14-1-2025

# Μοντέλο Πολλαπλής Παλινδρόμησης

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_K X_K + \varepsilon$$

$X_1, \ldots, X_K$ } ανεξάρτητες μεταβλητές / παράγοντες / predictors

$Y$ : εξαρτημένη μεταβλητή / response

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

Δείγμα

| | $X_1$ | $X_2$ | ... | $X_K$ | $Y$ |
|---|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{12}$ | ... | $x_{1K}$ | $y_1$ |
| 2 | $x_{21}$ | $x_{22}$ | | $x_{2K}$ | $y_2$ |
| ⋮ | | | | | |
| $j$ | $x_{j1}$ | $x_{j2}$ | | $x_{jK}$ | $y_j$ |
| ⋮ | | | | | |
| $n$ | $x_{n1}$ | | | $x_{nK}$ | $y_n$ |

$$\hat{y}_j = b_0 + b_1 x_{j1} + b_2 x_{j2} + \cdots + b_K x_{jK}$$

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

$$Y = b_0 + b_1 X_1 + \cdots + b_k X_k + \varepsilon$$

$X_1, \ldots, X_k$ = { διαφορετικά φυσικά μεγέθη
συναρτήσεις αρχικών $X_i$

π.χ. $X_2 = X_1^2$

(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 $\#b = k+1$

$$\min_{b_0, b_1, \dots, b_k} SSE \longrightarrow \underbrace{\hat{b}_0, \hat{b}_1, \dots, \hat{b}_k}_{\text{εκτιμώσες ελαχ. τετραγώνων.}}$$

$$E(\hat{b}_0) = b_0 \quad \hat{b}_0 \sim N(b_0, \sigma^2_{\hat{b}_0})$$
$$E(\hat{b}_1) = b_1$$
$$\vdots$$
$$E(\hat{b}_k) = b_k \quad \hat{b}_k \sim N(b_k, \sigma^2_{\hat{b}_k})$$

---

Πίνακας ANOVA.

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

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

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

$$SST = SSR + SSE$$

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

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

# 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 ο περιέχεται στις τιμές του $X$ στο δείγμα]

(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 δείγμα υπάρχουν παρατηρήσεις με όλα τα $X_i = 0$]

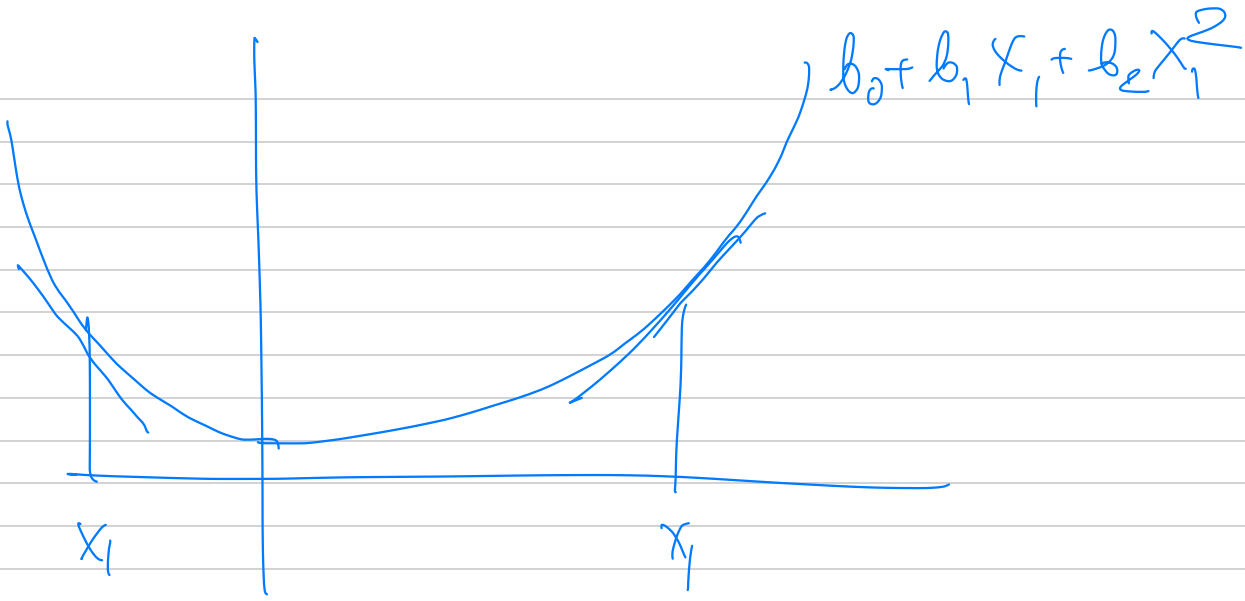
(3) $b_1$ = ρυθμός μεταβολής της $EY$ ως προς $X_1$ όταν τα $X_2, \dots, X_k$ μένουν σταθερά

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

(4) Παρατ: $EY = b_0 + b_1 X_1 + b_2 X_1^2$

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

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




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

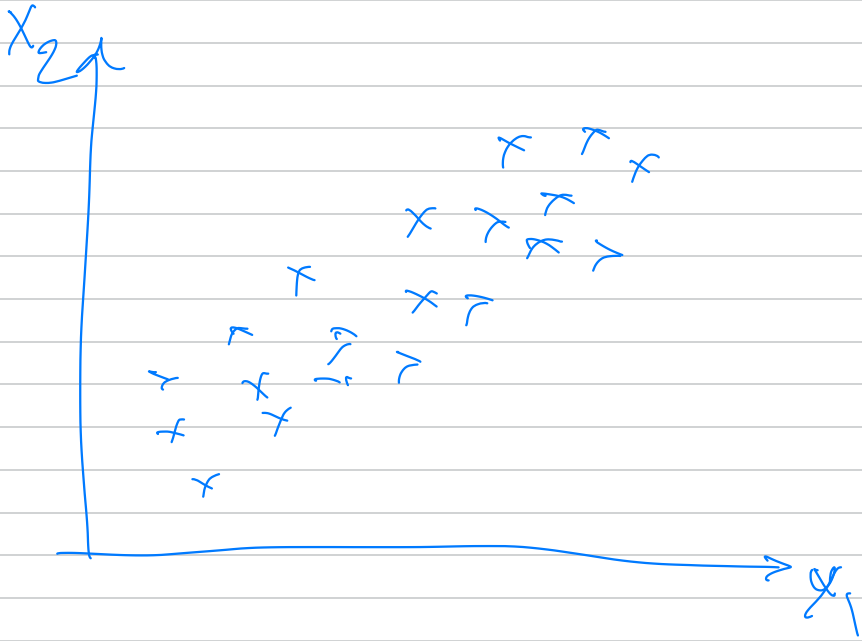
$X_1$ = age

$X_2$ = height (cm)

$X_1 \in [0, 10]$

$X_2 \in [30, 120]$

$b_1 = \frac{\partial EY}{\partial X_1}$ ( ? ) Μπορεί $Cor(X_1, X_2) > 0$



Αν $Cor(X_1, X_2) \approx 1$

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

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

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

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

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

$$= \dots$$

$$= \theta_0 + \theta_1 X_2$$

Πολυγραμμικότητα (Multicollinearity)

Συσχέτιση μεταξύ ανεξάρτητων μεταβλητών.

# Παράδειγμα Exercise. Rdata.

$Y$ = score.

$X_1$ = age

$X_2$ = extime.

---

① $Y = b_0 + b_1 \cdot age$

SSR = 10 255 910

SSE = 1 485 592

SST = 11 741 503

$R^2 = 0.8735$

$Y = b_0 + b_1 extime$

SSR = 10 701 573

SSE = 1 039 929

SST = 11 741 503

$R^2 = 0.9114$

$Y = b_0 + b_1 age + b_2 extime$

SSR 10808248

SSE 933254

SST 11 741 503

$R^2 = 0.9205$

$SSR(age) = 10255910$

$SSR(age, extime) = 10808248$

$SSR(age, extime) - SSR(age) = 552338$

$= SSR(extime \mid age)$

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

(γιατί)

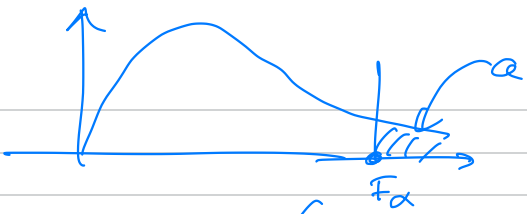
παρουσιαστικά !!

---

Η μεταβλητότητα του Y που εξηγείται από μια ανεξ. μεταβλητή X, η παραπάνω συμπεριφορά της X ως predictor εξαρτώνται από τις άλλες μεταβλητές που ήδη υπάρχουν στο μοντέλο παλινδρόμησης

(1)

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{df_{mod}}}{\frac{SSE}{df_{er}}} \sim F(df_{mod}, df_{er})$$




F statistic συναρτήσει μοντέλου

Έλεγχος F για συνολικό μοντέλο

$H_0: b_1 = b_2 = \dots = b_k = 0$ $\quad$ $H_1$: τουλάχιστον ένα $\neq 0$

$H_0: EY = b_0$ (όλο το μοντέλο μια σταθερά συμπυκνώνει)

$F > F_a \Leftrightarrow$ reject $H_0$ $\quad p < \alpha$

$F \leq F_a \Leftrightarrow$ accept $H_0$ $\quad p \geq \alpha$

---

Ειδική περίπτωση ($k=1$)

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

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{1}}{\frac{SSE}{df_{er}}} \sim F(1, df_{er})$$

$H_0: b_1 = 0$ $\quad$ $H_1: b_1 \neq 0$ —

Όπως είχαμε δει ως

$$\left\{ H_0: b_1 = 0 \quad H_1: b_1 \neq 0 \right\}$$ t-test για $b_1$.

ως μονοπαράγοντα

$$t = \frac{\hat{b}_1}{\hat{s}_{b_1}}$$

Στο μοντέλο $Y = b_0 + b_1 X$

$$F = \frac{MSR}{MSE} = t^2$$

Από πθ. $F(1, k) \overset{d}{=} t_k^2$

$$F_a = t_{a/2}^2$$

$$F \leq F_a \iff t^2 \leq t_{a/2}^2 \iff |t| \leq t_{a/2}$$

accept $H_0$ accept $H_0$

(2) Έλεγχος F για μέρος του μοντέλου

Model 1 (full model) $Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + b_{p+1} X_{p+1} + \cdots + b_k X_k$

Model 2 (partial model) $Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p$ , $(p<k)$

Model 2 nested στο Model 1 (εσωτερικό)

Ξέρουμε $SSR(X_1, X_2, \ldots, X_p, \ldots X_k) \geq SSR(X_1, \ldots, X_p)$

$\frac{SSE(X_1, X_2, \ldots X_p, \ldots, X_k)}{SST} \leq \frac{SSE(X_1, \ldots, X_k)}{SST}$

Επομένως $SSR(X_1, \ldots, X_k) - SSR(X_1, \ldots, X_p) =$

$= SSR(X_{p+1}, \ldots, X_k \mid X_1, \ldots, X_p) \geq 0$

$\Rightarrow R^2_{full} \geq R^2_{partial}$

$k-p$ συντελεστές

$H_0 : \beta_{p+1} = \beta_{p+2} = \cdots = \beta_k = 0$ , $H_1$ : τουλάχιστον ένα $\neq 0$

έλεγχος για μέρος του model 1

$SSR_{full} + SSE_{full} = SSR_{par} + SSE_{par} \Rightarrow$

$\Rightarrow SSR_{full} - SSR_{par} = SSE_{par} - SSE_{full}$

$df_{er,full} = n - (k+1) = n-k-1$

$df_{er,par} = n - (p+1) = n-p-1 > n-k-1$

$$df_{er,par} - df_{er,full} = k-p$$

$H_0 : b_{p+1} = b_{p+2} = \dots = b_k = 0$ , $H_1$ : τουλάχιστον ένα $\neq 0$

$$F = \frac{\frac{SSE_{par} - SSE_{full}}{df_{er,par} - df_{er,full}}}{MSE_{full}} \overset{k-p}{\sim} F(k-p,\ n-k-1)$$

accept $H_0$ αν $F \le F_\alpha(k-p,\ n-k-1)$

reject $H_0$ αν $F > F_\alpha(k-p,\ n-k-1)$

Εξ. περιττών 1

Ο έλεγχος F για όλο το Μοντέλο εδική περίπτωση

full $Y = b_0 + b_1 X_1 + \dots + b_k X_k$ full

part $Y = b_0$

$H_0 : b_1 = b_2 = \dots = b_k = 0$. $H_1$: τουλ ένα $\neq 0$

Το partial model εδώ $Y = b_0 \Rightarrow \hat{b}_0 = \bar{Y}$

$$SSE_{partial} = \sum (Y_i - \bar{Y})^2 = SST$$

$$SSE_{par} - SSE_{full} = SST - SSE_{full} = SSR_{full}$$

$$df_{er,par} - df_{er,full} = k.$$

$$F = \frac{\frac{SSR_{full}}{df_{mod,full}}}{MSE_{full}} = \frac{MSR_{full}}{MSE_{full}} = F$$

↓ για το test των μοντέλων

## Ειδική περίπτωση 2

full $Y = b_0 + b_1 X_1 + \cdots + b_p X_p + b_{p+1} X_{p+1}$

partial $Y = b_0 + b_1 X_1 + \cdots + b_p X_p$

$H_0 : b_{p+1} = 0 \qquad H_1 : b_{p+1} \neq 0.$

$\Updownarrow$

t-test $b_{p+1}$

Ο έλεγχος t για $b_j = 0$ ή $b_j \neq 0$ ελέγχει τη σημαντικότητα της $X_j$

δεδομένων όλων των υπόλοιπων μεταβλητών

(δηλ. όταν η $X_j$ μπαίνει τελευταία παρουσία όλων των άλλων).

(3) Έστω ότι το μοντέλο

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

οι μεταβλητές μπαίνουν με τη σειρά

$$X_1, X_2, \ldots, X_k$$

Model 1 : $X_1$ } $H_0: b_2 = 0$ $H_1: b_2 \neq 0$ το Model 2 — $F_1$

" 2 : $X_1, X_2$ } $H_0: b_3 = 0$ $H_1: b_3 \neq 0$ Model 3 — $F_2$

3 : $X_1, X_2, X_3$

⋮

k : $X_1, X_2, X_3, \ldots, X_k$ : $F_k$

$F_1, F_2, \ldots, F_k$ : Ftests typeI (sequential Ftests)

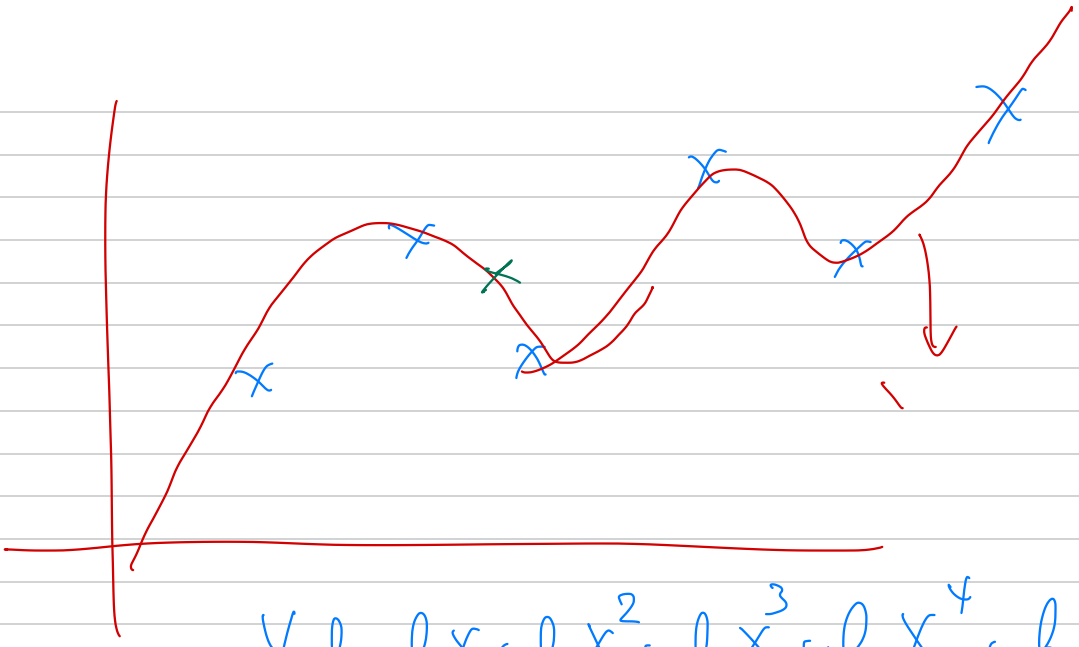
εξαρτώνται από τη σειρά εισαγωγής

(4) Ftests type III (partial)

$H_0: b_j = 0$ $H_1: b_j \neq 0$ στο Model k

(Η $X_j$ τελευταία) $\Leftrightarrow$ t-tests

δεν εξαρτώνται από τη σειρά εισαγωγής




$$Y = b_0 + b_1 X + b_2 X^2 + b_3 X^3 + b_4 X^4 + b_5 X^5$$

$$SSE = 0 \Rightarrow R^2 = 1$$

$$df_{er} = 0$$

$$MSE = \frac{0}{0} \quad ?$$

$$F = \frac{MSR}{MSE} \quad ??$$