβανόμενες μετρήσεις στο ίδιο άτομο

↓

longitudinal models

Με βάση τις υποθέσεις 1-4

μπορούμε να υπολογίσουμε αμερόληπτες εκτιμήτριες των b_0, b_1

ΔΕ ? Ελέγχους υποθέτω ?

π.χ. $H_0: b_1 = 0 \quad H_1: b_1 \neq 0$

χρειαζόμαστε επιπλέον υπόθεση για την κατανομή του ε

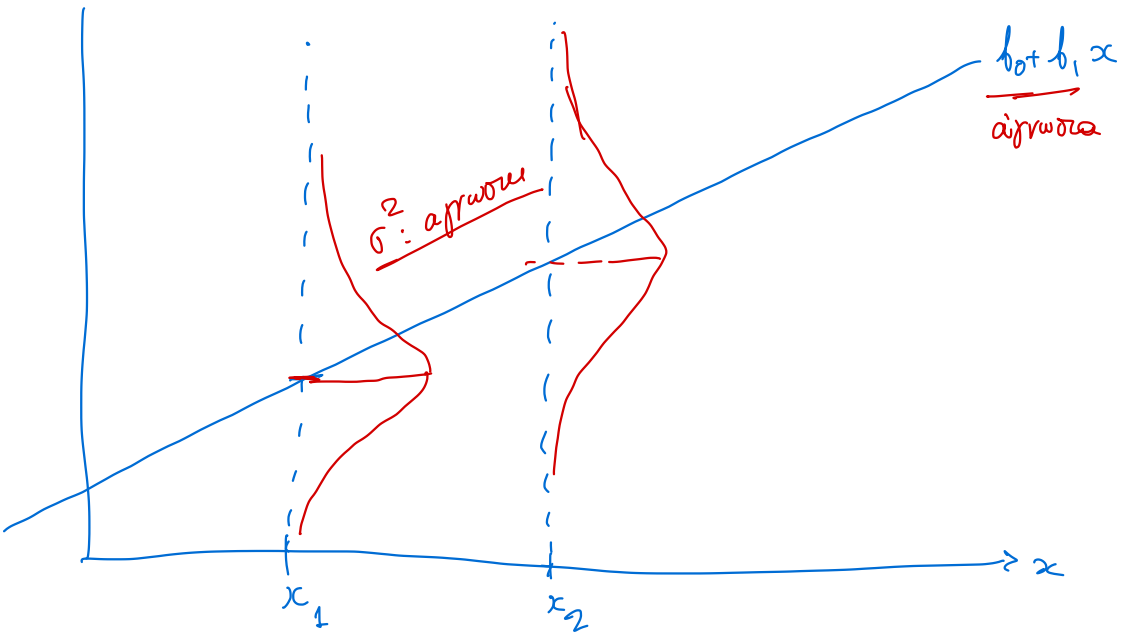
5) $\varepsilon \sim N(0, \sigma^2)$ υπόθεση κανονικότητας

↑ στα γραμμικά μοντέλα

Για άλλες κατανομές του ε → Γενικευμένα Γραμμικά Μοντέλα

Για $X = x \Rightarrow Y \sim N(b_0 + b_1 x, \sigma^2)$

$(Y = b_0 + b_1 x + \varepsilon)$

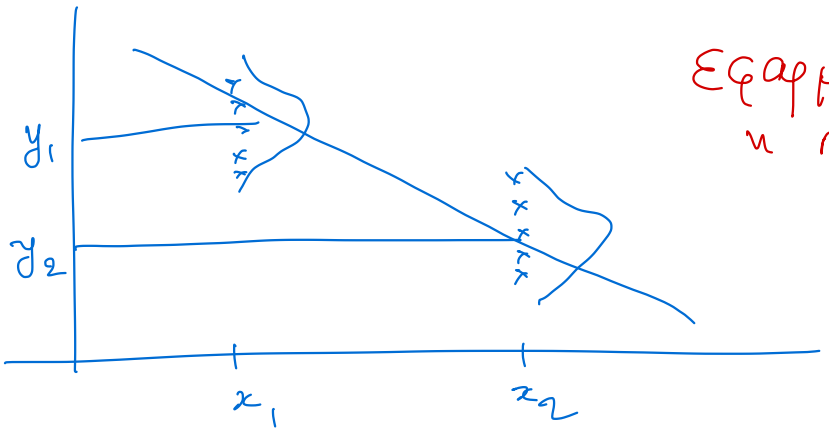


Όταν $X = x$ τότε $Y|X=x$ ίδια κατανομή

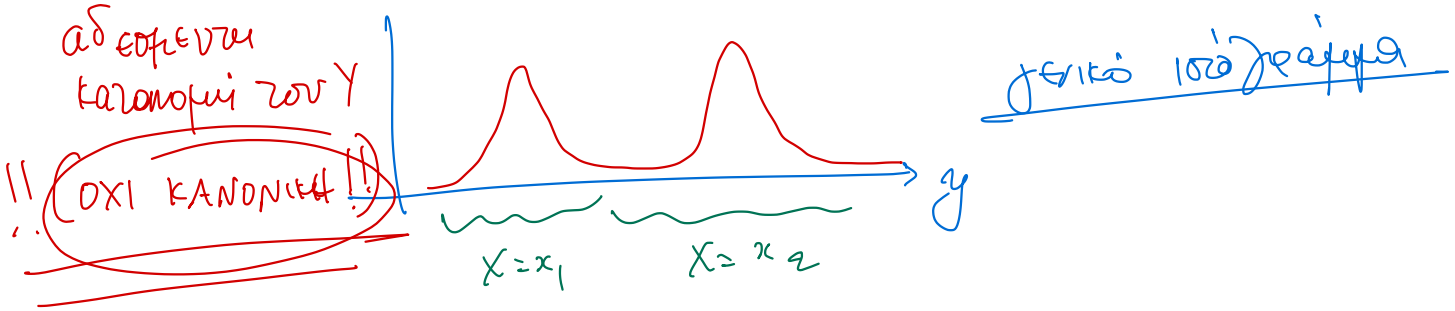
Για $X_1 = x_1$, $X_2 = x_2$

αξία καλαμιού αξία.

Έστω $X = x_1$ ή x_2 σε φορολόγιο



Εφαρμόζεται η παλινδρόμηση.

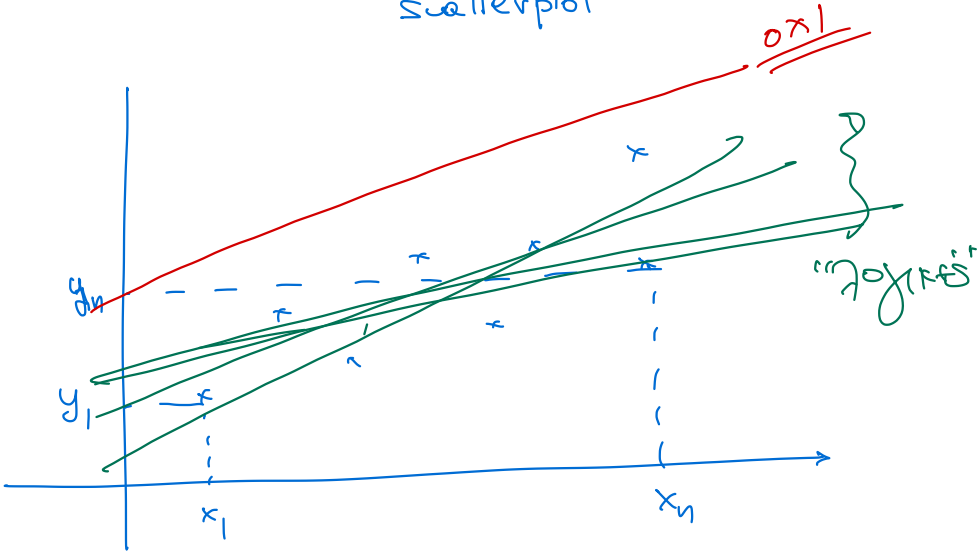


Εκτίμηση Παραμέτρων στο Μονοπαραγοντικό Μοντέλο

Δείγμα

X	Y
x_1	y_1
⋮	
x_n	y_n

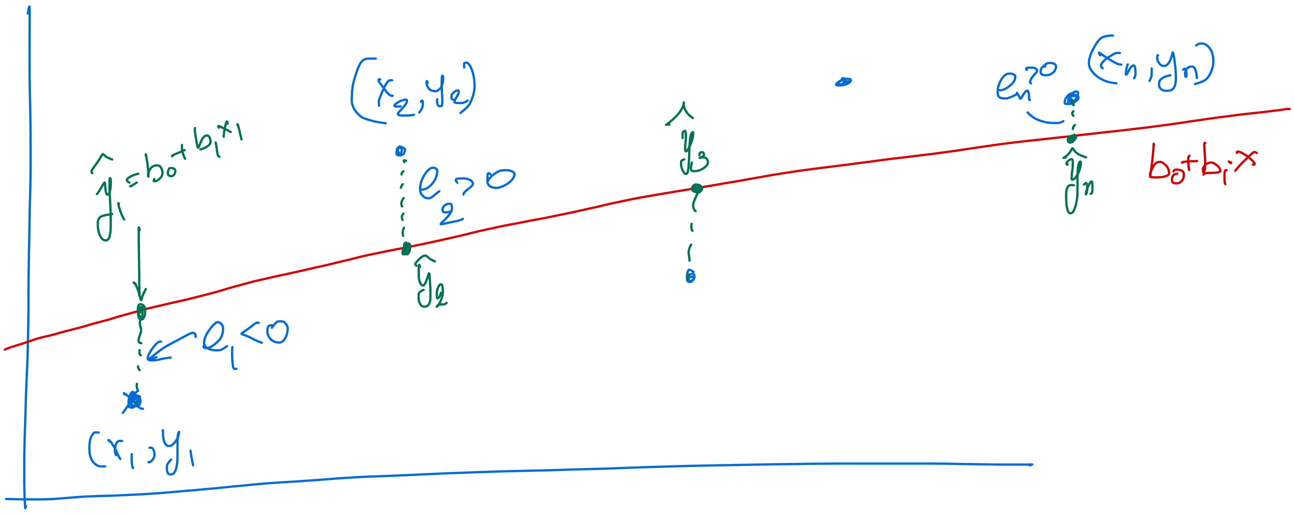
scatterplot



Κριτήριο (καλής Προσαρμογής)

Κριτήριο Ελαχίστων τετραγώνων

[ισοδύναμο με μέγιστης πιθανοφάνειας λόγω $N(0,\sigma^2)$]

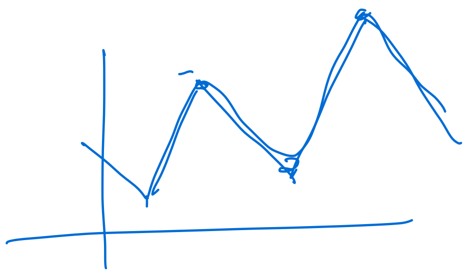


$\hat{y}_j$: προβλεπόμενες τιμές από μοντέλο b_0+b_1x

y_j : πραγματικές παρατηρήσεις

$e_j = (y_j - \hat{y}_j)$ κατάλοιπο (residual) της παρατήρησης j

Κριτήριο 1 : min $\sum_{j=1}^{n} |e_j|$



δύσκολο μαθηματικό πρόβλημα

absolute regression

[μηχ. μάθηση   Lasso]

Κριτήριο $\min \sum_{j=1}^{n} e_j^2$ [Μέθοδος ελαχίστων τετραγώνων (least squares method)]

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

↓ [Sum of Square errors]

Θέλουμε εκείνα τα b_0, b_1 έτσι ώστε

$SSE(b_0, b_1)$: ελάχιστο

Αρκεί να λύσουμε $\frac{\partial SSE}{\partial b_0} = 0$, $\frac{\partial SSE}{\partial b_1} = 0$ (κανονικές εξισώσεις)

$\Longrightarrow \cdots \Longrightarrow$ $\hat{b}_0, \hat{b}_1$ } εκτιμήτριες ελαχίστων τετραγώνων των b_0, b_1

$SSE(\hat{b}_0, \hat{b}_1) \leq SSE(b_0, b_1)$ για οποια-δήποτε (b_0, b_1)

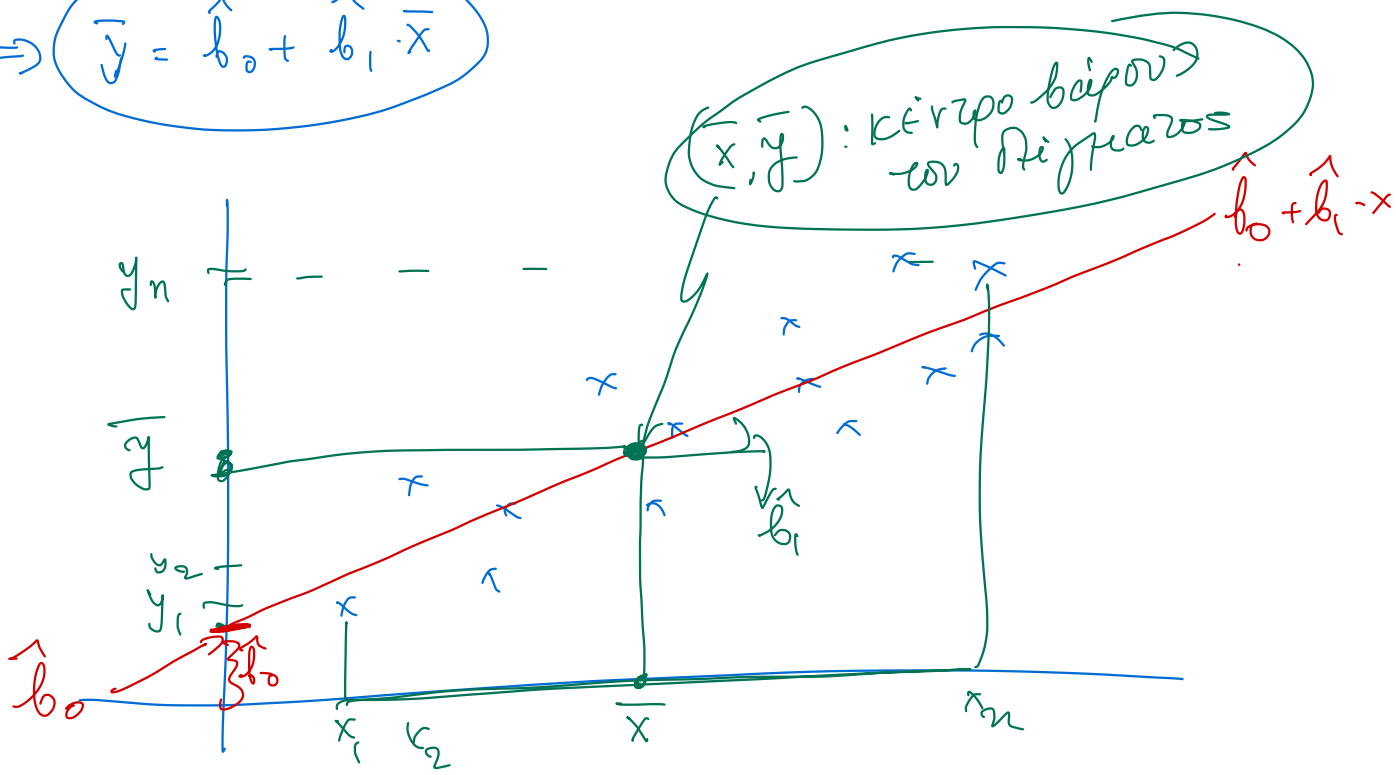
$$\hat{b}_1 = \frac{\sum_j x_j y_j - \frac{1}{n}\left(\sum x_j\right)\left(\sum y_j\right)}{\sum x_j^2 - \frac{1}{n}\left(\sum x_j\right)^2}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$$

$$\bar{Y} = \frac{1}{n}\sum y_j$$

$$\bar{x} = \frac{1}{n}\sum x_j$$

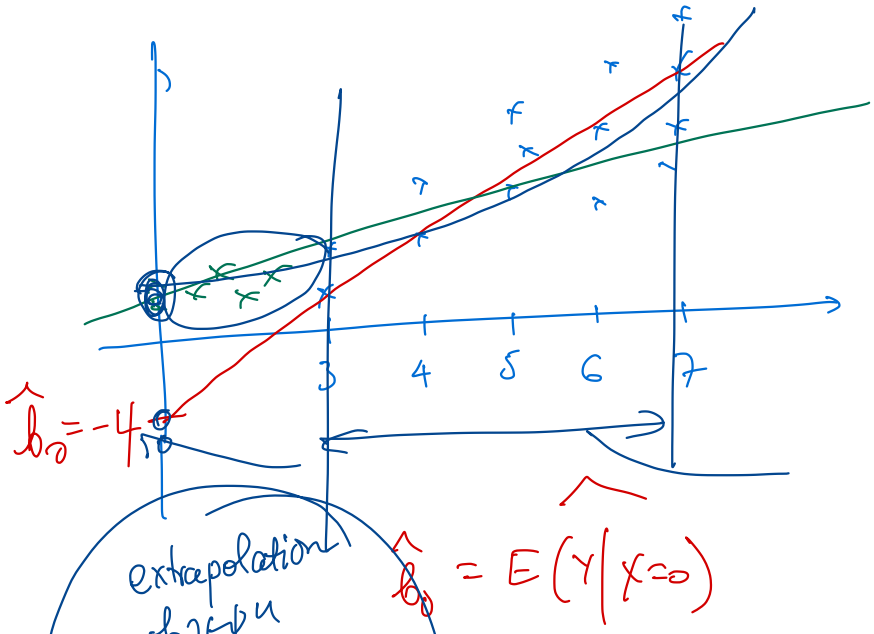
$$\Rightarrow \bar{y} = \hat{b}_0 + \hat{b}_1 \cdot \bar{x}$$



Ερμηνεία των $\hat{b}_0$, $\hat{b}_1$

$\hat{b}_1$: κλίση : Για κάθε αύξηση του x κατά 1 μονάδα η E(Y) αυξάνεται κατά $\hat{b}_1$ κατά μέσο ποσό (εκτίμηση)

$$\hat{b}_0 = E(Y \mid X=0)$$



X = age παιδιού

Y = βάρος

$n = 50$

x 3-7

extrapolation προβλήμα για μιά τιμή του x εκτός του δείγματος

$\hat{b}_0 = E(Y|X=0)$