28-2-2023

n.x. dataset exercise

Y = score

X = age

$$Y = b_0 + b_1 X + \varepsilon$$

R function lm : (linear model)

lm(model, ...)

model : formula object ($y \sim x$)

$\boxed{y \sim x} \iff \boxed{y = \underline{b_0} + \underline{b_1} x}$

απλώς lm(score ~ age)

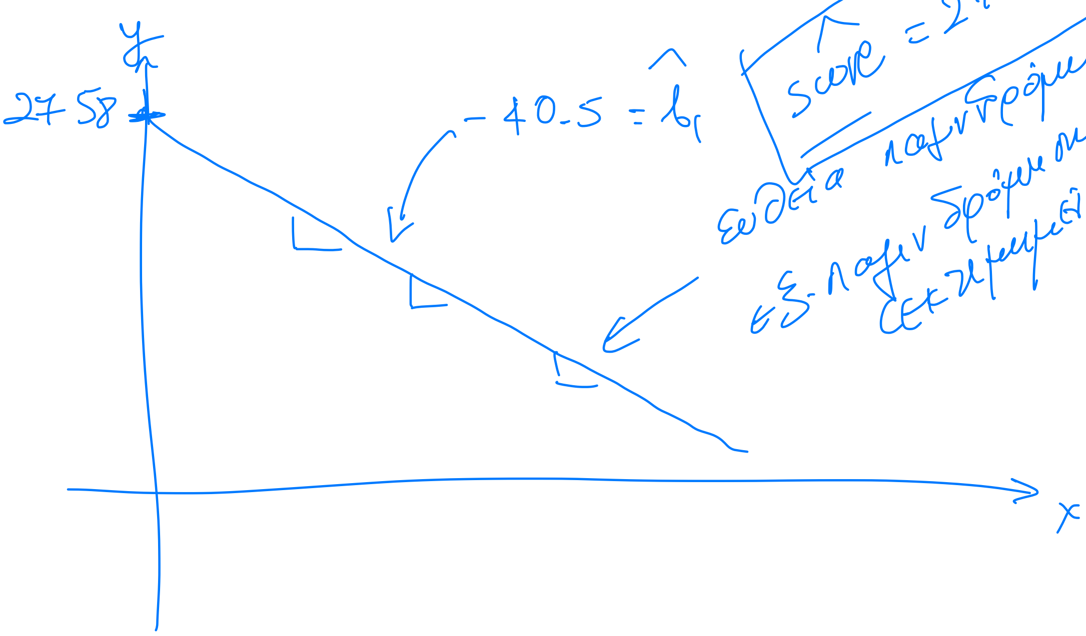
Εναλλακτικά

model1 = score ~ age
lm(model1)

Παρατηρήσεις

$\hat{b}_0 = 2758.22$

$\hat{b}_1 = -40.51$



$\boxed{\widehat{score} = 2758 - 40.5 \cdot age}$

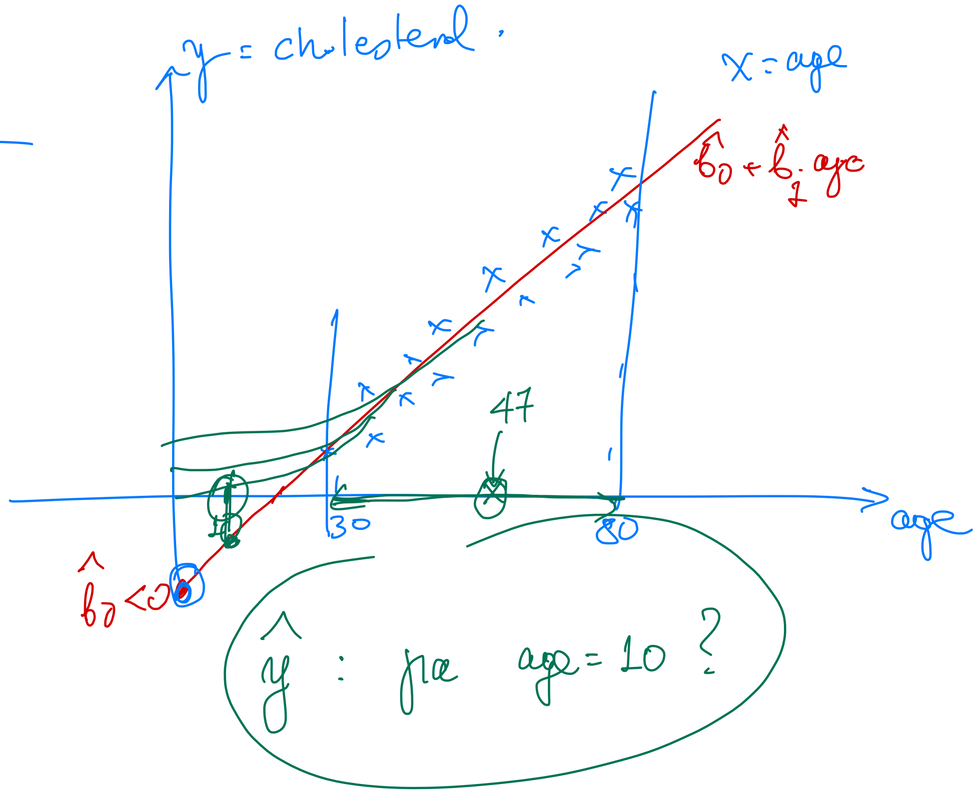
ευθεία ελαχ. τετραγώνων
εξ. παλινδρόμησης (εκτιμημένη)

① **Ερμηνεία** $\hat{b}_1 = -40.51$

Για αύξηση της ηλικίας κατά 1 έτος η μέση τιμή του Score μειώνεται κατά 40.51

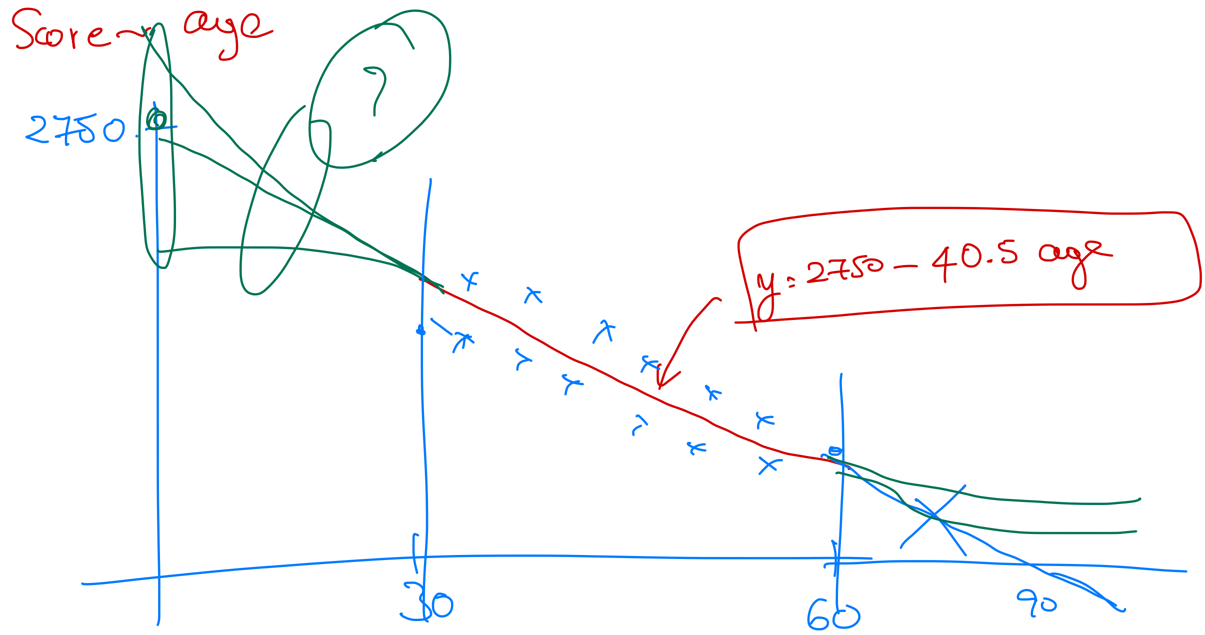
Ερμηνεία $\hat{b}_0 = 2758$ (για age = 0 ⟹ $\boxed{\widehat{Score} = 2758}$)

Άλλο παράδ.



Πρόβλεψη του E(y) για x εκτός της περιοχής τιμών του δείγματος ⟹ extrapolation × × !!

$\hat{b}_0$: έχει φυσική ερμηνεία μόνο όταν υπάρχουν τιμές x=0 (ή κοντά στο 0) στο δείγμα.



Επιπλέον πληροφορίες

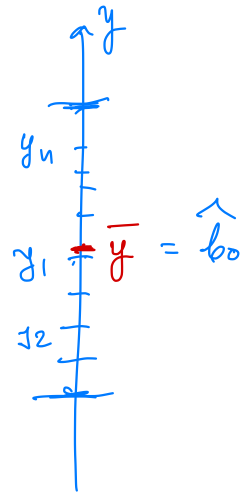
(Ανάλυση διασποράς κ' Πειραματική)

(1) Ανάλυση διασποράς (διακύμανσης) (ANOVA)

Y εξαρτ. μεταβλητή

δείγμα $(y_1, y_2, \dots, y_n)$

$$\left.\begin{aligned} Y &= b_0 + \varepsilon \\ \varepsilon &\sim N(0,\sigma^2) \end{aligned}\right\} \Leftrightarrow$$



$\Leftrightarrow Y \sim N(b_0, \sigma^2)$

b_0, σ^2 άγνωστα

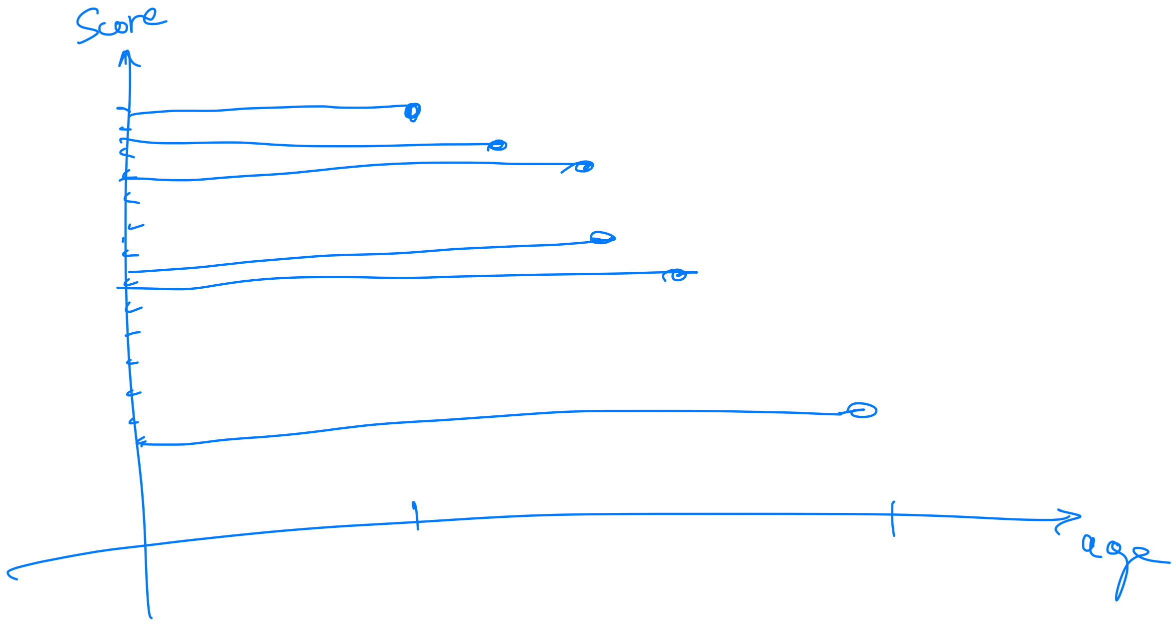
$\hat{b}_0 = \bar{y}$

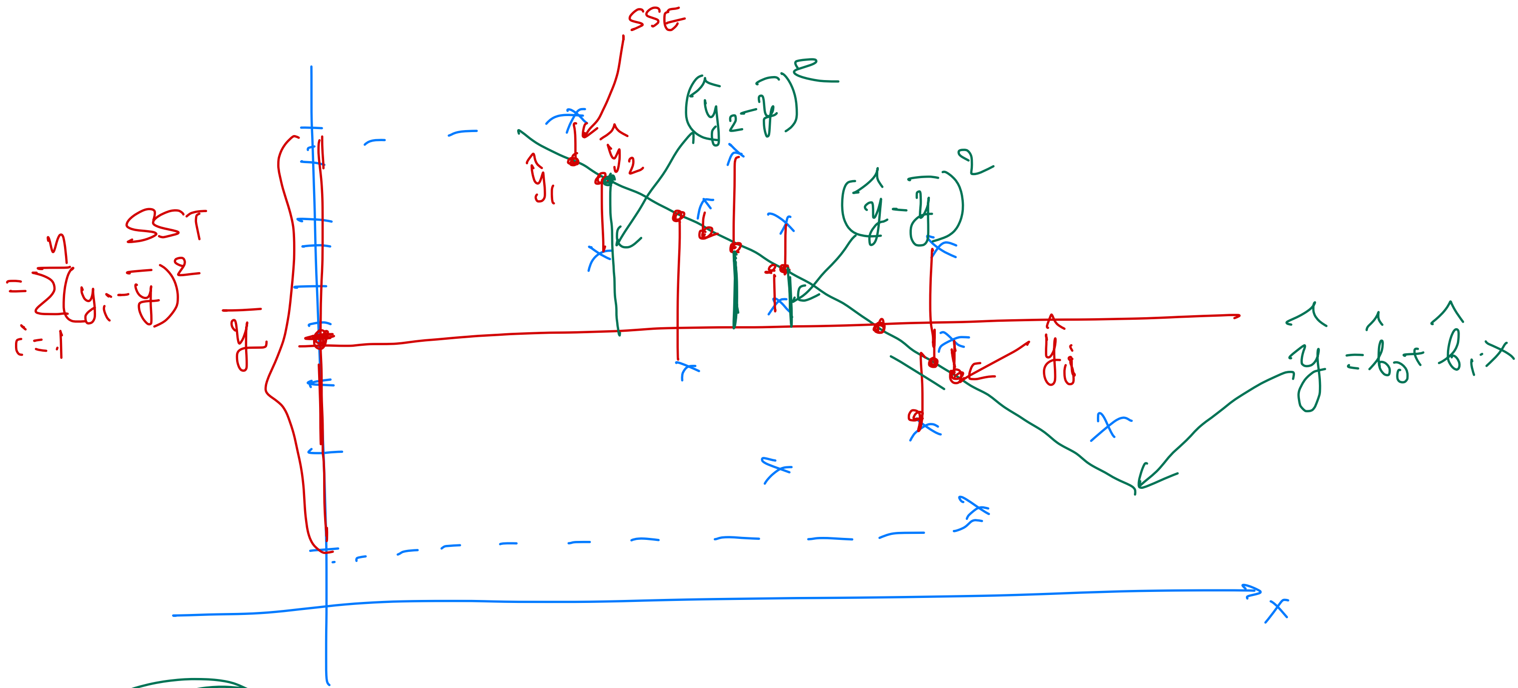
(MLE – LSE)

$\hat{\sigma}^2 = \dfrac{SST}{n-1} \sim \chi^2_{n-1}$

Διασπορά $\dfrac{1}{n-1} \underbrace{\sum (y_i - \bar{y})^2}$

συνολική διακύμανση του Y στο δείγμα μας $\Big\}$ Sum of Squares total (SST)





$$SSR = \sum(\hat{y}_j-\bar{y})^2$$

Αν $SSR \approx 0 \Rightarrow$ η ευθ. παλ. σχεδόν οριζόντια.

(εξηγημένη μεταβλητότητα του Y από το X) = Sum of squares regression

$$SSR = SS_{model}$$

$$SSE = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 = \text{Sum of Squares error} = SS_{er}$$

(αν $SSE \approx 0 \Rightarrow$

μεταβλ. του Y στο δείγμα που παραμένει ανεξήγητη από το μοντέλο $Y = b_0 + b_1 X$

$$SST = \sum_{j=1}^{n} (y_j - \bar{y})^2 \quad \text{συνολική}$$

$$SSR = \sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2 \quad \text{εξηγημένη}$$

$$SSE = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 \quad \text{ανεξήγητη.}$$

ΘΕΩΡΗΜΑ Όταν τα $\hat{b}_0, \hat{b}_1$ εκτιμώνται με βάση LSE

τότε ισχύει $\boxed{SST = SSR + SSE}$

τύπος ανάλυσης διακύμανσης

Analysis of Variance formula.

Παρένθεση

Εκτός από το κριτήριο Ελαχ. Τετρ.

Θα μπορούσαμε π.χ. να ελαχιστοποιήσουμε

$$= \sum_{j=1}^{n} \left| y_j - (b_0 + b_1 x_j) \right|$$

συνολική απόλυτη απόσταση.

$\underbrace{b_0^*, b_1^*}$ absolute distance estimates

SSR, SSE με βάση b_0^*, b_1^*

$$SST \neq SSR + SSE$$

LSE :

$$SST = \underbrace{SSR} + \underbrace{SSE}$$

συντ. προσδιορισμού $R^2 = \dfrac{SSR}{SST}$: % μεταβ. του Y στο δείγμα που εξηγείται από το μοντέλο $Y = b_0 + b_1 X$

1) $0 \leq R^2 \leq 1$

2) $R^2 = r^2$, όπου $r = Corr(x, y)$ (στο μονοπαραγοντικό μοντέλο $Y = b_0 + b_1 X$)

$SSR = 10255910$

$SSE = 1485592$

$SST = 11741503$

$R^2 = \frac{SSR}{SST} = 0.873$

Πίνακας ANOVA

mean squares

n = sample size

	SS	Degrees of freedom df	MS
X (Model)	SSR	1	$MSR = \frac{SSR}{df_{reg}}$
Error	SSE	$n-2$	$MSE = \frac{SSE}{df_{error}}$
Total	SST	$n-1$ (df_{total})	

$Y = b_0 + b_1 x + \varepsilon$

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

$df_{er} = n-2$

$$\frac{SSE}{df_{er}} = \frac{SSE}{n-2} = \hat{\sigma}^2$$ αμερόληπτη

άγνωστα

$$\sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2 \longrightarrow \sum_{j=1}^{n} (y_j - \hat{b}_0 - \hat{b}_1 x_j)^2 = SSE$$

2 παράμετροι αντικ. από εκτιμήσεις μέσα από το δείγμα

$df_{er} = n - 2$

$$MSE = \frac{SSE}{n-2} = \hat{\sigma}^2 \sim \chi^2_{n-2}$$: αμερ. εκτ. του σ^2

Βαθμοί Ελευθερίας

Στο γενικό γραμμικό μοντέλο

$$Y = \underbrace{b_0 + b_1 X_1 + \cdots + b_p X_p}_{\text{αρ. αγνώστων παραμ.} = p+1} + \varepsilon$$

$$df_{er} = n - (\# b \text{ στο μοντέλο}) = n - (p+1)$$

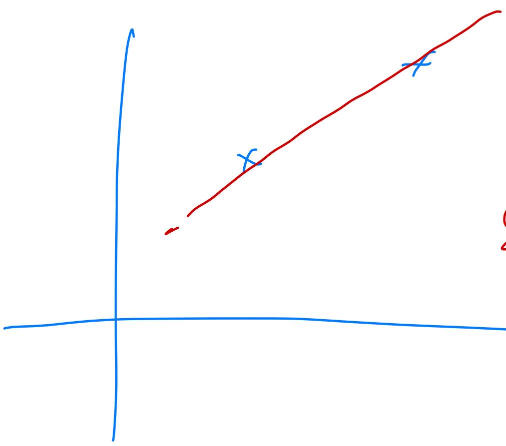
$$df_{reg} = (\# b \text{ στο μοντ.}) - 1 = p$$

$$df_{er} + df_{reg} = n - 1 = df_{total}.$$

$$SSE + SSR = \qquad = SST$$

$$df_{error} = n - (\# b)$$: "υποκατάστατο" ("ισοδύναμο") του μεγέθους δείγματος

Γενικά $\hat{\sigma}^2 = MSE = \dfrac{SSE}{df_{er}}$



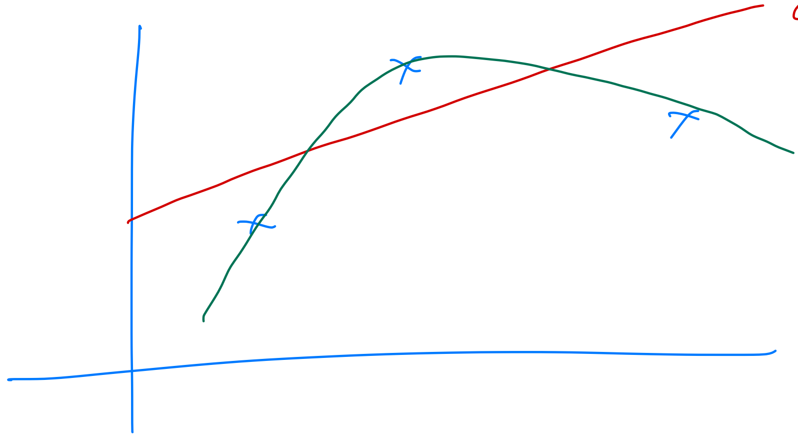
$n = 2$

$df_{er} = 0$

$SSE = 0.$

$\hat{\sigma}^2 = \dfrac{0}{0}$?

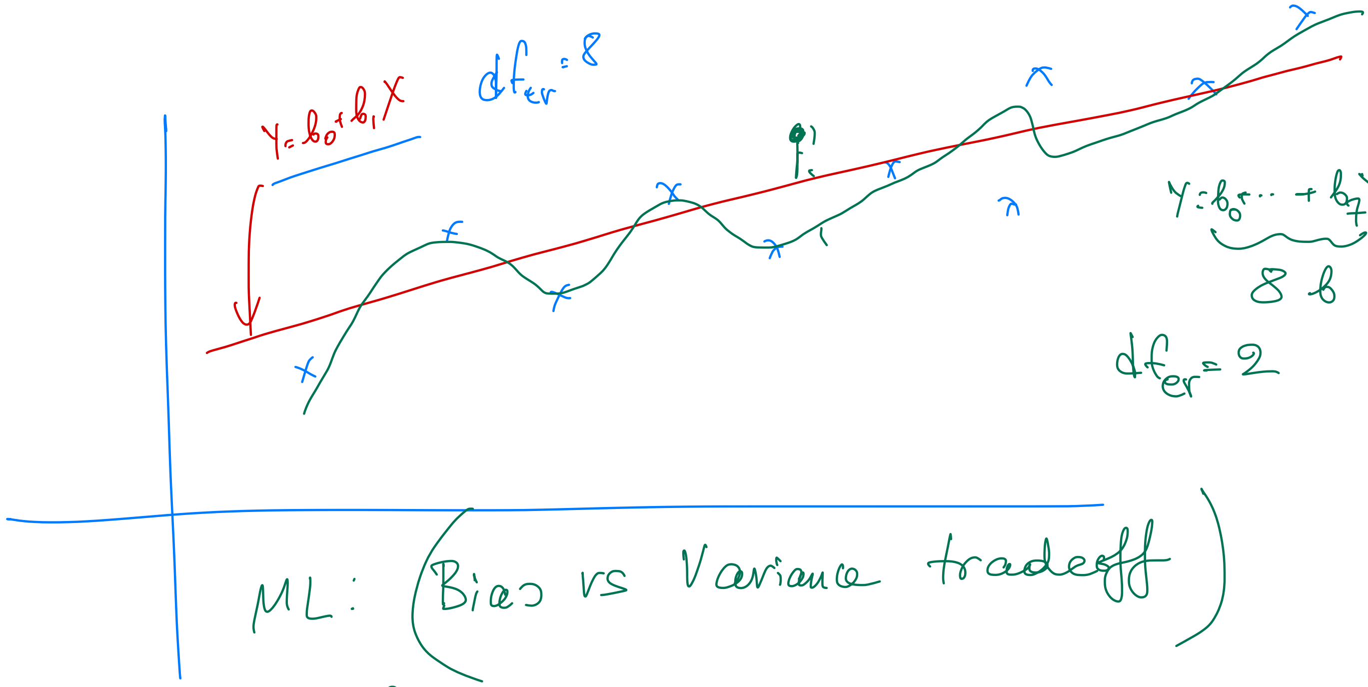
$df_{er} = 3 - 2 = 1$



$Y = b_0 + b_1 x + b_2 x^2 + \varepsilon$

$\frac{SSE = 0}{df_{er} = 0}$ $\quad \sigma^2 = ?$

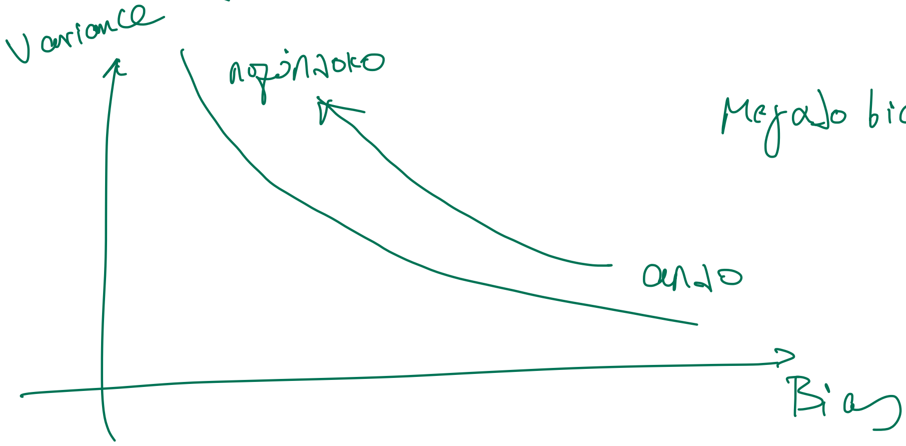
$n = 10$



$Y = \underbrace{b_0 + \cdots + b_7 X_7}_{8\ b}$

$df_{er} = 2$

ML: (Bias vs Variance tradeoff)



Μεγάλο bias: "κακό" fit

Εκτιμητές για τις παραμέτρους

$\hat{b}_1$: εκτίμηση του b_1 $\sim N(b_1, \sigma^2_{\hat{b}_1})$

$\hat{b}_0$: " του b_0 $\sim N(b_0, \sigma^2_{\hat{b}_0})$

(1) $\hat{b}_0, \hat{b}_1$: αμερόληπτες εκτιμήτριες των b_0, b_1

$E(\hat{b}_0) = b_0$

$E(\hat{b}_1) = b_1$

δικό δείγμα μας

Δείγμα 1: $\hat{b}_{1,1}$ — Δείγμα 2: $\hat{b}_{1,2}$ — ... — Δείγμα N: $\hat{b}_{1,N}$

με το ίδιο n
κ' τα ίδια $x_1, \dots, x_n$

$$\frac{1}{N}\left(\hat{b}_{1,1} + \dots + \hat{b}_{1,N}\right) \longrightarrow b_1$$

Για ΔΕ για b_1 :

$\hat{b}_1$: $E(\hat{b}_1) = b_1$

$s^2_{\hat{b}_1}$: δειγ. διασπορά $\hat{b}_1$

(αντιστ. $\sigma^2(\bar{y}) = \frac{s^2}{n}$)

$s_{\hat{b}_1} = \sqrt{s^2_{\hat{b}_1}}$: τυπ. σφάλμα του $\hat{b}_1$

$$\boxed{\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{df,er}} \Rightarrow \text{ΔΕ}$$

$$\boxed{\frac{\hat{b}_0 - b_0}{S_{\hat{b}_0}} \sim t_{df, er.}}$$

$$\Delta E \; b_1 : \left(\hat{b}_1 - t_{\alpha/2, df_{er}} \cdot S_{\hat{b}_1} \leq b_1 \leq \hat{b}_1 + t_{\alpha/2, df_{er}} \cdot S_{\hat{b}_1} \right)$$

$$\Delta E \; b_0 : \hat{b}_0 \; \ldots \leq b_0 \leq \hat{b}_0 \; \ldots$$

Ελεγχος Σημαντικότητας

$$H_0 : b_1 = 0 \qquad H_a : b_1 \neq 0$$

$$t\text{-test}, \quad t = \frac{\hat{b}_1 - 0}{S_{\hat{b}_1}}$$

accept H_0 : αν $|t| \leq t_{\alpha/2, df_{er}}$ $\quad p \geq \alpha$

reject H_0 : αν $|t| > t_{\alpha/2, df_{er}}$ $\quad p < \alpha$

p = pvalue