2023-12-15

Μοντέλο πολλαπλής παλινδρόμησης

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

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

$$Y | x_1, x_2 \ldots, x_k \sim N(b_0 + b_1 x_1 + \cdots + b_k x_k, \sigma^2)$$

Δείγμα

j	X_1	X_2	...	X_k	Y
1	x_{11}	x_{21}	...	x_{k1}	y_1
2	x_{12}	x_{22}	...	x_{k2}	y_2
⋮	⋮				
n	x_{1n}	x_{2n}	...	x_{kn}	y_n

$X_1, \ldots, X_k$ → διαφορετές μεταβλητές

→ συχνείτες αξίων μεταβλητών

π.χ. (1) $y = b_0 + b_1 \cdot age + b_2 \cdot height + b_3 \cdot weight$

x_1 x_2 x_3

(2) $y = b_0 + b_1 \cdot age + b_2 \cdot age^2$

x_1 $x_2 = x_1^2$

(3) $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$

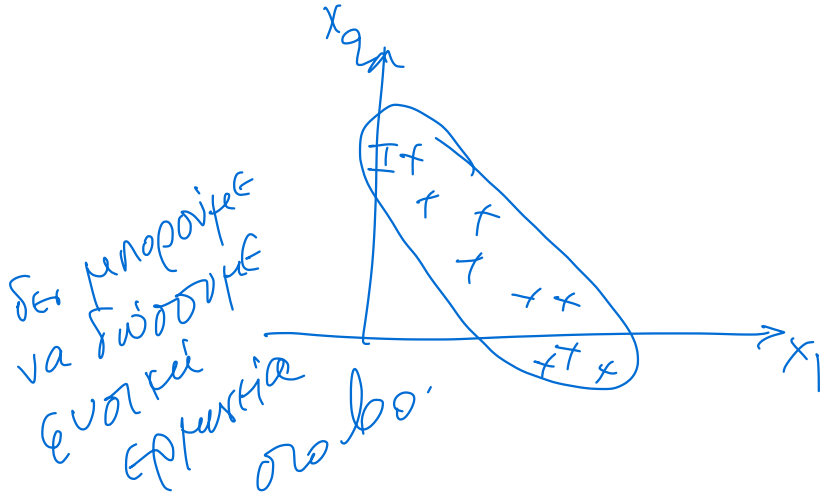
x_1 x_2 $x_3 = x_1 x_2$

Ερμηνεία

$$y_0 = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$$

(1) $b_0 = E(Y \mid x_1 = 0, x_2 = 0)$

έχει ερμηνεία όταν στο δείγμα υπάρχουν παρατηρήσεις με $x_1 = x_2 = 0$



δεν μπορούμε να δώσουμε ουσιαστική ερμηνεία στο b_0.

(2) $y_0 = b_0 + b_1 x_1 + b_2 x_2$

b_1 = μέση μεταβολή του Y αν το x_1 αυξηθεί κατά 1 μονάδα, κ' οι υπόλοιπες μεταβλητές παραμείνουν σταθερές

π.χ. $x_1 = age$ (έτη), $x_2 = height$

b_1 : ελάττωση του Y της αίγυπτος κατά 1 έτος, κρατώντας ύψος = σταθ.

$y = b_0 + b_1 x_1 + b_2 x_2$ $\quad b_1 = \frac{\partial y}{\partial x_1}$ ←

(3) $y = b_0 + b_1 x + b_2 x^2$

x_1 $\quad$ x_2

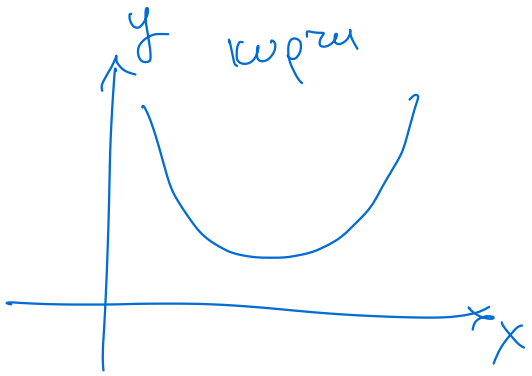
δεν είναι δυνατό να αυξηθεί το x_1 κατά 1 κ' το x_2 να μείνει σταθερό

$\frac{\partial y}{\partial x} = b_1 + 2 b_2 x$

b_1 δεν έχει φυσική ερμηνεία

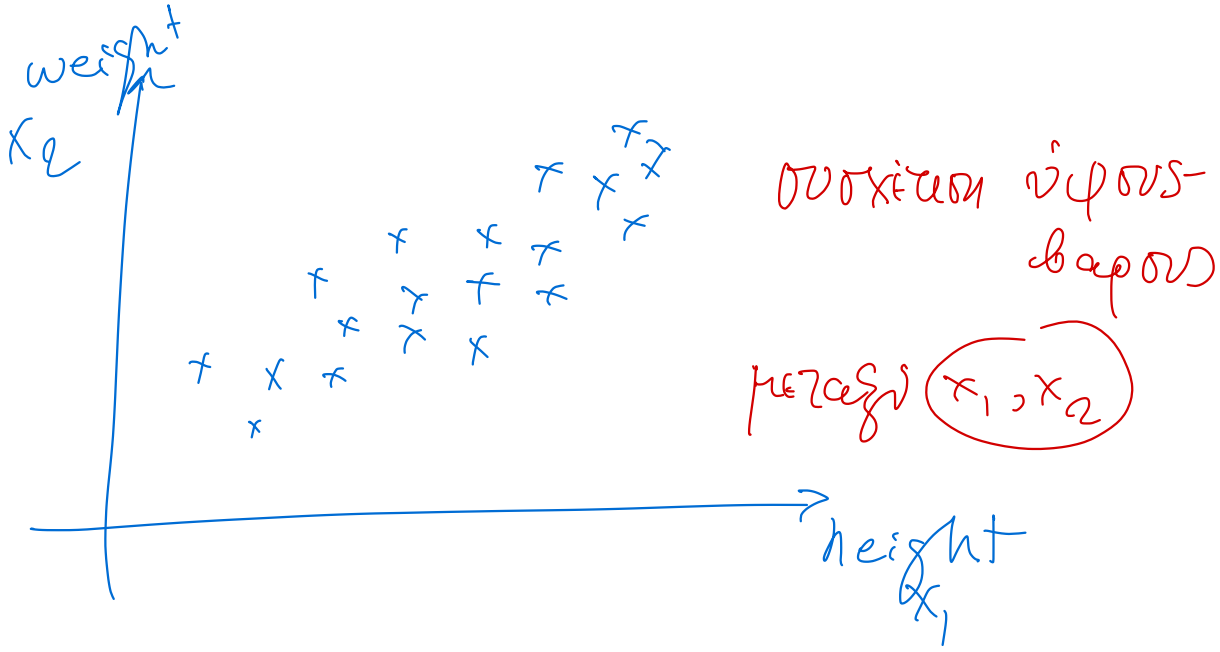
b_2 επίσης

$b_2 > 0$



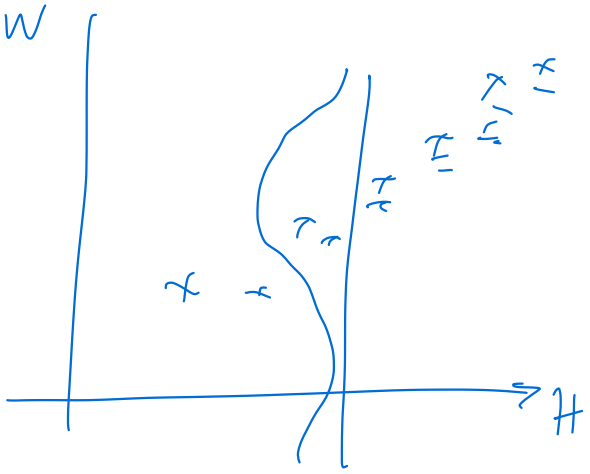
$b_2 < 0$

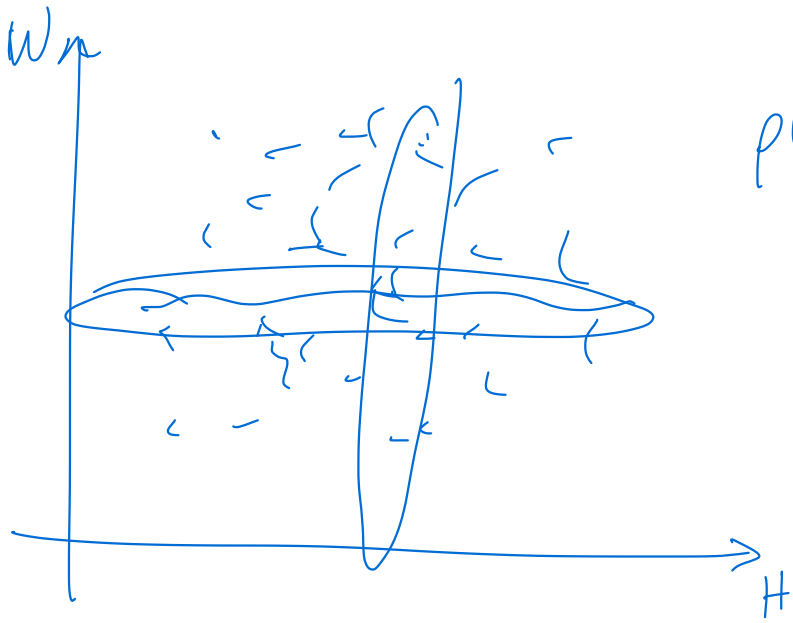
(A) $Y = b_0 + b_1 \cdot \text{height} + b_2 \cdot \text{weight} + \varepsilon$



συσχέτιση ύψους-βάρους μεταξύ (x_1, x_2)

αν $\rho(x_1, x_2) \simeq 1$





$\rho(X_1, X_2) \approx 0$

Μοντέλο $\quad Y_0 = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_K X_K + \varepsilon$

$b_0, b_1, \ldots, b_K$: άγνωστες παράμετροι $(K+1)$

Εκτίμηση ελαχ. τετραγώνων

$\hat{y} = \hat{b}_0 + \hat{b}_1 + \cdots + \hat{b}_K X_K \leftarrow$ προβλεπόμενης τιμές

$$SSE(b_0, b_1, \ldots, b_K) = \sum_{j=1}^{K} \left(y_j - b_0 - b_1 x_{1j} - b_2 x_{2j} - \cdots - b_K x_{Kj} \right)^2$$

$\min_{b_0, b_1, \ldots, b_K} SSE \Rightarrow \ldots - \hat{b}_0, \hat{b}_1, \ldots, \hat{b}_K$: εκτιμήσεις ελαχ. τετραγώνων

Ανάλυση Διασποράς

$$SST = \sum_j (y_j - \bar{y})^2 \quad \text{(συνεξ. μοντέλου)}$$

$$SSR = \sum_j (\hat{y}_j - \bar{y})^2$$: μεταβλητότητα του Y που εξηγείται από το μοντέλο

$$SSE = \sum_j (y_j - \hat{y}_j)^2$$: μεταβλητότητα που παραμένει ανεξήγητη

$$SST = SSR + SSE$$

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

Εκτιμητές για τα β

Πίνακας ANOVA

	SS	df	MS
Model	SSR	$df_{mod} = k$	$MSR = \frac{SSR}{df_{mod}}$
Error	SSE	$n-(k+1)$	$MSE = \frac{SSE}{df_{er}}$
Total	SST	$n-1$	

$SSR \sim \chi^2_k$

$SSE \sim \chi^2_{n-k-1}$

$SST \sim \chi^2_{n-1}$

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

$df_{er} =$ n − # παραμέτρων β στο μοντέλο $= n-(k+1)$

$\hat{\sigma}^2$: MSE αμερόληπτη εκτιμήτρια του σ^2

$E(MSE) = \sigma^2$

Έλεγχος F (Ftest) για το συνολικό μοντέλο

$H_0: b_2 = 0$ $\quad$ $H_1: b_2 \neq 0$ $\quad$ X

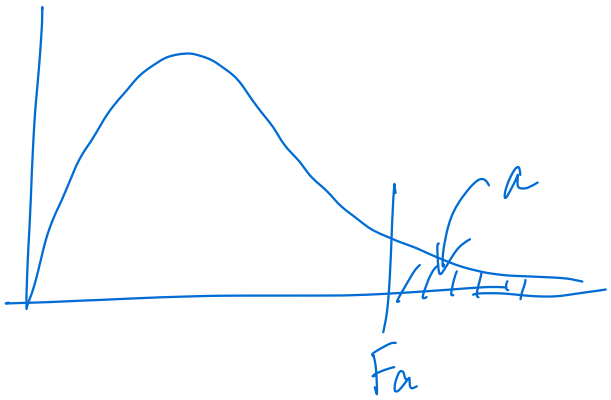
$H_0: b_1 = b_2 = \cdots = b_k = 0$ $\quad$ H_1: τουλάχιστον ένα από $b_1, \ldots, b_k \neq 0$

Στατιστικό ελέγχου $\quad F = \dfrac{MSR}{MSE}$

Accept H_0 if $F \leq F_\alpha$ } κρίσιμη περιοχή $p \geq \alpha$

Reject H_0 if $F > F_\alpha$ } $p < \alpha$

Αν H_0 : $F = \dfrac{MSR}{MSE} \sim F_{df_{mod}, df_{er}}$



Μονοπαραγοντικό Μοντέλο

F-test $H_0: b_1 = 0$ $H_1: b_1 \neq 0$

t-test $H_0: b_1 = 0$ $H_1: b_1 \neq 0$

} μπορεί να δώσουν διαφορετικά αποτελέσματα

Όχι

Στο μονοπαραγοντικό

ισχύει $F = \frac{MSR}{MSE} = t^2$ για το b_1

$= \frac{b_1^2}{s^2_{\hat{b}_1}}$ $\left(t = \frac{\hat{b}_1}{s_{\hat{b}_1}} \right)$

$$F_{\alpha, 1, n-2} = \left(t_{\alpha/2, n-2} \right)^2$$

$$F > F_\alpha \iff t^2 > t^2_{\alpha, n-2} \iff |t| > t_{\alpha/2, n-2}$$

Reject H_0 με το F-test

$\Longleftrightarrow$

reject H_0 με το t-test

Παράδειγμα y, x_1, x_2, x_3

Model 1 : $y = b_0 + b_1 x_1 + \varepsilon$

$H_0: b_1 = 0$, $H_1: b_1 \neq 0$

$p \approx 0$

SSR	3142
SSE	5235
SST	8378

$R^2 = 0.3751$

Model 2 $y = b_0 + b_1 x_3 + \varepsilon$

SSR	2377
SSE	6000
SST	8377

$R^2 \approx 0.28$

$H_0: b_1 = 0$ $b_1 \neq 0$ (x_3) ✗

$p \approx 0$

Model 3 $y = b_0 + b_1 x_1 + b_3 x_3 + \varepsilon$ xx

"Πρόχειρα" $SSR \approx 3142 + 2377$

$R^2 \approx 0,38 + 0,28 \approx 66\%$ (??)

Πραγματικά

$SSR = 3158$

$SSE = 5219$

8378

$R^2 = 0,3770$

$H_0: b_1 = 0$ $b_1 \neq 0$ $p \approx 0$.

$\rightarrow H_0: b_3 = 0$ $b_3 \neq 0$ $p = 0.588$ (x_3)

$$SSR(X_1, X_3) = 3158$$

$$SSR(X_1) = 3142$$

$$SSR(X_1, X_3) - SSR(X_1) = 16$$

Επιπλέον μεταβλητών Y που εξηγείται από την X_3 όταν προστεθεί στην X_1

$$SSR(X_3 | X_1)$$

$$SSR(X_3) = 2377$$

$$SSR(X_3 | X_1) = 16$$

Η πολλαπλή σημαντικότητα μιας μεταβλητής εξαρτάται από το μοντέλο στο οποίο περιλαμβάνεται

Κατά κανόνα αυτές οι διαφορές οφείλονται σε συσχετίσεις μεταξύ ανεξάρτητων μεταβλητών (πολυσυγγραμμικότητα multicollinearity)

Στο μοντέλο $y = b_0 + b_1 x_1 + \dots + b_k x_k + \varepsilon$

t-test $H_0 : b_1 = 0 \qquad H_1 : b_1 \neq 0$

Ελέγχει τη στατ. σημαντικότητα του b_1 (δηλ. της X_1) δεδομένου ότι προσθέτουμε στο μοντέλο που ήδη είχε όλες τις υπόλοιπες !!!

Θρησκεία

π.χ.

Y $\quad X_1 = age$

$X_2 = αρτ. πίεση$ (BP

$Y_1 = b_0 + b_1\, age + b_2\, BP$

$H_0 : b_2 = 0 \qquad H_1 : b_2 \neq 0.$

Αν η BP είναι στατιστικά σημαντική

Πληροφορία της ηλικίας στο μοντέλο

Σε αυτό το μοντέλο εξετάζουμε την επίδραση της BP στο Y έχοντας ήδη στο μοντέλο την ηλικία

Σενάριο

Y	BP
⋮	⋮

$Y = b_0 + b_1 BP$

$P \simeq 0.001$

Πως επηρεάζουμε αν το effect είναι όντως παραπλανητικό ή όχι?

① | Y | age | BP |
|---|---|---|

$Y_0 = b_0 + b_1 age + b_2 BP$

$H_0 : b_2 = 0 \quad H_1 : b_2 \neq 0$

$P \simeq 0.2$

② $Y_0 = b_0 + b_1 age + b_2 BP$: $H_0 : b_2 = 0$, $H_1 : b_2 \neq 0$

$P = 0.005$

BP είναι παρ. σημ. έχοντας ελεγχθεί ως προς την ηλικία

Μεταβλητές: age, ύψος, βάρος : "δημογραφικές μεταβλητές" //

Επιπτώσεις του πολυσυγγραμμικού παράδειγμα

$Y = b_0 + b_1 X_1$: $R^2 = 0.3751$ $P_{x_1} \approx 0$

$Y = b_0 + b_2 X_2$: $R^2 = 0.298$ $P_{x_2} \approx 0$

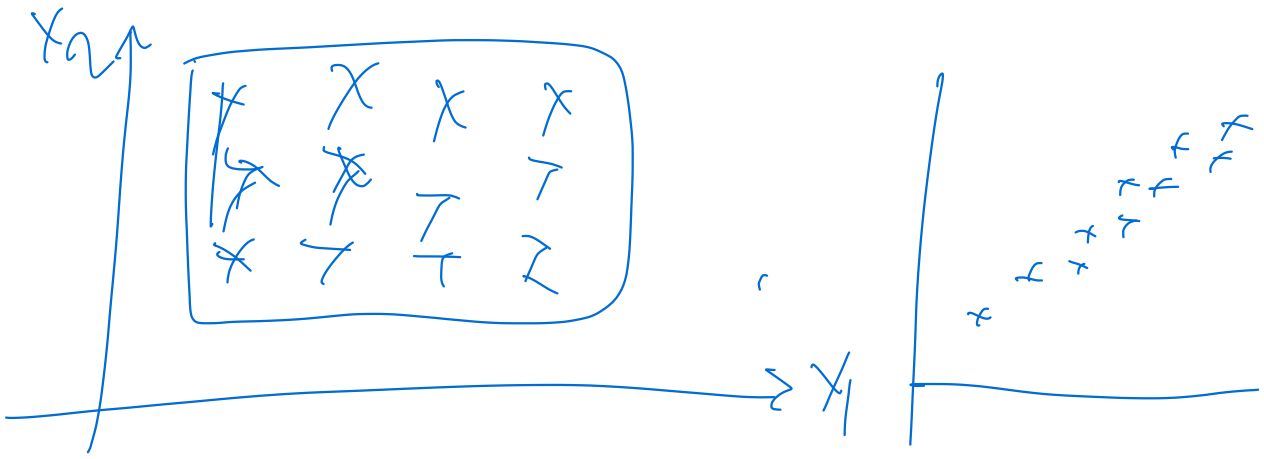
$Y = b_0 + b_1 X_1 + b_2 X_2$: $R^2 = 0.7255$ $P_{x_1} \approx 0$ $P_{x_2} \approx 0$

$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$

$= 6077 - 3142 \approx 2900$

$SSR(X_2) = 2444$

$\rho(X_1, X_2) \approx 0$ ← X_1, X_2 ορθογώνιες



$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

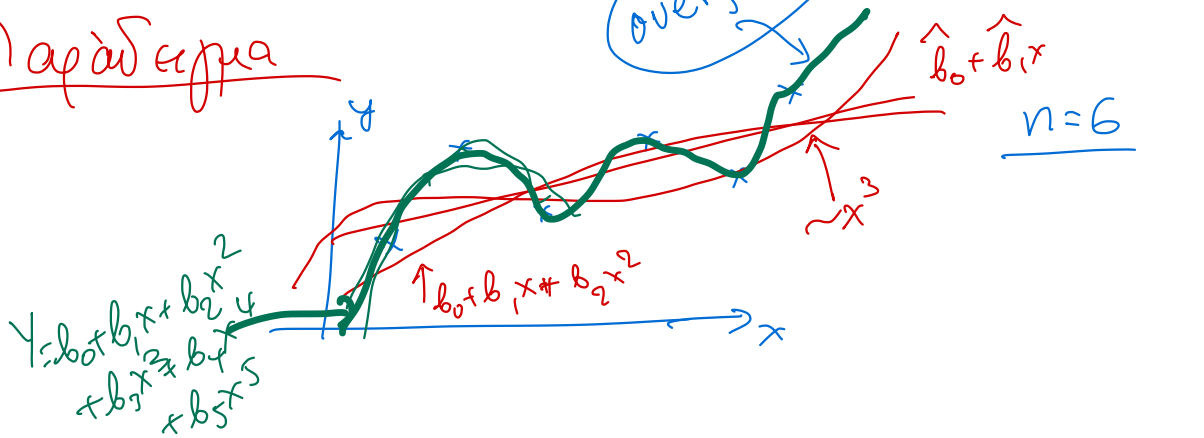
$R^2 = 0.7280$

$P_{x_1} \approx 0$

$P_{x_2} \approx 0$

$P_{x_3} \approx 0.342$

Παράδειγμα

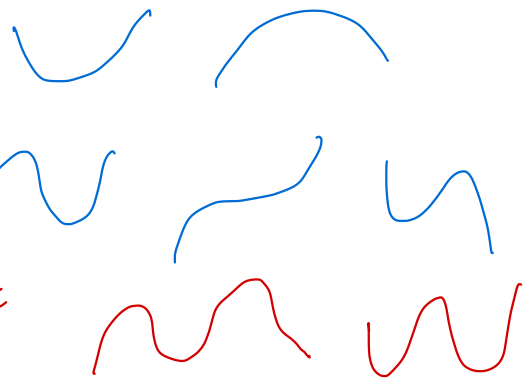


① $Y = b_0 + b_1 x + \varepsilon$

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

$Y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$

$+ b_4 x^4$



$Y = b_0 + b_1 x + b_2 x^2 + \dots + b_5 x^5$

$R^2 = 1$

$SSE = 0$

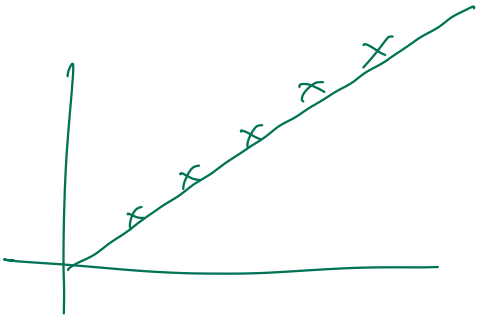
$SSR = SST$

$P_{b_1} \approx 1$

P_{b_2}

$df_{er} = n - \#b = 6 - 6 = 0$

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



$SSE = 0$

$df_{er} = n - 1$