

14-1-2025

Μοντέλο Πολλαπλής Παμνδρόμησης

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k + \varepsilon$$

$\left. \begin{matrix} X_1 \\ \vdots \\ X_k \end{matrix} \right\}$ ανεξάρτητες μεταβλητές / παράγοντες / predictors

Y : εξαρτημένη μεταβλητή / response

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

Δεδομένα

	X_1	X_2	$\dots$	X_k	Y
1	x_{11}	x_{12}	$\dots$	x_{1k}	y_1
2	x_{21}	x_{22}		x_{2k}	y_2
$\vdots$					
j	x_{j1}	x_{j2}		x_{jk}	y_j
$\vdots$					
n	x_{n1}			x_{nk}	y_n

$$\hat{y}_j = b_0 + b_1 x_{j1} + b_2 x_{j2} + \dots + b_k x_{jk}$$

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

$$Y = b_0 + b_1 X_1 + \dots + b_K X_K + \varepsilon$$

$X_1, \dots, X_K = \begin{cases} \text{διαφορετικά φυσικά μεγέθη} \\ \text{ανάπτυξης ως προς } X_j \end{cases}$

πχ. $X_2 = X_1^2$

① $Y = b_0 + b_1 X_1 + b_2 X_1^2$

② $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 \underbrace{X_1 X_2}_{X_3}$

③ $Y = b_0 + b_1 X_1 + b_2 \underbrace{\ln(X_1)}_{X_2}$

Μέθοδος Ελαχ. Τετραγώνων $\#b = k+1$

$$\min_{b_0, b_1, \dots, b_k} SSE \rightarrow \underbrace{\hat{b}_0, \hat{b}_1, \dots, \hat{b}_k}_{\text{Εκτιμήσεις Ελαχ. Τετραγώνων.}}$$

$$\begin{aligned} E(\hat{b}_0) &= b_0 & \hat{b}_0 &\sim N(b_0, \sigma_{\hat{b}_0}^2) \\ E(\hat{b}_1) &= b_1 & & \\ & \vdots & & \\ E(\hat{b}_k) &= b_k & \hat{b}_k &\sim N(b_k, \sigma_{\hat{b}_k}^2) \end{aligned}$$

Πίνακας ANOVA.

	SS	df	MS
Model	SSR	$\#b - 1 = df_{\text{mod}}$	$MSR = \frac{SSR}{df_{\text{mod}}}$
Error	SSE	$n - \#b = df_{\text{er}}$	$MSE = SSE / df_{\text{er}}$
	SST	$n - 1$	

$$SSR = \sum_{j=1}^n (\hat{y}_j - \bar{y})^2 \quad : \text{μεταβλ. που εξηγείται από το μοντέλο}$$

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

$$SST = SSR + SSE$$

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

$$E(MSE) = \sigma^2$$

Παρατηρήσεις

① $Y = b_0 + b_1 X + \varepsilon$, $E(Y|X=x) = b_0 + b_1 x$.

b_1 = κλίση (ρυθμός μεταβολής)

$b_0 = E(Y|X=0)$ [όταν 0 αφίχεται στις τιμές του X στο διάγραμμα]

② $E(Y|X_1=x_1, \dots, X_k=x_k) = b_0 + b_1 x_1 + \dots + b_k x_k$

$b_0 = E(Y|X_1=X_2=\dots=X_k=0)$

[όταν στο διάγραμμα υπάρχουν παρατηρήσεις με όλα τα $X_j = 0$]

③ b_1 = ρυθμός μεταβολής της EY ως προς X_1
όταν τα $X_2, \dots, X_k$ μένουν σταθερά

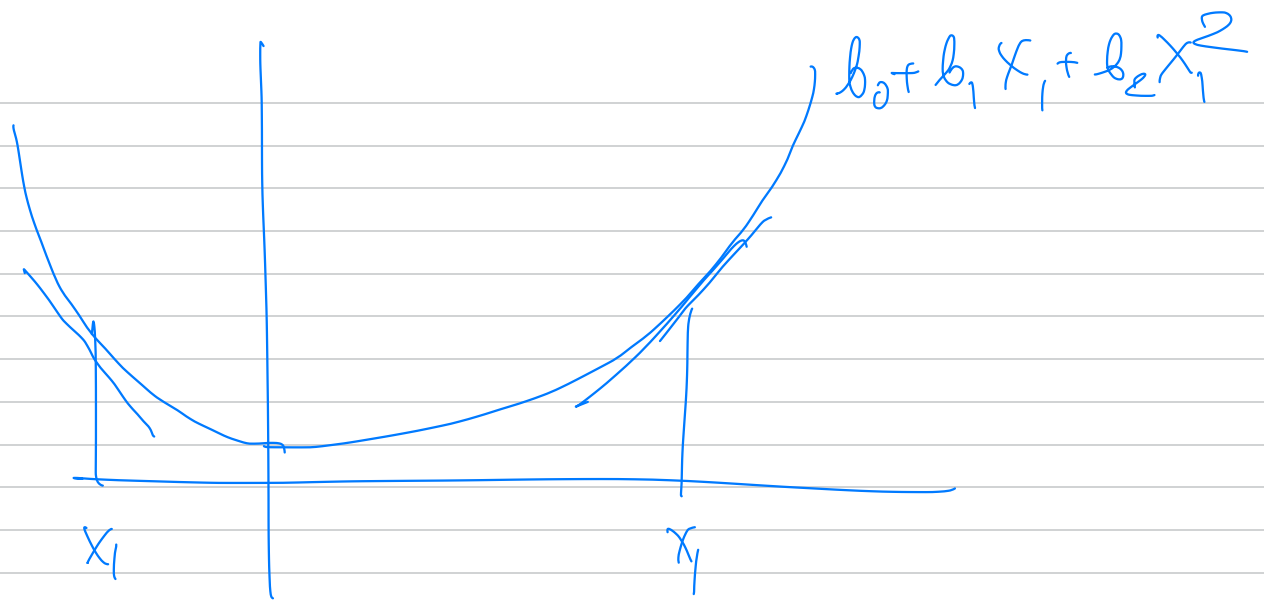
$b_1 = \frac{\partial EY}{\partial X_1}$, $b_2 = \frac{\partial EY}{\partial X_2}$...

(?)

④ Παράβ: $EY = b_0 + b_1 X_1 + b_2 X_1^2$

? $b_1 = \frac{\partial EY}{\partial X_1}$ ~~X~~

$\frac{\partial EY}{\partial X_1} = b_1 + 2b_2 X_1$!!



5) $EY = b_0 + b_1 \cdot X_1 + b_2 X_2$

$X_1 = \text{age}$

$X_2 = \text{height (cm)}$

$X_1 \in [0, 10]$

$X_2 \in [30, 120]$

$b_1 = \frac{\partial EY}{\partial X_1}$

(?)

Μπορεί $\text{Cor}(X_1, X_2) > 0$



$$\text{Ans } \text{Cor}(X_1, X_2) \approx 1$$

$$\Rightarrow X_2 = \gamma_0 + \gamma_1 X_1 \Rightarrow X_1 = \frac{X_2 - \gamma_0}{\gamma_1}$$

Τότε όμως : $Y = b_0 + b_1 X_1 + b_2 X_2 =$

$$= b_0 + b_1 X_1 + b_2 (\gamma_0 + \gamma_1 X_1)$$

$$= b_0 + b_2 \gamma_0 + (b_1 + b_2 \gamma_1) X_1$$

$$= \underline{\delta_0 + \delta_1 X_1}$$

$$= \dots$$

$$= \theta_0 + \theta_1 X_2$$

Πολυσυγγραμμικότητα (Multicollinearity)

Συσχέτιση μεταξύ ανεξάρτητων μεταβλητών.

Παράδειγμα

Exercise. Rdata.

$$Y = \text{score.}$$

$$X_1 = \text{age}$$

$$X_2 = \text{extime.}$$

①

$$Y = b_0 + b_1 \cdot \text{age}$$

$$SSR = 10\ 255\ 910$$

$$SSE = 1\ 485\ 592$$

$$\hline 11\ 741\ 503$$

$$R^2 = 0.8735$$

$$Y = b_0 + b_1 \cdot \text{extime}$$

$$SSR = 10\ 701\ 573$$

$$SSE = 1\ 039\ 929$$

$$\hline SST = 11\ 741\ 503$$

$$R^2 = 0.9114$$

$$Y = b_0 + b_1 \cdot \text{age} + b_2 \cdot \text{extime}$$

$$SSR = 10\ 808\ 248$$

$$SSE = 933\ 254$$

$$\hline SST = 11\ 741\ 503$$

$$R^2 = 0.9205$$

$$SSR(age) = 10255910$$

$$SSR(age, extime) = 10808248$$

$$SSR(age, extime) - SSR(age) = \underline{552338}$$
$$= SSR(extime | age)$$

Γενικά $SSR(extime | age) \neq SSR(extime)$

γιατί;

Παρουσιάζουμε !!

Η μεταβλητότητα του Y που εξηγείται από μια ανεξ. μεταβλητή X , ή η παρατηρούμενη ομοιογένεια της X ως predictor εξαρτάται από τις άλλες μεταβλητές που ήδη υπάρχουν στο μοντέλο παλινδρόμησης