2023-12-08

Αξιολόγηση Μοντέλου - Συμπερασματολογία

$Y = b_0 + b_1 X + \varepsilon$ $\qquad \varepsilon \sim N(0, \sigma^2)$

τυχαία απόκλιση

τυχαίο δείγμα

Δείγμα

X	Y
x_1	y_1
$\vdots$	$\vdots$
x_n	y_n

$\Longrightarrow$

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

$$= \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$$

$\min SSE(b_0, b_1) \Longrightarrow$ Εκτιμήτριες ελαχ. τετρ.

$\hat{\beta}_0, \hat{\beta}_1$

Ιδιότητες

$\left.\begin{array}{l} E(\hat{b}_1) = b_1 \\ E(\hat{b}_0) = b_0 \end{array}\right\}$ αμερόληπτες εκτιμήτριες (unbiased)

$\forall b_0 \ \forall b_1$

απόδειξη

Εκτίμηση μιας θ — αναδρομή-εκτιμήτρια

Δείγμα $(y_1, \dots, y_n) \Longrightarrow \hat{\theta}(y_1, \dots, y_n)$

$\downarrow$

τιμή του θ για $(y_1, \dots, y_n)$

''

Εκτίμηση

π.χ. $\overline{X}_n = \dfrac{X_1 + \dots + X_n}{n}$ εκτιμήτρια

για συγκεκρ. δείγμα $\overline{X}_n = 5.7 \leftarrow$ εκτίμηση του θ.

Διαφορετικό δείγμα $\Longrightarrow (y_1', \dots, y_n') \rightarrow \hat{\theta}_2 =$

$= \hat{\theta}(y_1', \dots, y_n')$

$\boxed{\hat{\theta} \text{ : τυχαία μεταβλητή}}$

$E(\hat{\theta})$: Επαναλαμβάνουμε τη διαδικασία $N \rightarrow \infty$ φορές με τις ίδιες συνθήκες εκτιμήσεις $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_N$

$$E(\hat{\theta}) \approx \frac{\hat{\theta}_1 + \dots + \hat{\theta}_N}{N}$$

Αν $E(\hat{\theta}) \neq \theta \leftarrow$ μεροληψία [συστηματικό σφάλμα] π.χ. Αν $\tilde{\beta}_0, \tilde{\beta}_1$: $\min_{b_0, b_1} \sum_{i=1}^{n} |y_i - b_0 - b_1 x|$

Τυχαίο δείγμα στο γραμμικό μοντέλο

Δείγμα x	y	Άλλο δείγμα
x_1	y_1	y_1'
x_2	$\vdots$	$\vdots$
$\vdots$		
x_n	y_n	y_n'

τα x ίδια !!

Τα $x_1, \dots, x_n$ θεωρούνται σταθερά

Η μόνη τυχαιότητα προκύπτει από $\varepsilon_1, \dots, \varepsilon_n$
δηλ. $y_1, \dots, y_n$

---

① $E(\hat{\beta}_0) = \beta_0$

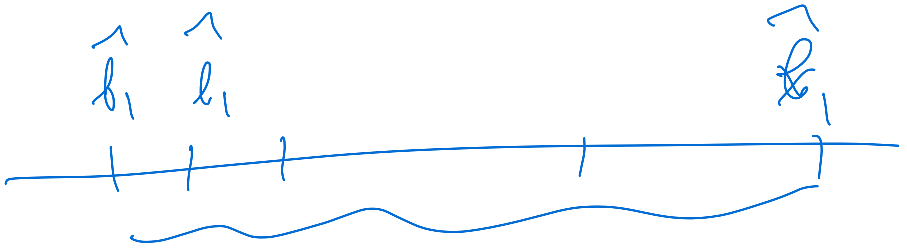
$E(\hat{\beta}_1) = \beta_1$

② Κατανομή των $\hat{b}_0, \hat{b}_1$?

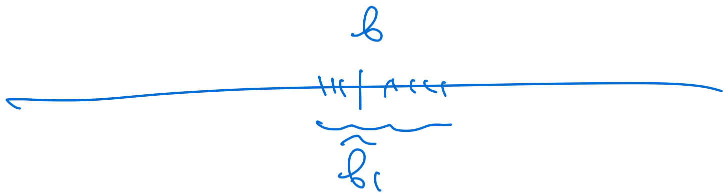
③ Διακύμανση / Διασπορά ?

Αν $\hat{b}_1$ μεγάλη διασπορά

$\sigma^2(\hat{b}_1)$

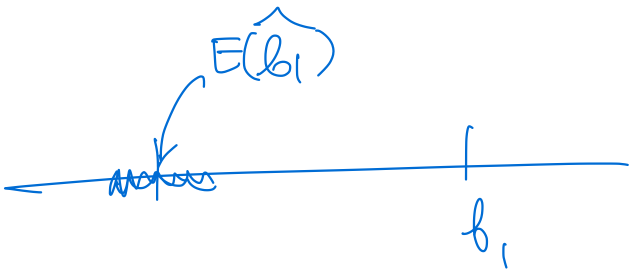


$\sigma^2(\hat{b}_1) \approx 0$



Bias/Variance

Biased, $\sigma^2(\hat{b}_1) \approx 0$



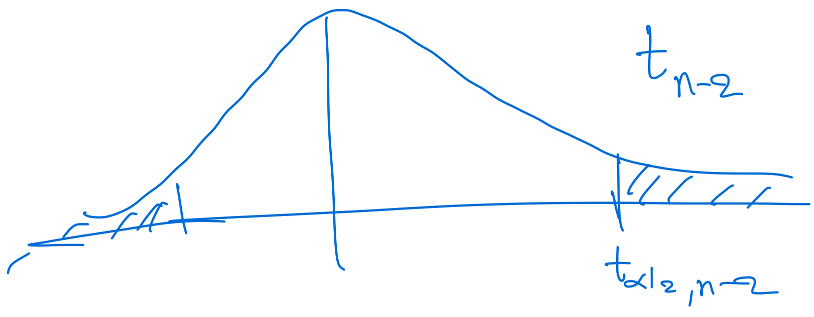
$$\hat{b}_0 \sim N(b_0, \sigma^2(\hat{b}_0))$$

$$\hat{b}_1 \sim N(b_1, \underline{\sigma^2(\hat{b}_1)})$$

$$s^2(\hat{b}_0) = \widehat{\sigma^2}(\hat{b}_0)$$

$$s^2(\hat{b}_1) = \widehat{\sigma^2}(\hat{b}_1)$$

ΔΕ b_1: $\hat{b}_1 - \underbrace{t_{\alpha/2, n-2}} \leq b_1 \leq \hat{b}_1 - t_{\alpha/2, n-2}\, s(\hat{b}_1)$

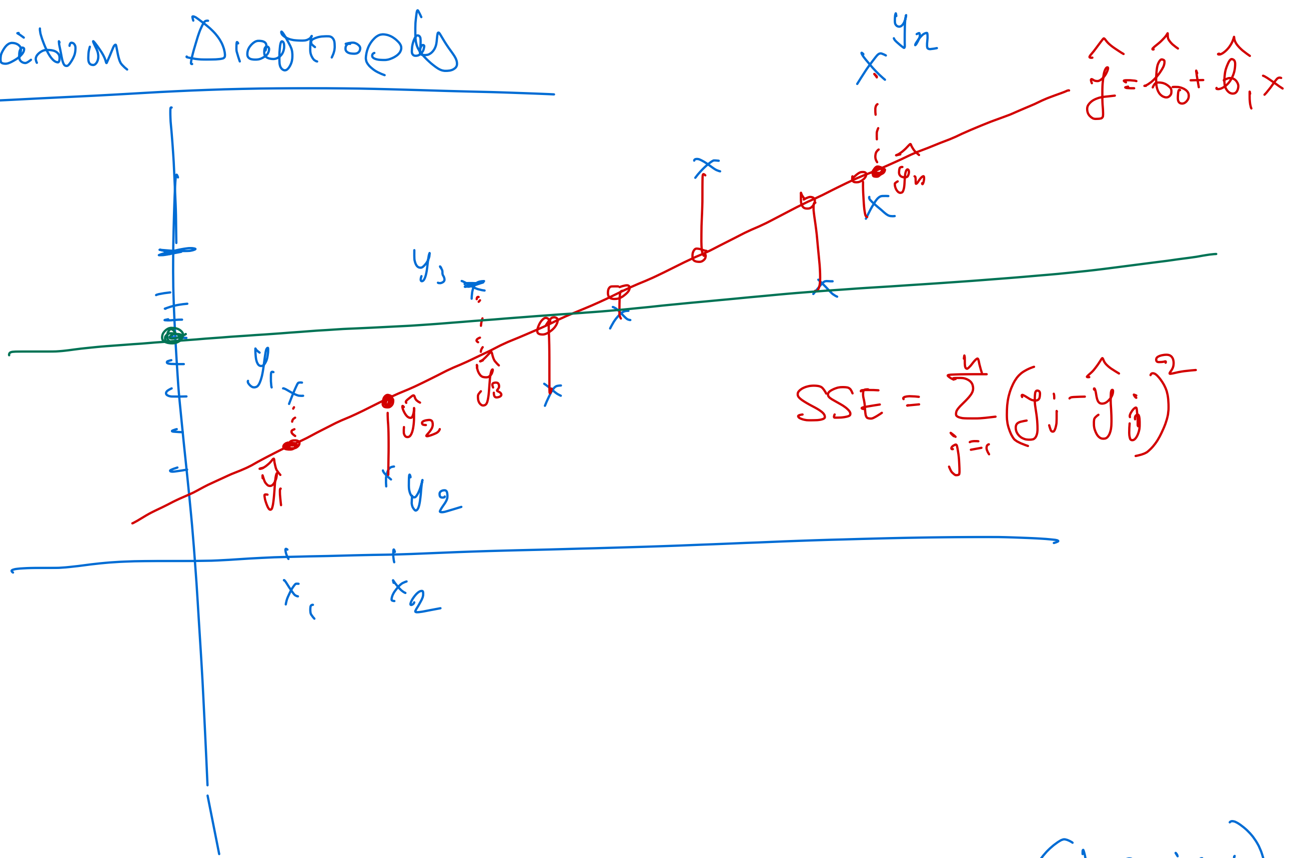


$H_0 : b_1 = 0 \qquad H_a : b_1 \neq 0$

Αν απορρ. $H_0 \Longrightarrow$ υπάρχει ένδειξη συσχέτισης μεταξύ x κ' y

έλεγχος στατιστικής σημαντικότητας του b_1 κ' ταυτ' ελέγχου του μοντέλου

Ανάλυση Διασποράς



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

"Μηδενικό Μοντέλο" $\quad Y = \mu + \varepsilon \quad$ (κανένα x)

$$\hat{Y} = \hat{\mu} = \bar{Y}_n$$

$$SSR = SS_{model} = \sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2$$

Sum Square regression (model)

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

sum of squares total

συνολική μεταβλητότητα του Y (ανεξάρτητο από το μοντέλο).

Θεώρημα (ANOVA)

Όταν τα $\hat{b}_0, \hat{b}_1$ από LSE

τότε $\boxed{SST = SSR + SSE}$

SSR: μεταβλητότητα εξηγείται από το μοντέλο

SSE: παραμένει ανεξήγητη

$$R^2 = \frac{SSR}{SST} = \%$$ μεταβλ. του Y που εξηγείται από το μοντέλο